TRAUNER VERLAG
UNIVERSITÄT

IDIMT-2024
Changes to ICT, Management, and Business Processes through AI

32nd Interdisciplinary Information Management Talks

With pride we present the proceedings of the 32nd annual IDIMT Conference.

Artificial Intelligence continues to make impressive advances and almost every month a new system is presented. But does this translate into a proportionate use in business – and as a second step, business advantages? In this conference we investigate various aspects of new developments to ICT itself as well as its use for management and business processes and their transformation. Ethical and security aspects as well as specific application areas (teaching, autonomous vehicles, supply chain management, social media) are touched too, to obtain an encompassing view of the topic.

We have chosen the following 11 topics for 2024:

- AI Support for Crisis Management (Neubauer, Rainer)
- Cyber Security (Sonntag)
- AI in Virtual Collaboration, Teaching & Learning (Jantos)
- Autonomous Vehicles and Smart Environments (Schoitsch)
- ICT Systems and Business (Doucek, Maryska)
- Digital transformation and digital business models (Pucihar)
- Social Media and the Role of AI (Pavlíček)
- Data and AI in Supply Chain Management (Delina, Tkáč)
- Academic Business Co-operation (Pitner)
- Ethical Integrity of Research in AI (Lisnik)
- Special session: Early Career & Student Showcase (Sonntag)

Based on a double-blind two-step review process we have selected 54 of the submitted papers with a totality of more than 126 co-authors. The program ran in two parallel streams.

The authors come from 11 different countries: Armenia, Austria, Czech Republic, Estonia, Germany, Greece, Netherlands, Poland, Slovakia, Slovenia and Spain.

DOI:10.35011/IDIMT-2024

ISBN 978-3-99151-527-2

www.trauner.at

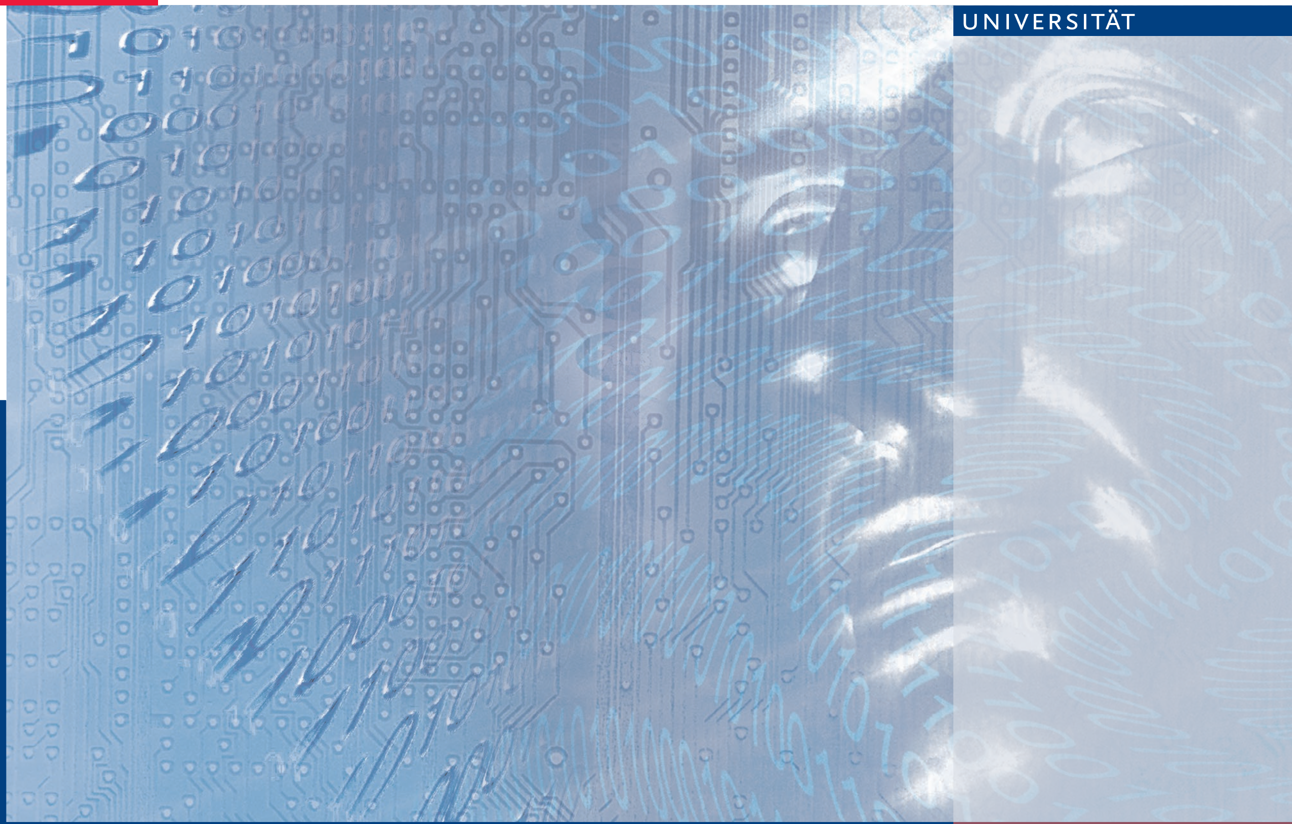
TRAUNER

IDIMT-2024

Doucek ■ Sonntag ■ Nedomova (Editors)

Band 53

TRAUNER VERLAG
UNIVERSITÄT



Schriftenreihe Informatik

53

Doucek Petr ■ Sonntag Michael ■ Nedomova Lea (Editors)

IDIMT-2024
Changes to ICT, Management, and Business Processes through AI

32nd Interdisciplinary Information Management Talks

Sept. 4–6, 2024

Hradec Králové, Czech Republic

JKU
JOHANNES KEPLER UNIVERSITÄT LINZ

t.

TRAUNER VERLAG

UNIVERSITÄT

Schriftenreihe Informatik

53

DOUCEK PETR ■ SONNTAG MICHAEL ■
NEDOMOVA LEA (EDITORS)

IDIMT-2024

Changes to ICT, Management, and Business Processes through AI

32nd Interdisciplinary Information Management Talks

Sept. 4–6, 2024
Hradec Králové, Czech Republic

Impressum

Schriftenreihe Informatik

Doucek Petr ■ Sonntag Michael ■
Nedomova Lea (Editors)
IDIMT-2024
Changes to ICT, Management,
and Business Processes through AI
32nd Interdisciplinary Information Management Talks

This publication was partially supported by the Prague University of Economics, Business – project IGA 409033 and institutional support for the long-term conceptual development of science and research at Faculty of Informatics and Statistics (IP400040), and the Johannes Kepler University Linz.

The Conference IDIMT-2024
took place September 4–6, 2024
in Hradec Králové, Czech Republic

Herausgeber:
Prague University of Economics
and Business
nám. W. Churchilla 1938/4
130 67 Praha 3
Czech Republic

Kommisionsverlag:
TRAUNER Verlag + Buchservice
GmbH
Köglstraße 14, 4020 Linz
Österreich/Austria

Herstellung:
paco Medienwerkstatt,
1160 Wien, Österreich/Austria

DOI: 10.35011/IDIMT-2024

ISBN 978-3-99151-527-2
www.trauner.at

TABLE OF CONTENS

AI SUPPORT FOR CRISIS MANAGEMENT

CYBER SECURITY

AI IN VIRTUAL COLLABORATION, TEACHING & LEARNING

AUTONOMOUS VEHICLES

ICT SYSTEMS AND BUSINESS

DIGITAL TRANSFORMATION AND DIGITAL BUSINESS MODELS

SOCIAL MEDIA AND THE ROLE OF AI

DATA AND AI IN SUPPLY CHAIN MANAGEMENT

ACADEMIC BUSINESS CO-OPERATION

ETHICAL INTEGRITY OF RESEARCH IN AI

EARLY CAREER & STUDENT SHOWCASE

ANNEX

SPONSORS OF IDIMT 2024

AI SUPPORT FOR CRISIS MANAGEMENT

ARTIFICIAL INTELLIGENCE IN CRISIS MANAGEMENT: POTENTIAL SOLUTIONS AND CHALLENGES

Karin Rainer, Viktoria Kundratitz, Melissa Hagendorn

Austrian Agency for Health and Food Safety GmbH
karin.rainer@ages.at

Georg Neubauer, Dražen Ignjatović

Austrian Institute of Technology GmbH
georg.neubauer@ait.ac.at

Ruth Kutalek

Medical University of Vienna
ruth.kutalek@meduniwien.ac.at

Nadine Sturm

Johanniter Ausbildung und Forschung GmbH
nadine.sturm@johanniter.at

DOI: 10.35011/IDIMT-2024-13

Keywords

Crisis management; decision making; technological tooling; responsible AI; pandemic management

Abstract

This article explores the potential solutions as well as challenges of AI-driven tools for crisis management. AI application in crisis management presents ethical, technical, ecological, and legal challenges, including biases, transparency, carbon footprint, and regulatory inadequacies, requiring robust frameworks and structures to ensure a sustainable, fair and responsible transfer in societal contexts. By drawing on examples from the ROADS to Health project and fundamental principles of crisis management, the paper advocates for a human-centered approach that places prioritization on user needs, cross sectional collaboration, and cultural sensitivity. The perspective of the evolution and further development of a tool stemming from the ROADS to Health project provides an illustration of how to bridge the challenges and facilitate targeted and comprehensive control measures in crisis scenarios. The proposed approach seeks to harness and maximize the potential of

AI while ensuring and safeguarding its ethical integrity and effectiveness in supporting crisis management efforts.

1. Introduction

Our societies are currently facing a rise in the frequency of global health crises, while immersing in a technologically advanced era with increasing dependency on digital technology than ever before. The use of Artificial Intelligence (AI) driven technology is often postulated as a promise of hope to deal with the complex and mostly pressing challenges in crisis management. The capabilities fostered by these applications can help to advance mitigation, preparedness, response and recovery in cases of health and complex crisis situations. Clearly, this pervasive technology contributes much added value to effective crisis management, ranging from the early detection of infectious disease outbreaks to tracing prediction, interactive dashboards, diagnosing risks, suggesting treatments, interventions, and cures, supporting social compliance and many other applications (Gaur et al. 2021). However, this new development and its implications are not yet well studied and understood in all of their complexities. Scholars warn of the risks and potential threat that these developments pose, as the reliance on AI and the broader implementation of digital technology adds another layer of interdependencies of socio-technical systems in crisis management and introduces additional vulnerabilities (Comes 2024, Gkeredakis et al. 2021). This paper aims to systematically address pertinent emerging issues and potential benefits, as well as gaps and challenges in the use of AI in health crisis management.

2. Added value and potential of AI use in crisis management

A crisis can be defined as an event or series of events that poses critical threat to the health, safety, security, or well-being of a community or other large groups of people usually over a wide area (Jeong & Yeo 2017, UNISDR 2009). Traditionally, crises have been connected to hazardous events such as natural or man-made disasters or armed conflicts, however, the United Nations (UN) identifies also vulnerable social conditions as crises. In this regard, health, energy, security, urbanization, population growth, poverty, inequalities, and climate change are accelerating drivers of crises and have to be taken into consideration within the complex field of crisis management (Jeong & Yeo 2017).

Crisis management is not only marked by a pressing urgency and complexity, but it must also be strategic and multi-faceted to mitigate, prepare, respond and recover to protect communities and infrastructure from any threat (Abid et al. 2021). These activities take place at the different stages of an occurring crisis and are suggested to be equally important for a successful crisis management process of coordinating complex relational systems and designing effective mechanisms (Kazemi 2023). Although, the common perception among researchers, decision makers, and government officials dealing with crisis management is to take proactive measures before a disaster occurs, all crises are linked to people who suffer the consequences (Abid et al. 2021) and in reality, modern crisis management places relatively larger emphasis on the response phase. Limitations of classical crisis management also stem from the urgency and need for quick decisions, which often lead to short-term rather than long-term interventions, resulting in uncoordinated and inefficient efforts. Traditional crisis management processes often struggle to cope with the high uncertainty, volatility and fast changing patterns of modern crises, which mostly rely on the cooperation of interdisciplinary actors, factors, and dimensions (Kazemi 2023). Nevertheless, the success and failure of crisis management depends on the planning and implementation of coordinated mechanisms, ranging from quick thinking to efficient resource allocation and effective decision making (Abid et al. 2021, Braten

2023). Recognizing the urgency as well as complexity of crisis management, AI has been portrayed as a potential way to support or even – at least partially – automate crisis management, given its potential to process and analyse large amounts of data to make targeted, rapid and efficient decisions (Comes 2024, Fontes 2023, Wong 2021).

While crises always entail negative effects on health and healthcare access, it also often serves as catalysts for significant breakthroughs in public health. The diverse responses to the COVID-19 crisis unleashed a technological acceleration that led to a "gold rush" of AI-related solutions, placing hope in these tools to address (health) crises (Curtis & Miailhe 2022, Fontes et al. 2023, Gkeredakis et al. 2021). Along with the ever-increasing prominence of AI, there is a multitude of definitions that aim to characterise the behaviours and functions of AI tools (Comes 2024), but none are universally accepted. In light of the thematic content of this paper, the OECD (2016) and UNCTAD (2017) definitions will be put forward: *"AI is defined as the ability of machines and systems to acquire and apply knowledge, and to carry out intelligent behaviour. This includes a variety of cognitive tasks (e.g. sensing, processing oral language, reasoning, learning, making decisions) and demonstrating an ability to move and manipulate objects accordingly. Intelligent systems use a combination of big data analytics, cloud computing, machine-to-machine communication and the Internet of Things (IoT) to operate and learn."* (Akhtar et al. 2020, p. 1)

AI and machine learning-enabled technologies act as force multipliers and are already being used in fields like medicine, transportation, robotics, science, education, military and many more. Even in crisis management, a wide range of tools are being developed to address, among other things, the following complex challenges of successfully combatting generic crises (Kazami 2023):

Response coordination and decision making: In emergency situations, where quick decisions and flexible reactions are critical, decision makers are under strain amid the urgency and high stakes of the situation, leading to biased decisions (Comes, 2024). In the response phase, timely action might even be a matter of life or death. To take a more practical example, a decision support tool being developed in the "ROADS to Health" project aims to specify the respective steps in a health crisis event and to adapt control measures to the given situation and anticipated scenarios. Advancing and further connecting the tool to an AI algorithm could automate decision making in infectious disease outbreaks to implement targeted and holistic control measures bypassing the cognitive and moral overload of decision makers in crisis management (Comes 2024). Lentz et al. (2019) argue that an AI automated decision making can provide greater transparency and replication by removing norms, biases, and politics from crisis decisions. However, an automated decision support system would shift the power from the strategic decision makers and their advisors to the programmers, developers, and designers of protocols, as well as the decisions made when developing data sets, algorithms or thresholds. Evidence-based and valid sources for the feeding of AI applications must be guaranteed to provide reliable and useful support to decision makers (see also point 2).

Situational assessment: The assessment of a crisis situation is often constrained by limited data, which are initially unavailable, uncertain, biased or conflicting, given limited access or data collection regimes (Comes 2024). In this regard, Big Data and other AI processes, such as machine learning or local and remote sensing data, decisively change traditional crisis management through the revolutionary possibilities of data processing. Consequently, AI is able to understand and identify patterns in a fast changing and complex environment, model empirical data and support the fight against infodemics (Fontes et al. 2023; Gaur et al. 2021; Quadir et al. 2016). On the other hand, AI can produce non-intentionally false information and could be used by malicious actors for disinformation. However, these processes not only facilitate the detection of, for example, infectious disease outbreaks or floods, but also advance the prediction of such events (Gaur et al. 2021). On the other hand, AI could also non-intentionally produce false informationLooking back at the COVID-

19 pandemic, clinical applications of AI such as medical imaging, image processing and even the analysis of coughing patterns were applied to help practitioners accurately identify and diagnose patients, thereby reducing the patient review time. In addition, contact tracing apps, smart watches and mobile phone data were used for disease tracking, allowing medical personnel to conduct remote follow-up, saving resources while monitoring and creating a clearer picture of COVID-19 cases (Fontes et al. 2023).

Resource allocation: Resources are frequently strained in crisis situations, necessitating effective allocation to meet urgent requirements (Kazami 2023). In this regard, AI can support programme and policy planning, by assessing the efficiency of the applicable workforce or the allocation of resources (Schwalbe & Wahl 2020). The facilitation of early warning and prediction models of severity and impact of a crisis allows for proactive adjustment of resource allocation and minimisation of resource waste (Anderson 2024).

Communication management: Clear and consistent communication is essential during crises to inform and reassure stakeholders and the public (Kazami 2023). AI-powered communication systems can use sentiment analysis and natural language processing to track public conversation and spot issues or fake information. AI technology equips policy makers with real-time insights into public sentiment and perceptions by examining social media trends, news stories, and public forums (Sumbaly et al. 2020). Virtual assistants and chatbots also provide individualized avenues of communication, sharing important information and answering questions quickly (Fontes et al. 2023).

Public perception: Understanding public perception during crises is crucial as public expectations can exert significant pressure on organizations as well as the selection of set mitigation measures. The (un)successful management and implementation of the previously mentioned points (1)-(4) shape public perception and should result in preserving public trust and credibility (Kazami 2023). To monitor and proactively respond to public sentiment during crises, machine learning and data mining of online social media networks and search engines can depict patterns of public perceptions and predominant trends (Schwalbe & Wahl, 2020). A "social listening" study on the public's confidence and hesitancy towards the COVID-19 vaccination has shown to generate valuable insights and understanding of perceptions for policy makers (Hou et al. 2021). AI is also a mean to assist in identifying and countering fake news, thus allowing for an early official position on disruptive social media posts (Sumbaly et al. 2020).

3. Gaps and challenges of AI use in crisis and pandemic management

While the multiple applications of AI imply great potential of entangling and dealing with the multifaceted complexities and demands in crisis management, this novel technological approach undoubtedly unravels its own intricate interplay of ethical, societal, technical, environmental, and legal considerations.

Decisions made during crises have far-reaching consequences, impacting the livelihoods of communities or even populations. Given this widespread effect, choices of decision makers in such situations are always explicitly or implicitly guided by values and also influenced by public opinion, frequently giving rise to moral dilemmas (Comes 2024). An automated decision support AI tool shifts the moral power out of the hands of decision makers towards the design of the protocols, data sets, algorithms, or thresholds that trigger response (Comes 2024). AI in crisis management could perpetuate or exacerbate vulnerability through *reinforcing existing biases* presented in the data used to programme and train AI algorithms (Bird et al. 2020). If biases reflect social inequalities, AI systems inadvertently amplify these disparities, disproportionately affecting marginalised communities. In crisis situations where the deployment of technological tools takes place at an unprecedented scale and speed, the unique needs and vulnerabilities of certain groups might be especially overlooked, leading to unequal access to resources or support (Tzachor et al. 2020). The shift of moral power must be recognised a considerable risk, and in any case that AI is applied in decision-making procedures, measures to counteract this dilemma must be implemented, as the moral responsibility cannot be shifted away from decision making organisations.

Regarding *technical challenges*, aspects like data quality and availability, scalability and adaptability, robustness and reliability, as well as the lack of transparency of AI tools in crisis management are of concern (Tzachor et al. 2020, Schwalbe & Wahl 2020, Fontes et al. 2023). Due to AI's complex and diverse algorithms and "black-box nature" as well as the dynamic nature of crises, these tools might struggle to provide clear explainability for their output, impeding stakeholders' traceability of steps and interventions, but more importantly the public's ability to trust and legitimize response measures taken. Considering the post-COVID-19 pandemic climate, where scientific scepticism is at its peak and scientific advancements such as genetic engineering are hampered by conspiracy beliefs, AI decision-support might make it close to impossible for the outcomes of decisions to be subject to public scrutiny (Tzachor et al. 2020).

Environmental concerns that are often not recognised in research and application relate to the large carbon footprint of AI. The entire life cycle of AI, which includes the design, training, development, validation, re-tuning, implementation, and the practical application of AI, is associated with high carbon dioxide emissions. Strubell et al. (2019) have illustrated the process of training a deep learning, natural language processing model produces roughly the same emissions as five cars produce over the average lifetime of a car (Van Wynsberghe 2021). AI systems, particularly deep learning models, require fuelling modern tensor processing hardware, resulting in high energy consumption and carbon emissions (Van Wynsberghe 2021). Furthermore, the manufacturing and disposal of hardware contributes to increased electronic waste and resource depletion, especially in low-and-middle income countries, which, combined with the high purchasing costs of the required hardware, raises ethical questions as to who has access to the tools and who bears the burden. Using AI to improve managing crises while providing the fuel to global warming consequently increasing the risk of the amount and severity crises, creates a potential paradox. A cost-benefit analysis of these applications, in light of the potential enhancements in crisis management by AI, must take these aspects into account.

Legal concerns and challenges related to AI include lack of clarity on liability, inadequate regulations and potential non-compliance with privacy and data protection rights (AI Act 2024). During the COVID-19 pandemic, the use of contact tracing apps and public mass surveillance in publicly accessible spaces endangered individual liberties and democratic values in terms of privacy intrusion (Fontes et al. 2023). The increasing ability to process and analyse data and the expanding use of facial and emotional recognition systems in public spaces, combined with inadequate regulation, opens the back door to exploitation (Barrett et al. 2019, McClain-Nyu 2023). These concerns, among other issues, arise in part from the complexity and rapid advancement of AI technology, which often outpaces existing regulatory frameworks. In the KRISAN project, these shortcomings were studied in the design of the smart AI emergency call and inquiry system for crisis situations (Hagendorn et al. 2024). In holistic and applicability-focused research, the inclusion of these topics from the very beginning in the design phase is essential, as also envisaged by the ROADS approach and the security research project KRISAN (FFG, 2024).The multitude of existing and arising challenges regarding the development and application of AI are uncertain as this development is new and its implications are not yet well understood, yet these advancements promise to be as pervasive as the internet or smartphones (Van Wynsberghe 2021). Questions are being raised about *the impact of autonomous* *technology on our shared values that govern social behaviour and politics* (Fontes et al. 2023). In particular, when rapid action is required, doubts have been raised about the meanings these tools will acquire and the direction they will take society. In response, organizations such as the European Union (EU) High-Level Expert Group on AI (AI HLEG) and the Institute of Electrical and Electronics Engineers (IEEE) have published ethical principles and values that should be followed along the design and deployment of such tools (van de Poel 2020). There are many scholars who recognize the pressing concerns of AI but do not want to block the path of this revolutionary technology. However, the frameworks proposed to address these pressing issues are insufficient if institutional structures lack the appropriate regulatory and ethical frameworks, as well as benchmarking standards to deploy and develop AI in a coordinated manner (Schwalbe & Wahl, 2020).

4. Potential benefits, future perspective, and directions: AI in the context of the project ROADS to Health

Clearly, the variety of potential and groundbreaking solutions raise hope to entangle or support the complex issues of crisis management. On the other hand, many difficult and unanswered questions create hesitation when it comes to dedicating time and resources to the further development and implementation of AI tools. There are many aspects to consider when designing a usable and applicable crisis management AI tool. In the following, the ROADS to Health research project will be used as a practical example to identify specific features, principles, and practices that should be incorporated and followed when designing an AI tool for crisis management. The initial idea and the subsequent development of the ROADS to Health project has its roots in the potential for optimization in national and global mitigation strategies that became evident during the COVID-19 pandemic. In Austria, control measures were at first set up without a validated background, exacerbating uncertainty and mistrust among the population. In response, the ROADS to Health project aims to develop a technological decision support tool that facilitates holistic and enhanced pandemic management from a societal perspective. The designed tool aims at offering decision makers a selection of potential targeted measures in a health crisis event and provides the opportunity to adapt interventions to the given situation and anticipated scenarios. Strategic parameters and goals are selected by decision-makers with a variety of targets that can be chosen according to the current needs. The matching system is intended to suggest possible measures for a specific point in time, given the knowledge and availability of the necessary resources, with regard to the officially set strategy of decision-makers at political and operational level.

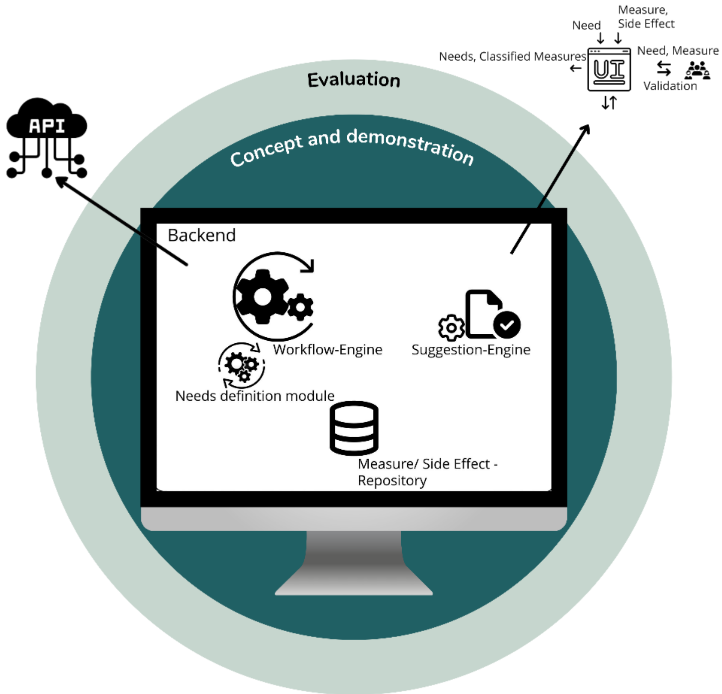


Figure 1. Concept and demonstration of the ROADS to Health decision support system

In accordance with an extensive assessment of vulnerabilities and critical infrastructure, emphasis is placed on counteracting systemic biases to protect these communities and locations, enabling usability and applicability of the decision support system. As mentioned above, AI systems can only provide proper support if fed with valid, reliable and well-structured data and background information, so the systematic and evidenced-based data reassessment in the ROADS to Health project can lay the foundation for enabling AI-assisted crisis decision-making. *Figure 1* shows the architecture of the ROADS decision support system. The backend makes up the largest part of the system and consists of several components: Workflow engine, which also contains a module for defining needs, a suggestion engine and a repository for measures and side effects. The workflow engine component is responsible for defining the necessary workflows (e.g. definition of goals/needs, definition of measures). The suggestion engine component is responsible for matching the needs with the measures that can be implemented to satisfy the needs. Finally, the measure/side-effect repository is used to store all relevant information in the system. In addition, APIs (Application Programming Interfaces) are defined in the backend, which enable the exchange of information (in both directions, sending and receiving) with other systems. The UI component (user interface) is used to enable users to communicate with the ROADS demonstration system. Different types of user roles (different levels of access) are defined. Depending on the role, users can access the stored information (basic user); define needs, measures and side effects or validate the information in the system (advanced user); or make system changes (admin user).

The evolution and further development of a tool emerging from the holistic approach of the ROADS to Health project promises to be connected to an AI algorithm, which enables processing large amounts of data and provides a range of even more targeted response measures and anticipated and unforeseen scenarios. In light of tackling the aforementioned challenges of AI, careful consideration of ethical, legal, ecological and local contexts is required. The design process should prioritise the needs, preferences and values of societal contexts, networks and end-users including decision makers and healthcare facilities but also involve all stakeholders of health crisis management, ranging from affected communities to first responders. The concept of human-centred AI sheds light on the design principles that need to be followed in order to ensure human control in complex networks interacting with AI (Comes 2024). Shneiderman (2020) claims that AI must *"amplify, augment, and enhance human performance in ways that make systems reliable, safe, and trustworthy"* to include the guiding principles of crisis management as well as the principles for digital development, to protect people with vulnerabilities (Comes 2024). These principles provide direction for best practices in the development of digital health technologies (Schwalbe & Wahl 2020). Moreover, the design and operation of the AI algorithm must ensure transparency. This includes providing clear documentation and explanations of the functioning and programming of the algorithm, the data and sources it relies on, and the rationale behind the set of suggestions presented to decision makers to make informed decisions. The design of the AI tool must take into account the location of use, as well as regional norms, values and cultural sensitivities. Recognizing and respecting the local context is crucial for ensuring that the tool is culturally appropriate and sensitive to the unique circumstances of different communities but also accepted and adhered to by the concerned population. However, putting forward fundamental principles leaves many questions unanswered about the concrete design. Van de Poel (2020) proposes *"an account of value embodiment in technology which can help in assessing whether designed AI systems indeed embody a range of moral values",* which should be acknowledged in testing and evaluation procedures (Comes 2024).

Many challenges remain in conceptualising a solid and usable AI application for crisis management, which must also account for sustainability issues and the practical transfer of guiding principles into technology. However, as regulatory and legal frameworks evolve, interdisciplinary teams and collaborative efforts will foster a holistic understanding and identify possible solutions to the remaining challenges. By incorporating these considerations into the design of the AI-driven decision support tool, we must strive to ultimately create a solution that is ethically sound, sustainable, and sensitive to the needs and contexts of crisis management.

5. Conclusion

The emergence of AI-driven technology in crisis management presents both unprecedented opportunities and profound challenges. The potential of AI to enhance response coordination, situation assessment, resource allocation, communication management, and public perception offers a promising avenue for more effective crisis mitigation and management. However, this advancement also unravels intricate ethical, technical, environmental, and legal considerations that demand careful navigation. The deployment of AI tools in crisis management introduces profound shifts in decision making dynamics, potentially exacerbating existing biases and inequalities. Technical challenges, such as data quality and explainability, further complicate the adoption of AI tools, raising concerns about public trust and accountability. Environmental implications, including the high carbon footprint of AI systems, add another layer of complexity, highlighting the need for sustainable development practices. Legal and regulatory frameworks that are crucial for the development of an AI-based tool lag behind the rapid advancement of AI technology, leaving critical issues such as liability and privacy unresolved. As organizations grapple with these challenges, ethical principles and values must guide the design and deployment of AI tools in crisis management.

Despite these hurdles, the potential benefits of AI in crisis management are undeniable. By prioritizing human-centered design principles, ensuring transparency, and respecting cultural contexts, AI technology can empower decision makers and communities to effectively navigate the complexity of crises. However, addressing these challenges requires collaborative efforts from policymakers, researchers, and stakeholders to develop robust regulatory frameworks and ethical guidelines. In moving forward, it is imperative to balance the transformative potential of AI with commitment to safeguarding human values, rights, and dignity. By embracing a holistic approach that integrates ethical considerations, technical expertise, and societal values, we can harness the full potential of AI to build more resilient and adaptive crisis management systems for the future.

Acknowledgement

The research leading to these results has received funding from KIRAS Cooperative R&D Projects 2021 with project number FO999899442.

References

Abid, S. K., Sulaiman, N., Chan, S. W., Nazir, U., Abid, M., Han, H., Ariza-Montes, A. & Vega-Muñoz, A. (2021). Toward an Integrated Disaster Management Approach: How Artificial Intelligence Can Boost Disaster Management. Sustainability (Basel), 13(22), 12560. https://doi.org/10.3390/su132212560

AI Act: European Parliament 'Corrigendum' of 16th April 2024. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf retrieved 26.04.4024

Akhtar, S., Mikic, M., Wong, J., Wang, T. & Ngernlim, P. (2020). Artificial Intelligence in Asia and the Pacific. UNESCAP. https://www.unescap.org/sites/default/files/ESCAP_Artificial_Intelligence.pdf

Anderson, J. (2024, 2. Januar). Predictive Analytics for Crisis Management: A Paradigm Shift. Bryghtpath. https://bryghtpath.com/predictive-analytics-for-crisis-management-a-paradigm-shift/

Barrett, L. F., Adolphs, R., Marsella, S., MartíNez, A. M. & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. Psychological Science in The Public Interest, 20(1), 1–68. https://doi.org/10.1177/1529100619832930

Bird, E., Fox-Skelly, J., Jenner, N., Larbey, R., Weitkamp, E., Winfield, A., Panel for the Future of Science and Technology, European Parliamentary Research Service & Scientific Foresight Unit (STOA). (2020). The ethics of artificial intelligence: Issues and initiatives. In STOA | Panel For The Future Of Science And Technology (report PE 634.452). https://doi.org/10.2861/6644

Braten, O. A. (2023). The Transformation of Crisis Management: How AI is Revolutionizing Best Practices (14.10.2023). https://www.linkedin.com/pulse/transformation-crisis-management-how-ai-best-practices-braten

Comes, T. (2024). AI for crisis decisions. Ethics And Information Technology, 26(1). https://doi.org/10.1007/s10676-024-09750-0

Curtis, S. & Miailhe, N. (2022, 22. April). Three takeaways from classifying AI systems designed to fight the COVID-19 pandemic - OECD.AI. OECD.AI. retrieved 05.04.2024, https://oecd.ai/en/wonk/three-takeaways-classifying-ai-systems-designed-to-fight-covid-19

FFG (2024). KRISAN. retrieved 25.04.2024, https://projekte.ffg.at/projekt/4757919

Fontes, A. C., Corrigan, C. C. & Lütge, C. (2023). Governing AI during a pandemic crisis: Initiatives at the EU level. Technology in Society, 72, 102204. https://doi.org/10.1016/j.techsoc.2023.102204

Gaur, L., Singh, G. & Agarwal, V. (2021). Leveraging Artificial Intelligence Tools to Combat the COVID-19 Crisis. In Communications in computer and information science (Print) (S. 321–328). https://doi.org/10.1007/978-981-16-1480-4_28

Gkeredakis, M., Lifshitz-Assaf, H. & Barrett, M. (2021). Crisis as opportunity, disruption and exposure: Exploring emergent responses to crisis through digital technology. Information And Organization, 31(1), 100344. https://doi.org/10.1016/j.infoandorg.2021.100344

Hagendorn, M., Rainer, K., Kundratitz, V., Leidwein, A., Fuhrmann, F. (2024) AI-based voice dialogue system for enhancing crisis management in the KRISAN project [Manuscript submitted for publication]

Hou, Z., Tong, Y., Du, F., Lu, L., Zhao, S., Yu, K., Piatek, S. J., Larson, H. J. & Lin, L. (2021). Assessing COVID-19 Vaccine Hesitancy, Confidence, and Public Engagement: A Global Social Listening Study. JMIR. Journal Of Medical Internet Research/Journal Of Medical Internet Research, 23(6), e27632. https://doi.org/10.2196/27632

Jeong, B. G. & Yeo, J. (2017). United Nations and Crisis Management. In Springer eBooks (S. 1–8). https://doi.org/10.1007/978-3-319-31816-5_850-1

Kazemi, A. (2023, 31. Oktober). Futurium | European AI Alliance - AI for Crisis Management: Impacts, Challenges, Best Practices. European Union. retrieved 05.04.2024, https://futurium.ec.europa.eu/en/european-ai-alliance/forum-discussion/ai-crisis-management-impacts-challenges-best-practices

Kee Wong, Y. (2021). Dealing crisis management using AI. International Journal of Computer Science, Engineering and Applications, 11(05), 15–22. https://doi.org/10.5121/ijcsea.2021.11502

Lentz, E. C., Michelson, H., Baylis, K., & Zhou, Y. (2019). A datadriven approach improves food insecurity crisis prediction. World Development, 122, 399–409.

McClain-Nyu, J. (2023). There's reason to worry about AI reading emotions. Futurity (19.12.2023). https://www.futurity.org/artificial-intelligence-emotions-3010572/

OECD. (2016). OECD Science, Technology and Innovation Outlook 2016. In OECD science, technology and innovation outlook (Online). https://doi.org/10.1787/sti_in_outlook-2016-en

Qadir, J., Ali, A., Rasool, R. U., Andrej Zwitter, Arjuna Sathiaseelan & Jon Crowcroft. (2016). Crisis Analytics: big data driven crisis response. In arXiv:1602.07813v1 [cs.CY] 25 Feb 2016. https://arxiv.org/pdf/1602.07813.pdf

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, safe & Trustworthy. In arXiv.org. https://arxiv.org/ftp/arxiv/papers/2002/2002.04087.pdf

Schwalbe, N. & Wahl, B. (2020). Artificial intelligence and the future of global health. Lancet (British Edition), 395(10236), 1579–1586. https://doi.org/10.1016/s0140-6736(20)30226-9

Singh, P. K., Veselov, G., Vyatkin, V., Pljonkin, A., Dodero, J. M., & Kumar, Y. (Eds.). (2021). Futuristic Trends in Network and Communication Technologies. (2021). In Communications in computer and information science. https://doi.org/10.1007/978-981-16-1480-4

Strubell, E., Ganesh, A. & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP (05.06.2019). arXiv.org. https://arxiv.org/abs/1906.02243

Sumbaly, R., Miller, M., Shah, H., Xie, Y., Chang Culatana, S., Khatkevich, T., Luo, E., Strauss, E., Szilvasy, G., Puri, M., Manadhata, P., Graham, B., Douze, M., Yalniz, Z. & Jegou, H. (2019). Using AI to detect COVID-19 misinformation and exploitative content. Meta. retrieved 25.04.2024, von https://ai.meta.com/blog/using-ai-to-detect-covid-19-misinformation-and-exploitative-content/

Tzachor, A., Whittlestone, J., Sundaram, L. & Héigeartaigh, S. Ó. (2020). Artificial intelligence in a crisis needs ethics with urgency. Nature Machine Intelligence, 2(7), 365–366. https://doi.org/10.1038/s42256-020-0195-0

UNCTAD (2017). Information Economy Report 2017. UNCTAD. http://unctad.org/en/pages/PublicationWebflyer.aspx?publicationid=1872

Van De Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. Minds And Machines, 30(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4

Van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. AI And Ethics, 1(3), 213–218. https://doi.org/10.1007/s43681-021-00043-6

AUTOMATING AI-SUPPORTED INFORMATION EXTRACTION IN NATURAL HAZARDS RECONNAISSANCE

Refiz Duro, Axel Weißenfeld, Medina Andresel, Veronika Siska, Dražen Ignjatović

AIT Austrian Institute of Technology GmbH
refiz.duro@ait.ac.at

Christoph Singewald

Syncpoint GmbH
christoph.singewald@syncpoint.io

DOI: 10.35011/IDIMT-2024-23

Keywords

Natural hazard reconnaissance; disaster response; multimodal data; AI

Abstract

Hazardous events and corresponding responses require the exploitation of information from reports in various modes (text, image, video, audio) coming from responders, affected population and officials. In addition to the underlying need for automated processing of such reports and information, we explore the potential of implementing AI services for information extraction and knowledge management. We demonstrate that we can transfer the practices and technology from the military reconnaissance to civilian applications, while integrating the technological advances from the AI-based object detection and Large Language Models, and providing a simple way to automate the processing of multimodal data to generate intelligence to support decision making.

1. Introduction

Accurate, comprehensive and timely information is essential for decision making in response to hazardous events (e.g., forest fires, earthquakes; Wolbers et al., 2021). It is necessary that the information is used to produce intelligence that can be shared by all involved stakeholders, i.e., authorities, crisis managers, first responders and other interest groups, to enable an effective sharing and management of resources and efforts. Different types of information sources are used to gain an overview of the situation. These range from open source data such as news articles and social media posts to expert reports and sensor data from various devices (Saroj & Pal, 2020; Velegrakis et al., 2024). The literature shows that ”information” is still the most valuable asset in disaster response (Seppänen & Virrantaus, 2015), and collecting it in the field can be done by reconnaissance-type of activities. It has always been difficult to extract relevant, reliable information from a large collection of heterogeneous, diffuse and ambiguous information, and it **often** requires human intelligence and experience to achieve relevant quality (Holzinger, 2021). Advances in te**c**hnology, particularly in the field of artificial intelligence, are enabling new approaches to this type of problems (e.g., data fusion, Large Language Models; Caufield et al., 2023; Li et al., 2022).

Reconnaissance is a set of activities designed to collect data and information through observations and sensors in order to provide insights to support decision making. It is usually associated with the military domain, where human scouts and advanced detection and surveillance technologies are exploited to produce information about adversary forces, the terrain and meteorology of the relevant area (Stilwell, 2020). Military reconnaissance is part of the Intelligence Cycle, a standardised process for gathering and processing information (NATO, 2016). It consists of four steps: 1) defining intelligence requirements, 2) collecting data and information through available means, 3) processing the collected data and information and producing intelligence products, and 4) disseminating the intelligence products. The last step closes the circle by assessing whether the initial requirements have been met, i.e., whether those responsible have received sufficient information to make decisions. Similar approaches to information processing for decision support can be found in civil protection. For example, the Austrian Guidelines for Leadership in Disaster Operations contain a very similar structure (SKKM Strategie 2020, 2009). The similarity can be traced back to the traditional and increasingly civil-military cooperation in disasters and crises, which is also the case in many other nations (Bollen & Kalkman, 2022).

Reconnaissance in the civil domain has mainly focused on the response and recovery phases of the Disaster Management Cycle[1] (not to be confused with the Intelligence Cycle). While the response phase is concerned with actions taken at the onset and during a disaster event with the aim to save lives and minimise other types of loss, the recovery phase is concerned with restoring pre-disaster living conditions. An example of this is the assessment of structural damage following the Haiti earthquake in 2021. Working with local activists and the Structural Extreme Events Reconnaissance Team, data and information was collected on more than 2.000 buildings. The data came in the form of images, reports and video recordings, which were then analysed to assess the impact of the earthquake and provide the information base for response and recovery (Whitworth et al., 2022). A similar approach was used to recon and asses the impact of the MW 6.9 earthquake that struck the Aegean coast in 2020 (Aktas & So, 2022). The team's objectives included "combining field and remote survey strategies [...] for damage assessment in buildings, critical infrastructure and geo-technical structures, [...] and investigating the extent to which other data sources can be used for remote reconnaissance in support of field work." These other data sources included geo-referenced images of buildings and infrastructure, YouTube videos of affected areas, both produced by individuals not part of the response team, and official channels such as relevant ministries and professional organisations. These two examples align well with the challenges, needs and success stories related to natural hazards reconnaissance, as collected by (Wartman et al., 2020). These include topics such as data fusion of multimodal data streams, handling of different data types/formats, time intensive data processing, reduction of large amounts of data into usable knowledge, analysis of disparate data sources (ground motion, site, building damage, building geometry and reinforcement), data gaps, etc. Addressing these gaps is by no means a new practice in disaster management, so that we can find research and applications related to data collection and management (Berman et al., 2020; Mohd Daud et al., 2022), machine learning for data fusion (Ochoa & Comes, 2021), and deep learning for disaster response (Algiriyage et al., 2021), to name a few. We are building on this existing knowledge to explore the potential for transferring/mapping the practices

[1] https://disaster-management.piarc.org/en/management-disaster-management/disaster-management-cycle

and technologies from military reconnaissance to civilian applications. We are targeting the domain of disaster management, exploring the integration of technological advances in AI and the potential of automating the processing of multimodal data. Our focus here is on the technological aspects, but it is necessary to keep in mind the framework requirements for the repurposing of military technology and applications for the civilian domain, e.g. ethics, security, safety and regulatory aspects. (e.g, Gospodinova & Dejond, 2021; Mandic & Punda, 2017).

The paper is structured as follows. After an introduction to the subject of natural hazards reconnaissance, its goals, uses, benefits and challenges, we continue in Section 2 with providing the context of our research, available data, implemented methods and a demonstration. Section 3 discusses and concludes the results, open questions and the application of AI.

2. Implementation

2.1 Scenario and data

Earthquake events in modern times are often very well documented. Documentation comes in the form of social media videos and photos, surveillance and security camera footage, satellite and UAV images and videos, audio recordings, written reports, as well as scientific and official records from seismic and tectonic monitoring. For our small study, we selected the 2023 Turkey–Syria earthquake, which had the magnitude of M_W 7.8 and caused massive destruction of infrastructure and loss of life[2] (Dal Zilio & Ampuero, 2023). The massive rescue and relief operations required timely and accurate information.

To demonstrate the potential of automatic data processing to create intelligence to support decision making, we searched online sources for videos from the earthquake event (e.g., YouTube). We had three criteria for the selection: 1) the audio and visual content of a video must be informative and relevant to a potential rescue operation; 2) the language must be local (i.e., in our case Turkish or Arabic); 3) the language must not be scripted (e.g., from the news), but natural. As for the length, we did not have any specific criteria, but the longer the video the more potential for more information. The good quality of the video was also not a requirement. Although the event is widely covered, the authors found it difficult to find videos observing the three criteria. A simple search for ”Turkey-Syria Earthquake of 2023”, or even more complex and targeted searches using different versions of ”turkey earthquake live footage”, including their Turkish translations, resulted mostly in news reports. We did, however, find one video that fits the mentioned criteria. It shows a young man trapped in the rubble of a collapsed building using his social media channel to plead for a rescue[3]. He succeeds and the situation has a happy ending, as described on the video page: ”Boran Kubat was trapped with his mother and two uncles when the second earthquake hit southern Turkey. Rescuers were able to locate Kubat’s family after he detailed their exact location.” The video has subtitles, presumably provided by professional translators for The Guardian, the news agency responsible for editing the video, and a transcription provided by YouTube (Figure 1).

[2] https://en.wikipedia.org/wiki/2023_Turkey%E2%80%93Syria_earthquakes; accessed on 28.03.2024.
[3] https://www.youtube.com/watch?v=5isWyLU7zOU

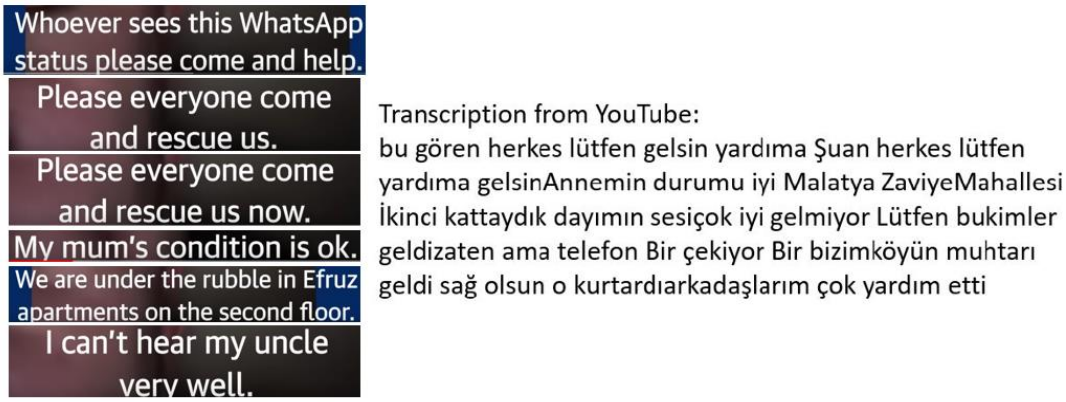


Figure 1. Screenshots of the subtitles, audio transcription provided by YouTube.

2.2 Experiment

In order to investigate the potential for transferring the practices and technologies from the military reconnaissance to civilian applications, and in particular the potential for automating the processing of multimodal data and integrating technological advances from the field of AI, we have set up an experiment based on the ongoing developments in the BOOST project (*BOOST*, 2022). The project aims to develop a flexible and adaptive prototype for the automated analysis and classification of heterogeneous and unstructured data from different sources using AI methods, and is as such suitable for our investigation. The project takes a practical approach, where a prototype consists of a set of data processing pipelines that are defined in order to produce intelligence products for decision support from the collected data. A processing pipeline includes the core functionalities such as data integration, data fusion, data and metadata management, display (e.g., maps), as well as advanced functionalities for image, text, video and audio processing. These are all packaged in modules using the Docker technology. For example, images and videos can be enhanced by applying modules for editing (e.g., changing contrast and resolution), or mined for information using functionalities in modules for, e.g., object detection or named entity recognition. The user, an analyst, only needs to select the modules that are suitable for her/his goals. An example of a predefined pipeline is shown in Figure 2, where a data input in form of a video is processed and analysed for information on geographic location, scene, objects and entities and related features (e.g., colour, licence plate numbers). For each new data input, regardless of their modularity (e.g., text vs. image), similar pipelines are activated and the extracted information is stored in the database. Further analysis is performed as the relevant information for intelligence lies in the entities and relationships in the data (e.g., people, places/locations, organisations). This enables, among other things, the creation of an operational picture for decision support, the discovery of networks of entities (organisations, people), and the recognition of event patterns.

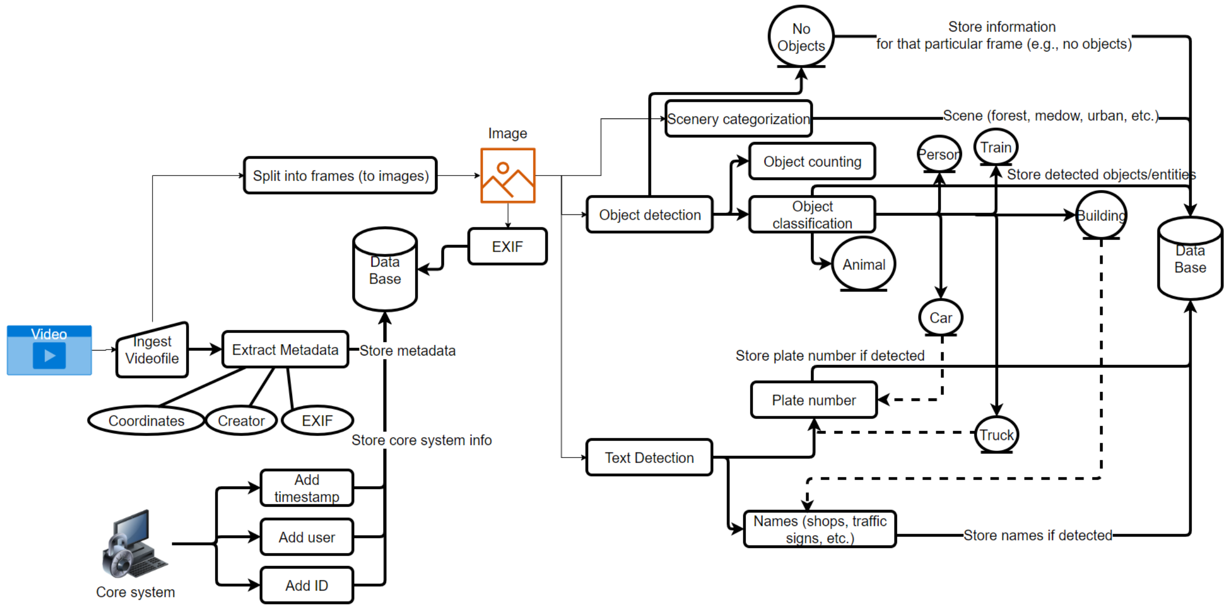


Figure 2. Once uploaded (ingested), the video undergoes a number of processing steps that serve to extract useful information and create intelligence for decision support

For the current study, we use the video described in Section 2.1, which contains all 4 types of modularity, i.e., video, image, audio and text, although the modularities become apparent at different stages of data processing. We demonstrate how the technology advancements can be used for efficient data processing and the production of intelligence. Figure 3 shows the selected modules that we focus on, which are part of a more complex pipeline. It is initialised by uploading a video file via the user interface, and is followed by the data preparation - splitting to images, audio extraction, transcription to text and its translation - and information extraction steps - object detection and Natural Language Processing (NLP) using Large Language Models (LLM).

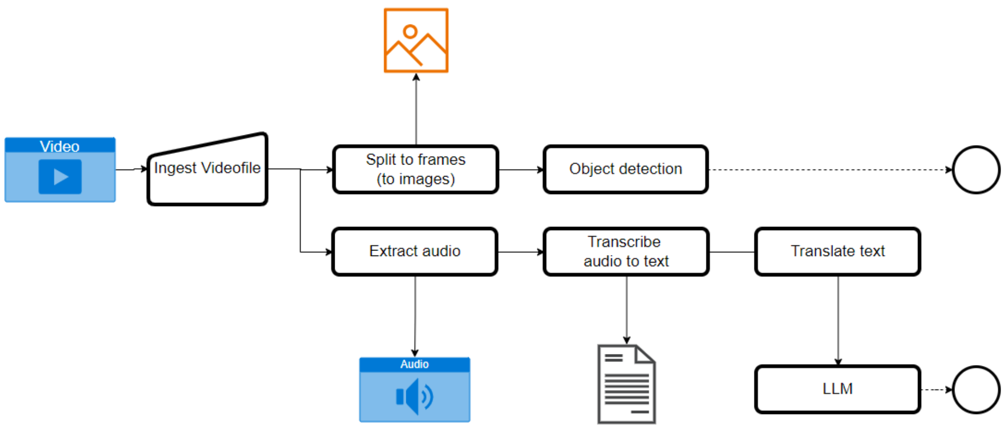


Figure 3. Selected modules as part of a data processing and analysis pipeline. An uploaded video is processed to provide images for object detection and an audio file for transcription, translation and advanced content analysis using Large Language Models.

2.2.1 Video to Image Frames to Object Detection

The module is rather a simple one, but very important as it enables the automatic data processing chain (pipeline). We exploit the fact that a video is a sequence of images (frames) converted into a digital format and displayed on a screen at a speed sufficient to create the illusion of continuity. We do this by providing a module that extracts all or only relevant frames from a video, allowing for a more targeted and potentially efficient approach to further processing. The logic behind extracting only relevant frames is related to the placement of valuable information in different parts of a video. For short videos, it is possible to quickly analyse them and extract the relevant information, i.e., to conclude whether there is any valuable information in the video, even without any advanced processing tools. However, if the video is of longer duration, it can be time- and resource consuming to skim through an entire video just to find out that the valuable information is only in the 35th minute of the video. The method we have implemented is the extraction of the so-called ”key-frames”, which are frames that contain important or critical information. The key-frames usually contain significant or new information, due to the recording device (e.g., camera) being moved to a new location, or some colour changes have been introduced in the new frame. For key-frame extraction we used the FFMPEG framework, which provides libraries for recording, converting and streaming audio and video[4]. The next module in the process takes the key-frames and applies AI-based methods to extract information. We applied YOLOv7 multi-class object recognition algorithm, which is one of the most widely used approaches for detecting objects in images (Redmon et al., 2016; Wang et al., 2022). In our case the algorithm easily recognises the ”object” in the video as a person. We discuss this simple, but potentially powerful result in other contexts (or video lengths) in Section 3.

2.2.2 Audio Transcription and Translation

Within the analysis pipeline, the whisper model (Radford et al., 2022) is employed for transcription, supporting almost 60 languages including Afrikaans, Arabic, Russian and Chinese. Whisper is an open source software and several pre-trained models are available, varying from 39M to 1550M parameters, are available. For our video, the Table 1 shows the output of the transcription step using one of these models, which is very similar to the transcription provided by the YouTube’s machine learning algorithms.

Since we are considering situations where international rescue teams may be involved, we assume that the target language may also be different from the original, transcribed language. Therefore, we need to translate the text using an additional processing step. We use three ”translators”, i.e., the pre-trained seamless model (Communication et al., 2023), which is able to translate (text-to-text) for nearly 100 languages, the open source ChatGPT 3.5 (https://chat.openai.com), and the whisper model, which can also be used as a translator, but with the support for fewer languages. The results are listed in Table 2, and we can already see some differences with respect to the baseline, i.e., the translation provided by The Guardian team.

Table 1. Audio transcription outputs.

Source	Transcription
YouTube	WhatsApp durumu hall etmez bu g¨oren herkes l¨utfen gelsin yardıma S¸u an herkes l¨utfen yardıma gelsin Annemin durumu iyi Malatya Zaviye Mahallesi ˙Ikinci kattaydık dayımın sesi ¸cok iyi gelmiyor L¨utfen.

[4] https://ffmpeg.org

Source	Transcription
whisper (large model with 1550M parameters)	”Whatsapp durum anne Bu g¨oren herkes l¨utfen gelsin yardıma S¸u an herkes l¨utfen yardıma gelsin. Annemin durumu iyi. Malatya Zaviye Mahallesi. ˙Ikinci kattaydık. Dayımın sesi ¸cok iyi gelmiyor. L¨utfen.”

Table 2. Outputs from the professional translation and LLM models applied on the transcribed text.

Model	Translation
Professional translation (as provided by The Guardian)	Whoever sees this WhatsUpp status, please come and help. Please everyone come and rescue us. Please everyone come and rescue us now. My mom’s condition is ok. We are under the rubble in Efruz apartments on the second floor. I can’t hear my uncle very well.
whisper (large model with 1550M parameters)	What’s up, mom? Everyone who sees this, please come to help. Everyone right now, please come to help. My mother’s situation is good. Malatya Zaviye neighborhood. We were on the second floor, my uncle’s voice is not very good.
ChatGPT 3.5	Whatsapp status mom Anyone who sees this, please come and help. Right now, everyone, please come and help. My mother is fine. Malatya, Zaviye Neighborhood. We were on the second floor. My uncle’s voice is not coming through very well.
Seamless (SeamlessM4TLarge-v2 model with 2.3B parameters)	WhatsApp situation mom everyone who saw this please come to the rescue now everyone please come to the rescue. My mother is fine, we were on the second floor of the Malatya Zaviye neighborhood, my uncle’s voice is not very good.

To assess the quality of the transcription and translation, we compute the semantic similarity between given texts. Semantic similarity comes from Natural Language Processing (NLP) and is a widely studied topic. It can be understood as the degree of likeness of the meaning between two given texts (Navigli & Martelli, 2019). This implies that the semantic similarity analysis goes beyond, e.g., word statistics and syntax to a more advanced understanding of the content. It is a simple approach that can be further used as a condition within the data processing pipeline. For semantic similarity calculations, we apply text (https://r-text.org/), which is ”an R-package for analysing and visualising human language using transformers from Natural Language Processing and Deep Learning”(Kjell et al., 2021). In particular, we use the *textSimilarity()* function, which computes the semantic similarity between two text embeddings and assigns it a score between 0 and 1, i.e., no similarity to identical meaning. For the transcription output of the audio file using the whisper model, and its comparison with the YouTube transcription (see Table 1), we obtain a similarity score of 0.94. The translations and the ”original text” (i.e., the professional) from the Table 2 are given as a string and we compare the original text with these model-based translations. The whisper and ChatGPT model outputs score 0.941 and 0.94 respectively, while seamless scores slightly worse with a similarity score of 0.91. With these results, we can assume that both the transcription and translation outputs (any model) can be used for further processing and information extraction.

2.2.3 LLM on the translated text

Large Language Models can be used as a zero-shot approach for extracting relevant information from a text (Caufield et al., 2023; Wei et al., 2022). They can extract relevant entities and their relations from any text, and then link this information together by creating a knowledge graph using data coming from multimodal data sources. The first step is to extract triplets of the form (subject, predicate, object) where subject represents the subject entity, object the object entity, and predicate represents the relation between them. The triplets can be further integrated into a knowledge base to create intelligence to support decision making.

We applied the out-of-the-box *Langchain* solution using GraphIndexCreator5 and the LLM model vicuna-7b-v1.5-16k6 and obtained the results presented in Table 3. We used the translations provided in Table 2 as the input. We can see that the results are not optimal in the sense of being accurate and thus potentially helpful in a crisis scenario. We also see, however, that some central information could be extracted, as in the output of the seamless model: (”mom”,”Malatya Zaviye neighborhood”,”was on the second floor”). The triple representation of the extracted information should certainly be improved in order to be really useful. Furthermore, more sophisticated solutions such as prompt engineering, retrieval augmented generation or fine-tuning of the LLM can be applied to obtain more accurate triple representations.

3. Discussion and Conclusion

Decision support intelligence plays a crucial role in both civil and military applications. The transfer of technologies and their applications between the two domains is to be expected, and thus we see civil-military cooperation in response to disasters and crises. For such events, the information extraction and creation of the intelligence is critical, with the emphasis on the accuracy of the intelligence and its timeliness (i.e., it needs to be sufficiently fast). In this context, we investigated the potential of transferring/mapping the practices and technologies of military reconnaissance to civilian applications, with the aim of extracting information from multimodal data and automating data processing, taking advantage of technological advances in the field of artificial intelligence. Automated processing of multimodal data is enabled by a combination of an architecture and a modular approach that allows us to define and implement a set of data processing pipelines to produce intelligence products. Automated data processing is important as it addresses many of the challenges, needs and gaps associated with natural hazard reconnaissance and disaster response (Section 1 and (Wartman et al., 2020). The benefits of automation are primarily in saving time and resources, and in dealing with data reduction and managing disparate data sets to provide actionable intelligence. Automation has been used in emergency, disaster and response for years, but it has often been focused on one or a few aspects of the entire processing chain (e.g., Imran et al., 2020). The automation we apply in our approach is largely agnostic in terms of data modularity, content and associated information, target domains and use cases. We illustrate this in practice by using exactly the same image, audio and text processing SW modules from the military reconnaissance use case (Duro et al., 2024) in the disaster response use case demonstrated in this paper. Our demonstration pipeline begins with the selection of information-relevant keyframes from the input video and the extraction of audio. This is followed by the advanced AI-based SW modules for information processing and extraction. The key-frames functionality can be easily exchanged with the module that already applies real-time object detection directly on the video, depending on what the end-users want to do (e.g., visualize, post-process the information or something else). Although there are already great benefits to applying AI-based object detection in images (as well as challenges, e.g., (Vijayakumar & Vairavasundaram, 2024), for example in search and rescue operations where people in distress are detected, added value can be provided to crisis responders by incorporating services for real-time intelligent video analysis for people counting, gender and age estimation (Abidi & Filali, 2023), and services for assessing a person’s state (e.g., lying down, not moving, or active) (Caputo et al., 2022).

In a parallel pipeline, the audio track of the video can be processed, and all relevant information can be extracted. Initially, transcription is carried out, which can then be translated into a target language. For example, English might be useful for international aid organisations. In recent years, there have been significant advances in speech processing models (Mehrish et al., 2023). It's important to note that the accuracy of these models depends heavily on the language chosen. For example, the word error rate evaluated on the Fleurs dataset ranges from 2.8% of Spanish to 50% of Bengali (Radford et al., 2022). It is therefore important to be aware of the limitations, especially as when dealing with critical decision making situations. Due to the agnostic nature of our approach (as mentioned above), any improved AI model (e.g., in terms of accuracy) made available by the scientific community can replace the models we used in the current experimental setup.

We are aware of the size of the data file analysed in this work, i.e., the video duration and the resulting text length, and the limitations this might impose. We chose it because of a) the difficulty of finding relevant video data (as discussed in the Section 2.1), and b) the need to include relevant information (i.e., information for rescue workers). We see, however, that even such small data could be useful to support decision making, although more work is needed to bring this to into a practice. This is further reinforced in the applications where the data sets are more diverse, especially in the field of military reconnaissance where the multimodal data diversity is very common, including both larger data sets and small reports, but with information-valuable content (e.g., Duro et al., 2023). Furthermore, we are aware of the limitations and challenges in collecting and using open source social media data, which range from low number of geotagged social media content, fraudulent content (e.g., footage of past events misidentified with the current event, fake information, etc.), to legal and ethical limitations in accessing or evaluating the data, to name a few. The research community is actively working to address most of these challenges, which can provide more robust and reliable decision support information (Hu et al., 2024; Imran et al., 2020; Lovari & Bowen, 2020; Ogie et al., 2022).

In conclusion, this study shows that there is a good potential to transfer the practices and technology from the military reconnaissance to civilian applications (and vice versa), while also integrating the technological advances from the field of AI. We also show that there is a potential for a simple way to automate the processing of multimodal data to generate intelligence to support decision making. However, the community has yet to fully assess the potential benefits, limitations, and possible harms of using the AI available today, especially as we address the decision-making applications where the stakes of harm to human life are likely to be higher. Until then, a combined approach with a human in or on the loop is advisable.

Acknowledgement

The research leading to these results is part of the "BOOST" project that has received funding from the FORTE programme. FORTE is a defence research, technology and innovation funding programme of the Republic of Austria, Federal Ministry of Finance (BMF). The Austrian Research Promotion Agency (FFG) has been authorised for the programme management. DeepL (www.deepl.com) has been used for editing and grammar enhancement of the paper.

References

Abidi, J., & Filali, F. (2023). Real-time AI-based inference of people gender and age in highly crowded environments. Machine Learning with Applications, 14, 100500. https://doi.org/10.1016/j.mlwa.2023.100500

Aktas, Y. D., & So, E. (2022). Editorial: Disaster Reconnaissance Missions: Is a Hybrid Approach the Way Forward? Frontiers in Built Environment, 8. https://doi.org/10.3389/fbuil.2022.954571

Algiriyage, N., Prasanna, R., Stock, K., Doyle, E. E. H., & Johnston, D. (2021). Multi-source Multimodal Data and Deep Learning for Disaster Response: A Systematic Review. SN Computer Science, 3(1), 92. https://doi.org/10.1007/s42979-021-00971-4

Berman, J. W., Wartman, J., Olsen, M., Irish, J. L., Miles, S. B., Tanner, T., Gurley, K., Lowes, L., Bostrom, A., Dafni, J., Grilliot, M., Lyda, A., & Peltier, J. (2020). Natural Hazards Reconnaissance With the NHERI RAPID Facility. Frontiers in Built Environment, 6. https://doi.org/10.3389/fbuil.2020.573067

Bollen, M., & Kalkman, J. (2022). Civil-Military Cooperation in Disaster and Emergency Response: Practices, Challenges, and Opportunities. Journal of Advanced Military Studies, 13, 79–91. https://doi.org/10.21140/mcuj.20221301004

BOOST. (2022, December 1). FORTE - Verteidigungsforschungsprogramm. https://www.forte-bmf.at/en/financed-proposals/detail/boost/

Caputo, S., Castellano, G., Greco, F., Mencar, C., Petti, N., & Vessio, G. (2022). Human Detection in Drone Images Using YOLO for Search-and-Rescue Operations. In S. Bandini, F. Gasparini, V. Mascardi, M. Palmonari, & G. Vizzari (Eds.), AIxIA 2021 – Advances in Artificial Intelligence (pp. 326–337). Springer International Publishing. https://doi.org/10.1007/978-3-031-08421-8_22

Caufield, J. H., Hegde, H., Emonet, V., Harris, N. L., Joachimiak, M. P., Matentzoglu, N., Kim, H., Moxon, S. A. T., Reese, J. T., Haendel, M. A., Robinson, P. N., & Mungall, C. J. (2023). Structured prompt interrogation and recursive extraction of semantics (SPIRES): A method for populating knowledge bases using zero-shot learning (arXiv:2304.02711). arXiv. https://doi.org/10.48550/arXiv.2304.02711

Communication, S., Barrault, L., Chung, Y.-A., Meglioli, M. C., Dale, D., Dong, N., Duppenthaler, M., Duquenne, P.-A., Ellis, B., Elsahar, H., Haaheim, J., Hoffman, J., Hwang, M.-J., Inaguma, H., Klaiber, C., Kulikov, I., Li, P., Licht, D., Maillard, J., … Williamson, M. (2023). Seamless: Multilingual Expressive and Streaming Speech Translation (arXiv:2312.05187). arXiv. https://doi.org/10.48550/arXiv.2312.05187

Dal Zilio, L., & Ampuero, J.-P. (2023). Earthquake doublet in Turkey and Syria. Communications Earth & Environment, 4(1), 1–4. https://doi.org/10.1038/s43247-023-00747-z

Duro, R., Andresel, M., Singewald, C., Siska, V., Weißenfeld, A., & Ignjatović, D. (2023). Boosting the automated Information Processing for Reconnaissance. 2023 IEEE International Workshop on Technologies for Defense and Security (TechDefense), 214–219. https://doi.org/10.1109/TechDefense59795.2023.10380937

Duro, R., Weißenfeld, A., & Singewald, C. (2024). Intelligence from Reconnaissance: Fusion of Multimodal Data. Data Fusion 2024, Venice.

Gospodinova, S., & Dejond, C. (2021). Synergies between civil, defence and space industries [Text]. European Commission. https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_652

Holzinger, A. (2021). The Next Frontier: AI We Can Really Trust. In M. Kamp, I. Koprinska, A. Bibal, T. Bouadi, B. Frénay, L. Galárraga, J. Oramas, L. Adilova, Y. Krishnamurthy, B. Kang, C. Largeron, J. Lijffijt, T. Viard, P. Welke, M. Ruocco, E. Aune, C. Gallicchio, G. Schiele, F. Pernkopf, … G. Graça (Eds.), Machine Learning and Principles and Practice of Knowledge Discovery in Databases (Vol. 1524, pp. 427–440). Springer International Publishing. https://doi.org/10.1007/978-3-030-93736-2_33

Hu, B., Mao, Z., & Zhang, Y. (2024). An overview of fake news detection: From a new perspective. Fundamental Research. https://doi.org/10.1016/j.fmre.2024.01.017

Imran, M., Ofli, F., Caragea, D., & Torralba, A. (2020). Using AI and Social Media Multimodal Content for Disaster Response and Management: Opportunities, Challenges, and Future Directions. Information Processing & Management, 57(5), 102261. https://doi.org/10.1016/j.ipm.2020.102261

Kjell, O., Giorgi, S., & Schwartz, H. A. (2021). The text-package: An R-package for Analyzing and Visualizing Human Language Using Natural Language Processing and Deep Learning. OSF. https://doi.org/10.31234/osf.io/293kt

Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., & Chanussot, J. (2022). Deep Learning in Multimodal Remote Sensing Data Fusion: A Comprehensive Review (arXiv:2205.01380). arXiv. https://doi.org/10.48550/arXiv.2205.01380

Lovari, A., & Bowen, S. A. (2020). Social media in disaster communication: A case study of strategies, barriers, and ethical implications. Journal of Public Affairs, 20(1), e1967. https://doi.org/10.1002/pa.1967

Mandic, S., & Punda, P. (2017). Defense Dual Use Technologies and Products. World Bank.

Mehrish, A., Majumder, N., Bhardwaj, R., Mihalcea, R., & Poria, S. (2023). A Review of Deep Learning Techniques for Speech Processing (arXiv:2305.00359). arXiv. https://doi.org/10.48550/arXiv.2305.00359

Mohd Daud, S. M. S., Mohd Yusof, M. Y. P., Heo, C. C., Khoo, L. S., Chainchel Singh, M. K., Mahmood, M. S., & Nawawi, H. (2022). Applications of drone in disaster management: A scoping review. Science & Justice, 62(1), 30–42. https://doi.org/10.1016/j.scijus.2021.11.002

NATO. (2016). AJP-2 Allied Joint Doctrine for Intelligence, Counter-Intelligence and Security. NATO Standardization Office (NSO).

Navigli, R., & Martelli, F. (2019). An overview of word and sense similarity. Natural Language Engineering, 25(6), 693–714. https://doi.org/10.1017/S1351324919000305

Ochoa, K. S., & Comes, T. (2021). A Machine learning approach for rapid disaster response based on multimodal data. ACM Symposium on Neural Gaze Detection, Woodstock.

Ogie, R. I., James, S., Moore, A., Dilworth, T., Amirghasemi, M., & Whittaker, J. (2022). Social media use in disaster recovery: A systematic literature review. International Journal of Disaster Risk Reduction, 70, 102783. https://doi.org/10.1016/j.ijdrr.2022.102783

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision (arXiv:2212.04356). arXiv. https://doi.org/10.48550/arXiv.2212.04356

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779–788. https://doi.org/10.1109/CVPR.2016.91

Saroj, A., & Pal, S. (2020). Use of social media in crisis management: A survey. International Journal of Disaster Risk Reduction, 48, 101584. https://doi.org/10.1016/j.ijdrr.2020.101584

Seppänen, H., & Virrantaus, K. (2015). Shared situational awareness and information quality in disaster management. Safety Science, 77, 112–122. https://doi.org/10.1016/j.ssci.2015.03.018

SKKM Strategie 2020. (2009). Staatliches Krisen- und Katastrophenschutzmanagemet.

Stilwell, A. (2020). Military Reconnaissance: The Eyes and Ears of the Army. Casemate Publishers. https://books.google.at/books?id=nnPhDwAAQBAJ

Velegrakis, A. F., Chatzistratis, D., Chalazas, T., Armaroli, C., Schiavon, E., Alves, B., Grigoriadis, D., Hasiotis, T., & Ieronymidi, E. (2024). Earth observation technologies, policies and legislation for the coastal flood risk assessment and management: A European perspective. Anthropocene Coasts, 7(1), 3. https://doi.org/10.1007/s44218-024-00037-x

Vijayakumar, A., & Vairavasundaram, S. (2024). YOLO-based Object Detection Models: A Review and its Applications. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-024-18872-y

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors (arXiv:2207.02696). arXiv. https://doi.org/10.48550/arXiv.2207.02696

Wartman, J., Berman, J., Bostrom, A., Miles, S., Olsen, M., Gurley, K., Irish, J., Lowes, L., Tanner, T., Dafni, J., Grilliot, M., Lyda, A., & Peltier, J. (2020). Research Needs, Challenges, and Strategic Approaches for Natural Hazards and Disaster Reconnaissance. Frontiers in Built Environment, 6. https://doi.org/10.3389/fbuil.2020.573068

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models (arXiv:2206.07682). arXiv. https://doi.org/10.48550/arXiv.2206.07682

Whitworth, M. R. Z., Giardina, G., Penney, C., Di Sarno, L., Adams, K., Kijewski-Correa, T., Black, J., Foroughnia, F., Macchiarulo, V., Milillo, P., Ojaghi, M., Orfeo, A., Pugliese, F., Dönmez, K., Aktas, Y. D., & Macabuag, J. (2022). Lessons for Remote Post-earthquake Reconnaissance from the 14 August 2021 Haiti Earthquake. Frontiers in Built Environment, 8. https://doi.org/10.3389/fbuil.2022.873212

Wolbers, J., Kuipers, S., & Boin, A. (2021). A systematic review of 20 years of crisis and disaster research: Trends and progress. Risk, Hazards & Crisis in Public Policy, 12(4), 374–392. https://doi.org/10.1002/rhc3.12244

THE IMPACT OF WORKFLOWS ON DECISION SUPPORT IN PANDEMIC MANAGEMENT

Georg Neubauer, Dražen Ignjatović, Gerald Schimak

AIT Austrian Institute of Technology GmbH
georg.neubauer@ait.ac.at

Georg Duscher, Lisa Winkelmayer, Karin Rainer, Alois Leidwein

AGES Austrian Agency for Health and Food Safety
karin.rainer@ages.at

Gerald Lichtenegger

Office of the Regional Government of Styria
gerald.lichtenegger@stmk.gv.at

DOI: 10.35011/IDIMT-2024-35

Keywords

Pandemic, West Nile Virus, BSL4, trials, processes, decision support

Abstract

According to the WHO zoonoses are a worldwide major public health challenge due to the closeness to animals in human made environments as well as in nature. The management of pandemics caused by zoonotic pathogens leads to increased complexity because of the need to manage both the human medical as well as veterinarian perspective of such a disease. Both human and animal cases may need to be handled involving different actors with diverging capabilities. Within the project MOBILISE a mobile laboratory for pathogen risk group 4 with an integrated decision support system is developed. The decision support system is validated in 4 national trials, the scenarios of these trials are taking the national processes for the management of zoonoses into account. In this paper we describe the federal workflows for the management of zoonoses in Austria and discuss their impact of the development of the decision support system of MOBILISE.

1. Introduction

According to the WHO (WHO, 2020) zoonosis are diseases that can be naturally transmitted from vertebrates to humans. The respective pathogens can be bacteria, viruses, parasites and may be transmitted due to direct contact, food, water or the environment. WHO estimates globally one billion cases and millions of deaths per year due to zoonotic diseases such as West Nile Virus (WNV) or Yellow Fever (WHO, 2024). In that context, arboviruses are of utmost relevance because they can be transmitted by arthropod vectors such as ticks, mosquitoes or sandflies between human and animals. Examples of pathogens transmitted by such vectors are Crimean-Congo haemorrhagic fever virus (CCHFV) or the Chikungunya virus. In that light it is recognized that detection capabilities of zoonotic diseases in human, animals and the environment need to be improved, use of mobile laboratories in combination with decision support tools are promising solutions for that purpose.

1.1 Mobile laboratories for pathogen detection – why, when and where

The main reason of the increasing relevance of mobile laboratories in pandemic management is their potential to provide testing capabilities in remote areas as well as anywhere where testing capacities are limited or missing. This is both true for the equipment and the highly qualified staff. Worldwide disadvantageous developments such as climate change and reduced wildlife habitats make the occurrence of pathogens such as a CCHFV in Europe more likely or increase their prevalence in case of already occurring pathogens such as WNV, but also travellers may transport pathogens. The WNV is a virus transmitted by gnats, birds are its natural reservoir. It is endemic in multiple regions such as Africa or India. In Europe, each year individual cases as well as regional outbreaks are reported (AGES, 2024). In that light, within the Horizon Europe project MOBILISE, a high biosafety level (BSL) mobile laboratory is developed to close this existing diagnostic gap by establishing a quality-assured, mobile one health laboratory (MOHL) solution, with the potential capacity to handle pathogen up to risk group 4, available to many European countries (MOBILISE, 2024). Such capacities are scarce, only 3 % of about 200 identified European mobile labs reach a BSL4 level (Neubauer et al, 2023). From the perspective of shared European capacities within the European Civil Protection Pool out of 121 offered response capacities such as search and rescue units, flood containment or medical air evacuation, only one mobile laboratory is provided (ECPP, 2023) also stressing the relevance of more available mobile risk group 4 testing capacities.

1.2 Requirement for decision support

The application of a rare and expensive high risk group laboratory requires strategic planning and decision making. For that purpose, the highly qualified staff of the laboratory needs to communicate and share appropriate user designed information with different other actors. Such actors may be the staff from other mobile laboratories, critical infrastructure operators such as hospital or other health related stakeholders such as first responders. Moreover, it is imperative that authorities are involved such as regional and federal authorities as well as health agencies. The communication exchange process needs to be aligned with the legal mandates and workflows of all involved organizations, in case of Austria the legal frame is defined by the Austrian Federal Law for the Monitoring of Zoonoses and Zoonotic Pathogens (BGZ 2005). Mandates arising from legal requirements need also to be considered, when designing data management procedures. Each involved entity should know to which partners they provide selected data under predefined conditions. This can be achieved by applying identity services to enable trust between the involved actors. A certificate management of the involved organizations can protect against malignant actors intending to misuse health related data shared by the mobile laboratories or confidential data from health organizations. To set up a trustful and reliable process for communication and data exchange it is necessary to describe and understand in detail the regulatory workflow in case of potential threats for pathogens such as CCHFV or WNV. In the next chapter, such workflows are described for Austria to have a substantial basis for the design of the decision support system of the MOBILISE risk group 4 laboratory.

2. Workflows for the management of zoonotic events – the Austrian use case

At this stage it is important to note, that the BGZ 2005 regulates among other aspects also the exchange on information on zoonoses and zoonotic pathogens, the implementation of this law is a duty of the provincial governors. Both on federal as well on provincial level commissions with the mandate to manage zoonoses are established in Austria. Due to variations of the implementation of the federal law in the Austrian provinces, the workflow in the provinces may also differ. In this chapter we describe the workflows from the federal perspective, they reflect the practical implementation of the specific mandates of the different protagonists acting in the frame of the BGZ 2005.

In case of a human patient potentially infected by a zoonotic pathogen such as WNV a workflow is set in motion, that differs from the one of a potentially infected animal patient. For the Austrian use case, both workflows have several communalities, e.g., the involvement of National Reference Laboratories and National Reference Centres in central roles as well as data transfer both to national as well as European databases, another similarity is the involvement of Austrian Agency for Health and Food Safety (AGES) in the process. Apart from certain similarities there also a multitude of differences that are getting apparent in the next chapters.

2.1 Baseline workflows of zoonotic management in Austria – the human medical perspective

In Figure 1 the flowchart illustrates the sequence of occurring actions in case of an identified human patient with symptoms, potentially infected by WNV. Diamonds represent different types of actors such as the patient or an organization e.g. a ministry, rectangles stand for actions such as case acknowledgement or properties, e.g. symptoms of a patient and finally cylinders represent data repositories. The colour code represents different aspects: the colour green stands for the source of a zoonotic pathogen such as human or animal patient or blood samples from a blood donation campaign. The colour yellow reflects the involvement of any involved authority or other persons or institution involved in the management of a zoonotic event such as a medical doctor or a national reference centre.

In the first use case a patient with symptoms that may indicate infection by WNV visits a doctor where a first clinical assessment can be triggered, according to the judgement of the doctor. The doctor takes a sample and sends it to an analytic laboratory, where tests are performed. The doctor is assumed to register the potential case in the national database. In any case, also in case of a negative outcome of the test, the analytic laboratory must inform the National Reference Centre (NRC) on the outcome of the test and to provide a sample to the NRC. The laboratory also provides the information on the result to the patient and medical doctor that triggered the analytic test. The workflow is almost the same for the second use case, where blood samples are taken in a blood donation campaign. In such a case the analytic laboratory that analyses the samples from the campaign needs to act in the same way as in use case one. The third use case represents the treatment of a patient in a hospital with diagnostic capabilities, where a first medical assessment is done followed by an entrance in the national database by the laboratory staff from the hospital. In this case the hospital laboratory staff needs to inform the NRC and gives feedback on the outcome of the test to the doctor of the hospital in charge. In all three use cases, the NRC notifies the outcome of the second test performed by the National Reference Laboratory (NRL) or another certified laboratory in the national database.

The information on positive samples is explicitly provided to the Federal Ministry of Health, in addition provincial authorities, the AGES as well as the Federal Ministry can extract information form the national database at any time. Depending on their mandates, actors like the ministry, AGES or provincial authorities may need to acknowledge the positive case, the authorities decide afterwards on potential preventive or protective actions and measures such as further testing, social distancing, or provision of medication to potentially affected persons may be implemented. AGES reports positive case(s) in a European database.

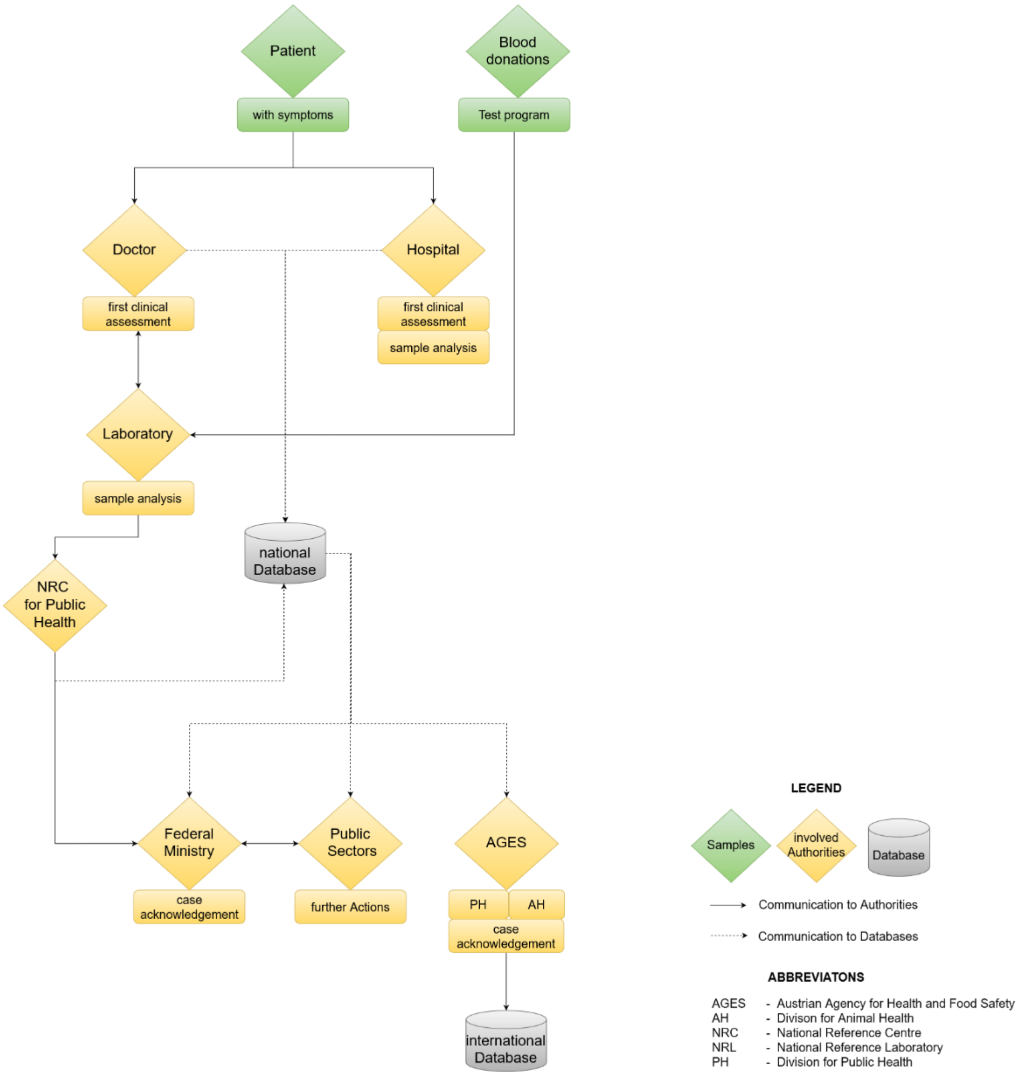


Figure 1. Workflows in the event of the occurrence of zoonoses for the human medical perspective

2.2 Baseline workflows of zoonotic management in Austria – the veterinarian perspective

The processes of managing the occurrence of pathogens such as WNV in case of their detection in animals differs in several ways compared to the human medical workflow. A major difference is that AGES is acting as national reference laboratory (NRL) in this case.

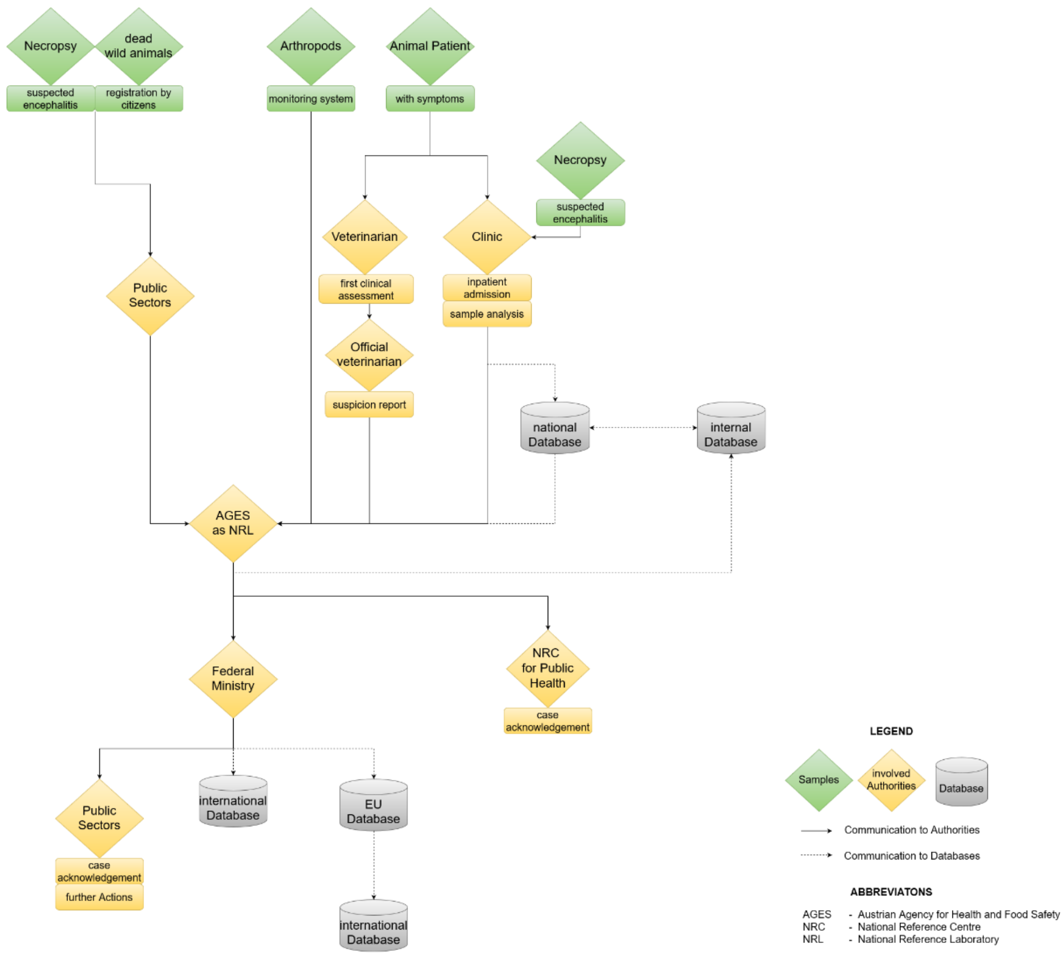


Figure 2. Workflows in the event of the occurrence of zoonoses for the veterinarian perspective

For the explanation of the meaning of symbols and colours, please refer to chapter 2.1. Looking at the beginning of the workflows in Figure 2, there are four starting points for diagnosis:

1. Necropsy in case of suspected encephalitis,
2. Dead wild animals reported by citizens,
3. Arthropods potentially acting as vectors identified by the monitoring system,
4. Animal patients with symptoms.

Path 1 and 2, both representing processes where information is coming from the public are very similar. In suspicious cases, the incident is reported to the NRL for confirming the results, afterwards documented in the legacy database of AGES and finally transferred to the national database. Path 3 reflects an initiative from AGES dedicated to the monitoring of arthropods for the purpose of pathogen detection. In this case AGES directly reports positive cases in the database, afterwards the information is also transferred to the national database. The 4th path shows the handling of animal patients with symptoms. There are two options of diagnosis: in the first case a first clinical assessment is performed by a veterinarian, in case of suspicious symptoms the official veterinarian arranges sample analysis, provides a suspicion report and informs the NRL. The alternative path starts with a clinical first level diagnosis and in case of a suspicion outcome an inpatient admission and sample analysis are performed. Also, in this case the information is transferred by the same process in the national database.

Independent of the selected path, AGES is a core actor and in the event of a positive WNV patient, also informs the other responsible sector, whether human or veterinary. Moreover, AGES informs the federal ministry in charge in case of positive diagnosis. The ministry consequently starts the process of informing the regional authorities, where a case acknowledgement is done and potential further actions such as mitigation measures may be initiated. The ministry also reports the cases in the international database of the World Organization for Animal Health (WOAH) as well as in the EU database and the international database from ECDC.

2.3 Enhanced interoperability between human and animal health management

As already indicated at the beginning of this chapter, commissions for the management of zoonoses are established both on federal as well as provincial level in Austria. A requirement arising from the BGZ 2005 is the implementation of integrated, risk-based surveillance programs also encompassing the surveillance of zoonoses and related pathogens (see BGZ 2005, § 5(3)) and the duty to provide to report on the evolution and sources of zoonoses and related pathogens on a yearly basis.

Systematically, timely shared knowledge on the occurrence of positive cases by decision makers involved in the management of zoonoses both on the veterinarian and the human side could help to enhance the reaction to zoonotic events in very early stages. Assuming that several infected animals due to a zoonotic pathogen such as WNV are detected in a district, it would be very beneficial if this body of information is provided as early as possible to authorities in charge engaged in human health management in the same and neighbouring districts, to give an example. Such a process would substantially benefit from a systematic information exchange between central actors from the human medical side on one hand and the veterinarian side on the other. Basically, such an approach could be based on the continuous information exchange between the NRCs of the human medical and the veterinarian side. In the frame of the Austrian trial of MOBILISE, this approach is currently examined.

3. Decision support system for a Mobile One Health laboratory

Due to multiple reasons mainly discussed in chapter 1.2 and 2.3 management of zoonoses can be substantially improved both by the availability of mobile laboratories as well as shared knowledge among human medical and veterinarian stakeholders. The mobile laboratory of MOBILISE and its integrated decision support system have the potential to fulfil both requirements at once. The concept of the decision support system enables decision makers such as NRCs to share relevant information such as on regions, where possible cases of zoonotic pathogens have been detected as well as the frequency and severity of cases (see Figure 3). These decision makers can exchange information with the staff of the Mobile Laboratory via their Mobile Laboratory Data System (LIMS), additional information for decision support can be supplied by data experts making predictions on potential evolution of the pathogen frequency and distribution possible and serving as basis for multiple decisions such as future regional needs for testing capabilities. The Mobile Laboratory can support the process of medical sample taking in case of lack or limitations of adequate testing capacities and speed up the decision making process due to secure, trustful and fast interconnection between the laboratory in the field and authorities in charge. Within MOBILISE four trials taking place in Austria, Greece, East Africa, and Romania are currently in the preparation phase.

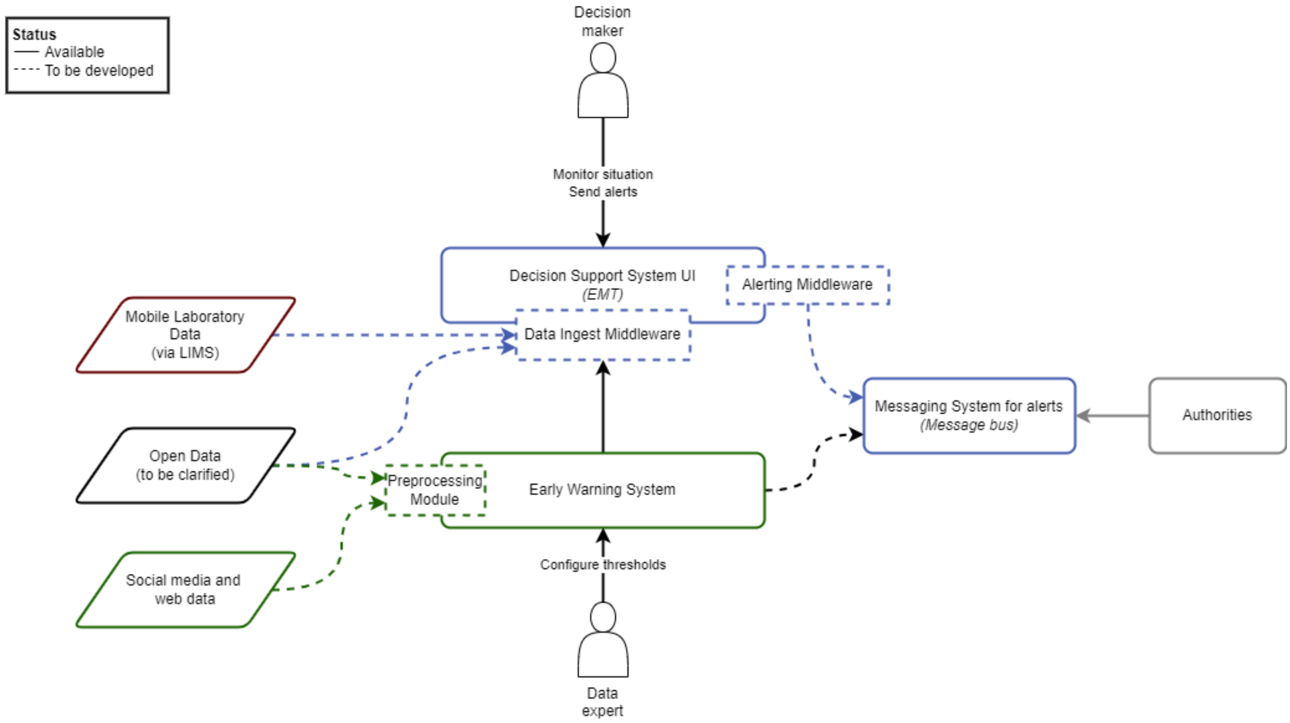


Figure 3. The concept of the decision support system

4. Conclusion and outlook

The workflows for the management of zoonoses shown in this paper represent only the federal perspective from Austria for the specific pathogen West Nile Virus (WNV). One must bear in mind that the workflows as well as the actors will at least partially differ on provincial and regional level and also vary between the different provinces due to differences in the regional implementations of the Austrian Federal Law for zoonoses management (BZG 2005). In addition, the processes for the management of other zoonoses will be distinct from the ones for WNV, because of disparities between pathogens, e.g., variations in contagion paths, severity of symptoms, different epidemiological characteristics such as reproduction rate or mortality among other parameters. Finally, in other states a different legal framework with different actors and other mandates will again lead to different conditions for the management of zoonoses.

Within MOBILISE such variabilities are considered, e.g., by taking the specific workflows of the province of Styria in the design of the Austrian Trial into account. For the other trials in MOBILISE there is need to consider the workflows of the states and regions of interest, i.e., Greece, Africa, and Romania.

For future application of a mobile lab as the MOBILISE project develops it, a mapping of specifically vulnerable areas regarding zoonoses and vector borne outbreaks would be a valuable support for timely deployment and cross-border cooperation. This has to include specifically for the European area not only common, well-known factors for spill over and spreads like in tropical areas due to the progress of human habitats (WHO, 2020), but also the new developments of re-naturation and increase of biodiverse wet-lands (see e.g. MOIST 2024). This aspect has to be researched more in-depth together with the benefits of this development in the industrialized regions of Europe. In addition, human mass movement in the course of uncontrolled migration can also be taken into account as a prominent aspect for the location of a mobile lab to support the safety in case of cross-border and dynamic outbreak scenarios (Koinova et al, 2023).

Acknowledgment

The research leading to these results has received funding from the European Union`s Horizon Europe research and innovation programme under Grant Agreement n°101073982 (MOBILISE).

References

AGES (2024), West Nile Virus [online], https://www.ages.at/en/human/disease/pathogens-from-a-to-z/west-nile-virus

BGZ (2005), Bundesgesetz zur Überwachung von Zoonosen und Zoonosenerregern, [online], https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20004373

ECCP (2023), European Civil Protection Pool, [online], https://reliefweb.int/map/turkiye/european-civil-protection-pool-offeredregistered-capacities-dg-echo-daily-map-09112023

Koinova, M., Düvell, F., Kalantzi, F., de Jong, S., Kaunert, C., Marchand, M.H., International Politics of Migration in Times of Crisis and Beyond the Covid – 19, Pandemic, Migration Studies, vol.11, Issue 1, March 23, pp 242, [online], https://academic.oup.com/migration/article/11/1/242/7010550?login=false

MOBILISE (2024), A novel and green mobile one health laboratory for (re-)emerging infectious disease outbreaks, Cordis Website, [online], https://cordis.europa.eu/project/id/101073982

MOIST (2024), Erfassung degradierter Moorflächen Österreichs und Beurteilung Ihrer Eignung zur Regeneration, [online], https://geographie.univie.ac.at/arbeitsgruppen/geooekologie/forschung/laufende-projekte/moist-erfassung-degradierter-moorflaechen-oesterreichs-und-beurteilung-ihrer-eignung-zur-regeneration

Neubauer, G., Schimak, G., Ignjatovic, D., Peham, J. R., Gehre F., Affara, M., Rainer, K. Duscher, G., Leidwein, A., Vlachostergiou, A., Jaho, E (2023). Decision Support for Application of Laboratory Platforms in Pandemics, IDIMT 2023 – New Challenges for ICT and Management, Vol. 52, pp 13, [online], https://idimt.org/wp-content/uploads/2023/08/IDIMT-2023-proceedings.pdf

WHO (2020), Zoonoses. [online], available at: https://www.who.int/news-room/fact-sheets/detail/zoonoses

WHO (2024), sixty first session, zoonotic disease: emerging public health threats in the region. [online], https://www.emro.who.int/about-who/rc61/zoonotic-diseases.html

AI-BASED VOICE DIALOGUE SYSTEM FOR ENHANCING CRISIS MANAGEMENT IN THE KRISAN PROJECT

Melissa Hagendorn, Karin Rainer, Viktoria Kundratitz, Alois Leidwein

Austrian Agency for Health and Food Safety GmbH
melissa.hagendorn@ages.at

Ferdinand Fuhrmann

JOANNEUM RESEARCH Forschungsgesellschaft mbH
ferdinand.fuhrmann@joanneum.at

DOI: 10.35011/IDIMT-2024-43

Keywords

Crisis management; Artificial Intelligence; automated call retrieval; real-time data analysis; assistive technology; workload management

Abstract

The COVID-19 pandemic exposed the challenges decision makers involved in a crisis face in obtaining a reliable picture of the situation. Collecting data rapidly and efficiently, incorporating diverse sources of information, and merging information into a near-real-time situational picture are crucial for managing critical situations effectively. The primary objective of the KRISAN research project is to develop the model for an AI-based assistance system that can automatically manage calls to hotlines and control centres and create a near-real-time situational picture during a crisis. This system aims to significantly enhance crisis management capabilities by providing decision makers with timely and accurate information. After the transfer into a product, the AI-based voice dialogue system will process hundreds of simultaneous calls. Two selected use cases, in which the phone calls are simulated, will be used to analyze and evaluate these calls in real-time using automated call centre support. Consequently, the system will dynamically generate relevant questions based on the course of the call to extract relevant information from the caller. By streamlining the call-handling process and leveraging AI technology, the KRISAN project aims to improve crisis response effectiveness, resource management, and contribute to better outcomes during health emergencies.

1. Background

When dealing with crises such as pandemics, operation centres, public agencies, and authorities are put to the test from organizational, logistical, and emotional perspectives. Protecting the population is a key priority that is very complex. Side effects must be considered when implementing broad-scale countermeasures. In the event of a crisis, it is challenging for those involved to quickly gain an overview of the situation and obtain a reliable picture (Cosgriff et al. 2020). In particular, the COVID-19 pandemic outbreak has illustrated the importance of rapid data collection (Gao & Janssen 2022), the development of additional information sources (Afzal 2020), and intelligent consolidation of data into a near-real-time information picture to manage critical situations effectively and make responsible decisions.

Call centres are a traditional tool and have been instrumental in facilitating quick and transparent communication to and from the public (World Health Organisation, WHO), both globally and in Austria since the beginning of the COVID-19 pandemic. They serve as vital hubs for individuals to seek information and clarification on various aspects of the pandemic.

In a crisis, countless phone calls are received from the public every minute, which cannot be processed simultaneously due to technical and resource limitations. When multiple callers call a hotline simultaneously, all other callers are put on hold. Once a call is answered, the call taker can enter the relevant information provided by the caller into a documentation system. Currently, rapidly generating a reliable picture of the crisis situation is time- and resource-consuming. Consequently, valuable information from local data sources and directly from the population often cannot be utilized immediately.

The call statistics of the Austrian Coronavirus Infoline during the COVID-19 pandemic illustrate the scale of the problem: on peak days, more than 50,000 calls per day were registered. The main activity on the infoline was between 06:00 and 22:00. During these 16 hours of peak activity, the employees of the helpline had to process around 52 calls per minute. In 2021, call takers from the Austrian Health and Food Agency (AGES) at the Coronavirus Infoline answered a total of 1,848,566 calls. This corresponds to around 5,000 calls per day with an average duration of 3.86 minutes. In total, AGES had to allocate about 70 full-time equivalents to handle the calls in 2021.

This high demand for information from reliable sources vividly depicts the need for a technologically supported call handling system. Additionally, there is a critical need to relieve the workload on call centre operators. To address these needs, an AI-based voice dialogue system will be implemented and tested in two different use cases. One use case will focus on handling an infoline during a pandemic crisis situation. The second use case will involve a service scenario where callers request specific services, such as booking appointments for vaccination or testing, or requesting emergency services. This system aims to reduce the workload on human operators in call centres by automating the processing of these calls.

2. Objectives of the project KRISAN

The project aims to develop a German-language assistance system for automated call retrieval and near-real-time generation of an information situational picture, based on incoming calls to hotlines and control centres in the event of a crisis, as shown in Figure 1. The automated evaluation of calls allows hundreds of calls to be answered simultaneously by an AI-based voice dialogue system in a short period of time. An automated call query also makes it possible to analyze the information from many calls in parallel and in a technically sound way. Callers can provide information or questions in colloquial language, and the system automatically extracts the necessary information. The system generates the appropriate questions in real-time, depending on the course of the call, in order to obtain any missing information from the caller. Ideally, the caller will not recognize any difference between the conversation with a human operator and the voice dialogue system.

To ensure transparency and clarity, callers are informed that they are speaking to an AI voice assistant. If the automated voice dialogue system, as the 1st level responder, cannot assist a caller, human operators (2nd level responders) are still available. This handoff may occur for callers with speech impediments, voice disorders, extremely pronounced dialects, limited German language skills, or complex requests that the AI cannot handle adequately. In such cases, the call is forwarded to a human operator to ensure proper assistance. Additionally, callers can always request to speak to a human operator to strengthen trust and user acceptance.

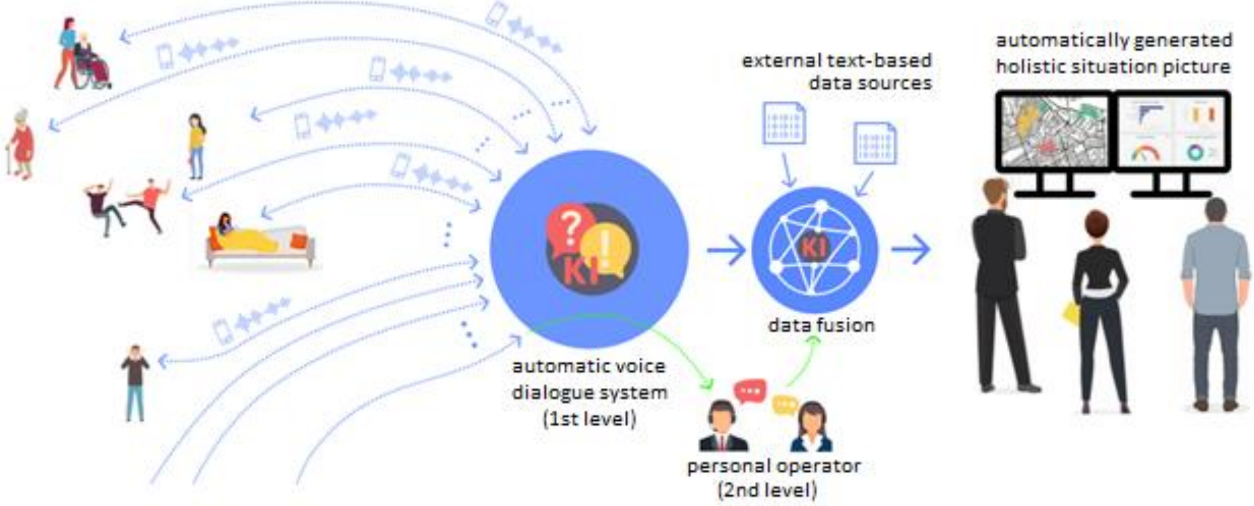


Figure 1. Representation of the information flows of the voice dialogue system

The details and information provided by different callers can increase the density and thus the reliability of the information received. This data can be combined with other external data and used as basis for a reliable and quickly compiled information picture, which can serve as a basis for decision makers. Data protection and other applicable legal regulations concerning the use of this information must be embedded in the early design phase of the development to ensure legal compliance of the future service tools stemming from the project results.

In this project, an AI-based call support system will be designed and developed based on two specific, prioritized use cases from the healthcare sector. These use cases were determined by stakeholders through a voting process. The selected use cases are:

5. a supplementary support tool for the dispatch management of the Ambulance Service of the Austrian Red Cross Branch of Vienna and
6. an infoline support for pandemic and health crises for AGES and the Austrian Ministry of Health in particular.

The selection of these use cases was based on various factors, including aspects of human resources such as average call intervals and volumes during crisis and non-crisis periods, as well as call duration and the likelihood of application. Human operators at the Vienna Red Cross call centre manage both emergencies and patient transports, including return transports. Leveraging AI for return transports will reduce their workload during crises, potentially alleviating the increased demand for emergency dispatch.

3. AI application in KRISAN use cases

3.1 Requirements and potential impacts

The AI-supported assistance system must fulfil several requirements and bridge the gap between transparency, socio-cultural, ethical and legal requirements and technical limitations (Kiseleva et al. 2022). General requirements and constraints on potential impacts were considered in the development and implementation of the use cases.

The development of a prototype for a voice dialogue system aims to significantly *reduce staff workload* in the healthcare sector *and to function reliably* during high call volumes (López Jiménez & Ouariachi 2021). For this reason, a total of seven use cases were identified and assessed by stakeholders regarding various aspects.

The reliability of the AI in information transfer is essential to minimize misinformation for both callers and operators (Shneiderman 2020). An AI must embody ethical and moral values, such as respect for autonomy, fairness, transparency, explainability, and accountability (van de Poel 2020). Furthermore, it must ensure that the conversation is transparent, using the simplest and clearest possible language to ensure accessibility and inclusivity (Aumayr 2021).

The assessment of legal requirements for the project's systems must be considered from the phase of development until operation. Each phase harbours specific risks that need to be recognized and assessed. A preliminary risk analysis was conducted during the development of use cases to ensure legal aspects are adequately addressed. Legal requirements were evaluated for each phase, particularly based on the regulatory regimes of the AI regulation (AI Act 2024), data protection law, equal treatment law, and liability law. AI systems interacting directly with people require additional transparency from operators.

Recent advancements in *deep learning* have significantly enhanced AI systems in speech processing. Techniques such as convolutional neural networks and recurrent neural networks, including Long Short-Term Memory networks, have improved the accuracy and efficiency of speech recognition and natural language understanding. These models are capable of processing large datasets to learn complex speech patterns, making AI-driven voice dialogue systems more robust and responsive. These advancements underpin the AI-based call support system in this project, enabling it to handle high call volumes and provide accurate, timely information (Mehrish et al. 2023).

User acceptance and ethical aspects are critical components that need to be addressed. The hook model (Eyal 2014) will be the basis for developing the needs of future users of the AI-supported voice dialogue system, illustrated in Figure 2. The hook model describes users' interactions with a product by going through four phases: a trigger that leads to the use of the product, an action to satisfy the trigger, a variable reward for the action (e. g., in the form of a response), and some kind of investment that ultimately makes the product more valuable to the user. These four phases form an iterative loop: as the user repeatedly goes through these phases, they develop habits.

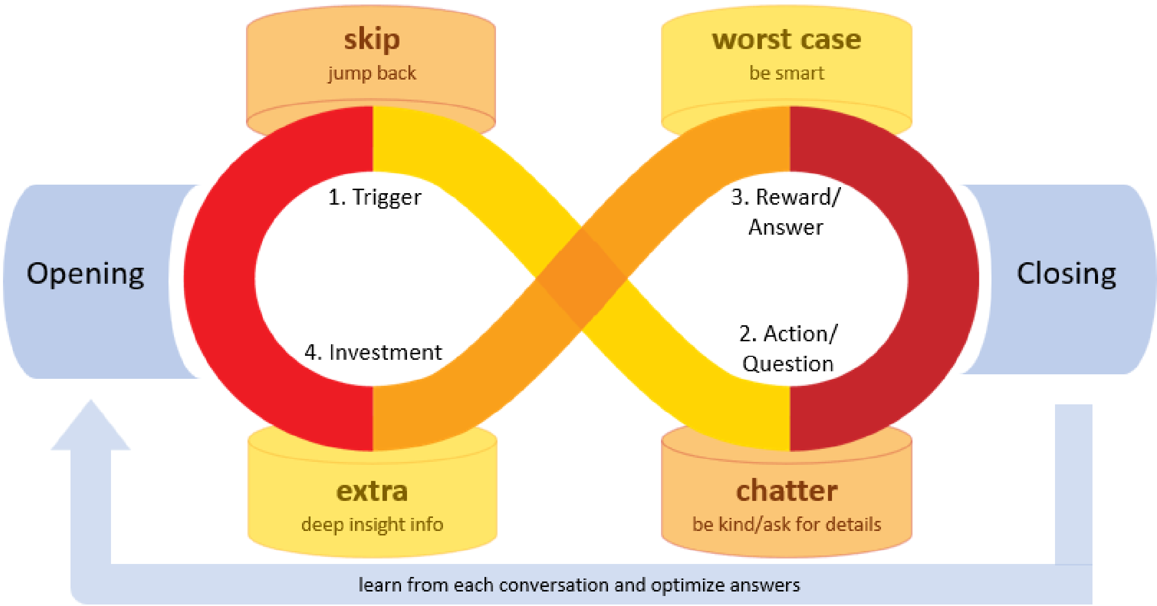


Figure 2. The hook model forms an infinity loop of four phases (modified from Eyal 2014)

Behavioural intention and willingness to use an AI-based assistance system are crucial for the system's development (Kelly et al. 2023). Communication must be simple and understandable, and a pleasant and clear dialogue must be facilitated, while adhering to all necessary transparency measures. This includes, for example, informing users that they are communicating with an AI-supported system. A brief explanation should be given at the beginning of each conversation to promote acceptance, clarifying that AI is being used to optimize the use of human resources.

Gender and diversity aspects must also be taken into account (González & Rampino 2024). In particular, attention must be paid to age- and gender-sensitive language to ensure the language used is as universally understandable as possible. Minorities and vulnerable, underrepresented groups, such as people whose first language is not German and people with language deficits, must also be sufficiently considered by providing the option to bypass the AI system and easily reach a human operator.

3.2 Description of the use cases

3.2.1 Service use case "patient transport"

"Patient transport" is a so-called service use case and can be requested by any individual. The caller orders an ambulance for themselves or another person by providing various personal and location-related data. They specify the departure and destination addresses while speaking to the AI-based assistance system (1st level) or, alternatively, a personal operator (2nd level), see Figure 3. Any special equipment required is also requested. In the early planning phase of the project, this use case included planned patient transports, such as to to medical appointments and examinations with varying degrees of relevance, as well as home transports, for example, return transport to the care facility or home.

Emergency transports are explicitly excluded here due to the high urgency and time factor and the associated susceptibility to risk when using the AI-based system. This use case is technically less complex than the infoline use case described below, as the structure of the call is easier to map.

Due to the resources of the project and the potential for resource saving, it was necessary to focus on return journeys from medical facilities. These return journeys are usually made from one clinic to another, to a specialist or to the patient's home, and the communication with the dispatch call centre is conducted exclusively by health care staff.

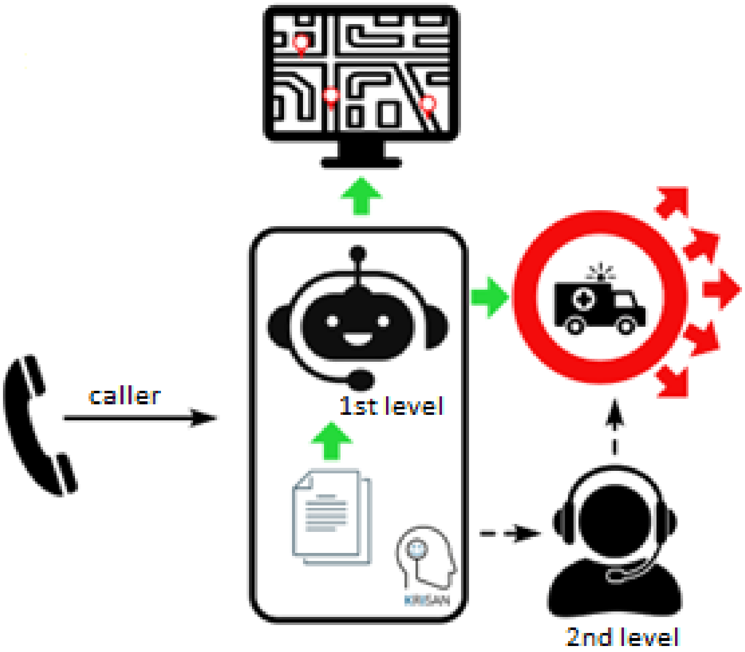


Figure 3. Schematic overview of the communication process in the service use case

The aim is to develop a system that is as independent as possible by generalising information from as many examples as possible, such as test calls carried out as part of the project. The technical system is being developed iteratively, as not all the information required during the calling process when ordering a transport will be known from the outset.

A real-time demonstrator will be developed for the service use case and integrated into the existing framework of the Red Cross Vienna. At the end of the project, this will facilitate an evaluation in practice with the users while considering the legal aspects of the process. In this phase, all possibilities of the call flow must be covered, including specific and difficult cases, to make a well-founded assessment of the applicability of the technology.

3.2.2 "Infoline" use case

The health crisis infoline is intended for use in a future pandemic or health crisis in case the call volume becomes too high for quick and efficient handling by human operators. Due to the large number and variety of possible questions, this use case will only be addressed conceptually in order to develop the basis for a possible subsequent project.

The plan includes creating a timeline and discussing trigger values for the set-up of such an AI-supported system. Based on the experience gained from the COVID-19 crisis, an exemplary roadmap will be drawn up with milestones that, for example, map the availability of vaccinations against the pathogen. The definition of key terms and prioritisation of relevant information areas must be clear, and an initial set of supporting information on generic topics (e. g., reliable data sources) could be prepared. The system must be managed by the operator's experts to ensure timely and rapid growth of this section in response to specific questions on the current scenario. External data sources such as the WHO can also be incorporated early in this process. This includes creating frequently asked questions (FAQs) and regularly updating them based on developments during the pandemic phases or the health crisis.

4. Benefits and challenges

Until now, voice dialogue systems have primarily been used in call centres within the industrial sector (e. g., purchasing tickets, reporting faults). The innovation of this project lies in developing, for the first time, an automated voice dialogue system designed to assist both the healthcare sector and the general population in managing crisis situations such as a pandemic. This system's novelty stems from the combined use of automatic speech recognition, natural language understanding, and the integration of external AI systems through an interface into existing command and control systems for situational awareness.

It is important to note that current AI-based systems are not yet capable of developing algorithms that can be generalized for all crisis situations. Therefore, it is necessary to focus on specific scenarios where these systems can be effectively deployed and utilized in the (crisis) situations for which they have been developed.

4.1 Potential benefits

Resources freed up: There is currently no application that is capable of processing calls at a scale of hundreds per minute. The proposed system will automatically recognize the information it needs from the caller and ask generated questions in order to obtain all necessary details. This reduces the burden on human operators, freeing up resources for other critical tasks.

The voice dialogue system is being developed in a resource-efficient manner to *adapt quickly* to rapidly evolving events. In the event of a crisis situation, this adaptability represents an opportunity to tailor and enhance the system to meet specific situational needs (Gkeredakis et al. 2021).

Real-time, *automatic generation of a holistic situational picture*: Current situation reports in control centres typically have two primary characteristics: one view focuses primarily on number-oriented representations of active or processed calls using various diagrams, while the other visualizes resulting deployments or tasks on a map. The aim of this project is to develop a holistic situational picture visualization. Automatically generating a situational picture in real-time based on calls to information hotlines or emergency numbers is a novel advancement. Providing decision makers with reliable real-time information about the national situation would be immensely supportive. This development would have aided not only in the recent pandemic situation but will also enhance societal support in terms of situational awareness and improve future crisis management. Thus, the creation of a real-time situational picture represents a significant innovation over the current state of the art.

4.2 Countering risks and challenges

The AI must gather caller information and allow for switching to human operators when needed, ensuring partial processing by the AI before forwarding to humans. The system must be reliable to minimize false responses.

In the patient transport use case, the system will be used exclusively by qualified medical staff and will operate on a separate phone line. Furthermore, only non time-critical patient return transports from hospitals or doctors to home or care facilities will be requested. This minimizes the risk of a false or misleading response by the voice dialogue system.

For the infoline use case, answers are only permitted if they are based on validated sources of information. If a question is not understood correctly, the AI will forward the caller to a human operator. Consequently, the risk of the AI giving misleading answers is eliminated; the voice dialogue system may only give answers that are certainly correct, though they may not always be helpful for the caller.

In crisis situations, an AI must still be accompanied by human input. While AI offers advantages in speed and accuracy, the more cognitively complex and emotional aspects still require human operators (Aung et al. 2021).

When using external data sources, the project will ensure compliance with legal requirements (data protection, liability, and equal treatment law). These external data sources may include examination data, vaccination records, previous illnesses, data from the central register of residents, side effects of medication, and location and property data from hospitals.

5. Discussion and outlook

The initial phase of the project has been successfully completed. During this stage, the two selected use cases were meticulously elaborated, and detailed plans were developed, outlining all necessary requirements. Additionally, several hundred test calls were performed with test callers and human operators using self-developed user stories to train the AI.

In the upcoming phase, the focus will shift towards a comprehensive approach to achieving legal compliance and ethical standards, ensuring real-time data transparency and enhanced security in the post-project period. Central to the technical implementation is the acceptance and willingness of future users to engage with the AI-based assistance system, which will be promoted by the early integration of usability considerations along with ethical, legal, and socio-behavioural needs.

Planned *evaluation processes* will involve further test calls conducted by the same test callers and with the same user stories from the first stage. These calls will be handled by the so-called VoiceBot instead of human call takers. This evaluation will include a quantitative analysis to determine the success rate of the calls, and a qualitative evaluation through interviews with the test callers.

Once the initial evaluation is successfully completed, the system will be tested in live operations at the Red Cross Vienna call centre in a supervised mode, with the VoiceBot supervised by a human call taker. This evaluation will also proceed in two steps: a quantitative analysis that provides statistics on the number of successful calls, followed by a qualitative analysis based on feedback from the accompanying call takers.

In the future, KRISAN’s adaptable AI-based voice dialogue system will enable the simultaneous processing of hundreds of telephone calls. Caller information will be recorded quickly through automatic transcription and summarized in an information map. This will not only effectively prevent long waiting times in queues but also ensure that the information from the calls is immediately accessible to those responsible for assessing the current situation.

Moreover, it is of particular interest for authorities and decision makers to generate comprehensive situation reports from the information provided by callers, offering a comprehensive overview of the prevailing situation.

However, the project highlights that the application of AI in dynamic scenarios requires extensive background research to meet future needs. A holistic approach is essential for the further development of a successful tool and its adoption and implementation in current practice.

Acknowledgement

The KRISAN project, leading to these results, has received funding from KIRAS Cooperative F&E Projects 2021 with project number FO99901586.

References

Afzal W. (2020). What we can learn from information flows about COVID-19: Implications for research and practice. Proc Assoc Inf Sci Technol. 2020;57(1):e245.

AI Act: European Parliament 'Corrigendum' of 16th April 2024. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf retrieved 26.04.4024

AI Act explorer. https://artificialintelligenceact.eu/ai-act-explorer/, retrieved 09.04.2024

Aumayr, G. R. (2021). Ethical Evaluation Standard for Security Research. DCNA Disaster Research Days 2021, 41.

Aung, Yuri Y. M., Wong, David C. S., Ting, Daniel S. W. (2021). The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare, British Medical Bulletin, Volume 139, Issue 1. https://doi.org/10.1093/bmb/ldab016

Cosgriff, C. V., Ebner, D. K., Celi, L. A. (2020). Data sharing in the era of COVID-19 (2020). Correspondence Volume 2, Issue 5, E224. https://doi.org/10.1016/S2589-7500(20)30082-0

Eyal, N. (2014). Hooked: How to Build Habit-Forming Products. Portfolio Penguin, 2014, ISBN 978-1591847786.

Gao, Y. and Janssen, M. (2022). The Open Data Canvas–Analyzing Value Creation from Open Data. Digit. Gov.: Res. Pract. 3, 1, Article 5, 15 pages. https://doi.org/10.1145/3511102

General Data Protection Regulation GDPR https://gdpr-info.eu/#:~:text=General%20Data%20Protection%20Regulation%20GDPR, retrieved 12.04.2024

Gkeredakis, M., Lifshitz-Assaf, H. & Barrett, M. (2021). Crisis as opportunity, disruption and exposure: Exploring emergent responses to crisis through digital technology. Information And Organization, 31(1), 100344. https://doi.org/10.1016/j.infoandorg.2021.100344

González, A. S., Rampino, L. (2024). A design perspective on how to tackle gender biases when developing AI-driven systems. AI Ethics. https://doi.org/10.1007/s43681-023-00386-2

Kelly, S., Kaye, S. A., Oviedo-Trespalacios, O. (2022). What factors contribute to the acceptance of artificial intelligence? A systematic review. Telematics and Informatics, Volume 77, 2023, 101925. https://doi.org/10.1016/j.tele.2022.101925

Kiseleva, A., Kotzinos, D., De Hert, P. (2022). Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations. Frontiers in Artificial Intelligence, Volume 5. https://doi.org/10.3389/frai.2022.879603

López Jiménez, E. A. and Ouariachi, T. (2021). "An exploration of the impact of artificial intelligence (AI) and automation for communication professionals". Journal of Information, Communication and Ethics in Society, Vol. 19 No. 2, pp. 249-267. https://doi.org/10.1108/JICES-03-2020-0034

Mehrish, Ambuj & Majumder, Navonil & Bharadwaj, Rishabh & Mihalcea, Rada & Poria, Soujanya. (2023). A review of deep learning techniques for speech processing. Information Fusion. 99. 101869. 10.1016/j.inffus.2023.101869.

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. International Journal of Human–Computer Interaction, 36(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

van de Poel, I. Embedding Values in Artificial Intelligence (AI) Systems. (2020). Minds & Machines 30, 385–409. https://doi.org/10.1007/s11023-020-09537-4

CYBER SECURITY

PROBLEMS OF DATA OBFUSCATION AND POTENTIAL SOLUTIONS, APPLIED TO THE EXAMPLE OF 3D DATA

Michael Sonntag

Johannes Kepler University Linz
Institute of Networks and Security
michael.sonntag@jku.at

DOI: 10.35011/IDIMT-2024-55

Keywords

Obfuscation, anonymization, obfuscation methods

Abstract

For personal data anonymisation is often desirable or required. But for other kinds or types of data similar aims exist, e.g. the obfuscation of 3D data: scans of environments may be required for various uses and need to be stored and perhaps partially passed on, but attackers should be unable to use this information for performing or improving their attacks. This paper discusses how obfus-cation can be performed in general (methods/approaches) and applies it to a specific example, 3D scans of cable ducts. As can be seen from this, obfuscation 3D data is very difficult or impossible, as many elements depend on each other or can be recreated from other sources.

1. Introduction

Obfuscation of data, i.e. rendering it unusable for unauthorized/unintended recipients, is important in many aspects, the most prominent being anonymization of personal data. In this area mathemat-ical measures have been developed to assess the "quality" of anonymization, and methods how to perform it are known (see Lison et al 2021). Other types of data exist too, where similar activity is desirable: a part of it might be disclosed, but it should be impossible for the recipient (or third par-ties) to identify the rest or use it undesirably. I.e. we talk about partitioning some data into two sets: one that is disclosed to another party, and another one to be kept secret. And from the public part it should be impossible, even with additional data, to recreate/recognize/select the private one, even partially (i.e. some interesting aspect). The disclosed part can be either a direct subset of data elements, but also a "coarser" view on the "full" data, e.g. omitting details, after modifica-tions etc.

When arranging data like a table in individual objects as rows and their aspects as columns of var-ious type, disclosure can take the form of disclosing several rows in total, from all rows only a subset of columns, or a subset of both dimensions. Here we are only interested in the aspect of disclosing a subset of attributes of the objects, i.e. a "column subset". If several objects with all their associated data are to be published, other requirements will arise, e.g. preventing deductions on the whole based on the set (e.g. if the subset is a representative sample of an attribute, the aver-age value of that attribute will still be the average of the whole population).

This partitioning/modification of the attributes to achieve obfuscation should be performed according to several requirements to be secure:

- The disclosed data should be useful to the recipient. Obviously reducing the data to an empty set is secure, but useless, while disclosing the complete data is very useful but inse-cure. Some balance is therefore needed, which includes defining exactly what parts/aspects/… need to be kept secret. This is a problem of the application domain and re-quires an attacker profile (what they will be interested in).
- From the disclosed data it should be impossible to infer the rest of the data not disclosed. Note that here no additional knowledge is assumed, i.e. from the data itself no conclusions on other data should be possible. An example would be disclosing the height and weight of a person, which will indicate the (undisclosed) BMI of that person. For this we do not need a list of all persons in the data set, identify the person on that list, and then "transfer" the BMI to the other list. Neither do we need general information unrelated to the actual data set, e.g. a list of "first name gender", from which the determination of the (undisclosed) gender based on the (provided) first name will be possible. This category therefore pertains to direct redundancy in the complete set of data, where the disclosed subset contains enough for reconstruction (in the example a simple calculation/table lookup). This is inde-pendent of the resources of the attacker and only depends on intrinsic aspects of the data. So if attackers are intelligent enough, respectively possess domain knowledge, de-obfuscation is guaranteed (typically with little resource requirements).
- From the disclosed data together with some additional external knowledge (speci-fied/assumed in the attacker profile; could be generally known or strongly restricted) the missing data cannot be reconstructed. I.e. the data on the object is available with different attributes (but some overlap) from other sources and we have to align both lists to obtain a comprehensive view and assign the additional attributes of both lists to the same element. This is exemplified by de-anonymization, where a set of data is attributed again to a specif-ic person. An alternative is "guessing" missing data based on general information. This de-pends on the resources and pre-knowledge of the attacker, and obviously can only be ful-filled if less information is disclosed as required for the specific assumed knowledge.

Note that this obfuscation not only depends on the general properties of disclosure (i.e. which attributes are selected for publication), but also the specific values in the actual dataset. E.g. for anonymisation of personal data the aspect to be kept secret is the identity of the person it refers to. But this "identity" stems from a combination of elements and their uniqueness. Therefore it needs an assessment, whether the obfuscation is "good enough" for each and every row: e.g. is the name "John Miller" precise enough? It is probably quite common and without additional knowledge no person can be uniquely identified. Compare that to "Crescencia Unterhabichtsberger", which is technically the same data (first + family name only), but probably much more unique. In the ex-ample of cable ducts, a voltage display for mains current will disclose very little information, but a flow-meter/pressure gauge will probably be much more useful to attackers.

1.1 Related work

A similar problem of obfuscating 3D data exists in 3D printing: the theft of models. For this, approaches have been developed to prevent successful manufacturing except under special circumstances (Gupta et al 2017). However, these pertain specifically to 3D printing. E.g. if the model of a cable duct has small "gaps" at corners, this is no "loss" for an attacker. Additionally, scan shad-ows will in many cases produce such gaps anyway. The simplest form proposed is remote render-ing

(Koller/Levoy 2005), where the 3D data itself is simply not accessible to attackers – which we cannot assume here. However, the alternative listed in that paper, rendering on secure hardware, is what was finally implemented through SGX (see at the end).

Obfuscation of 3D data has been investigated especially in the case of buildings (Google Street view) before (Kada et al. 2009). Problems there are related, but mostly cover abstraction of build-ings, removing windows etc. Removing license plates/signs/house numbers are relevant aspects: signs exist in ducts too and are related to gauges (dynamic aspects are irrelevant for obfuscation). Algorithms for recognizing such elements and their spatial boundaries are therefore transferable.

Obfuscating the location in 3D space, especially for drones (Naeem et al 2021) or mobile devices (Ardagna et al 2009) is another research area. However, this is not very relevant here, as only a single location for a cable duct is relevant, the point of origin for the model – and that is mostly irrelevant and easily deleted. The robot location, e.g. at scan time, is not relevant for attackers.

The closest approach to the aims in our practical example is "visual privacy" for virtual reality (Tabet et al. 2023). However, their aim is to "remove" a 3D object (selected by a user by enclosing it with a 3D primitive) in a way, that changing the position does not change the hiding. However, human intervention to select elements to hide is complex and requires a lot of work.

Targeting specifically point clouds and machine learning is another aspect (Sun et al 2023, Zhang et al 2023). In our example the actual use would be to pose as "attackers" to poison any machine learning model "real" attackers as considered here (in the publications: model creators/owners) might construct. Note that constructing the machine learning model might undo/detect any obfus-cation measures, so this is worth considering.

An interesting approach is presented by Jolfaei et al (2015), direct encryption of 3D data. Coupled with object recognition this could be used to selectively encrypt parts of a cable duct, rendering them useless to attackers. Problematic would be how to detect/mark, which areas are encrypted and therefore need decryption, as the encrypted points would be intermingled with unencrypted ones. This is not a problem in another approach, where encrypted objects remain within their bounding box (Éluard/Maetz/Doërr 2013). This approach again requires identification of the bounding box for the elements to be secured.

2. Problems of obfuscation

A difficult problem in obfuscation is, that many elements may depend on each other or may be related. In this way any change to one element requires changes to potentially multiple others, or the modification is easily recognizable as such, distinguishable from normal data, or even undoable. In 3D data this applies e.g. to the individual measurements: a surface will generally be flat, so moving individual points is severely limited (small bumps may occur, but not a spike of 20 cm, with the wall 2 cm beside it being completely flat…). The same applies to larger structures, where pipes, their supports (size & distance) as well as valves etc all match (pipes of a certain diameter need a certain strength of support, valves must have specific size to be able to withstand the pressure etc).

This problem exists also, but only to a lesser degree, with classical personal data and anonymization. An example there would be the first name, which in most cases will determine the gender of the person. Also height and weight are (loosely) correlated, similarly age and name to a picture of face etc. Anonymization is usually aiming to prevent unique identification. I.e., a set of data (=single row) exists - can we identify/point to the single person it belongs to? Anonymisation then tries to ensure that at least several (actual; otherwise verifying whether each potential individual exists in reality at all will lead to the correct one!) persons will fit the remaining/altered data. So while attackers may possess data about a person, but cannot decide which person it is, or in reverse, know a person but cannot decide which row (set of attribute values) pertains to it.

With obfuscation, especially in this project, the aim is similar but covers more aspects. We want to prevent the localization of the duct entrance, but also its exact trajectory etc. But we cannot readily generalize data to a large set of actual entrances as possible alternatives: in many cases no entrance should be disclosed at all. Alternatively, we can provide a large number of "possible entrance locations": but then again verifying them and ruling those out not existing in reality reduces the utility of the approach as it only becomes more work and time-consuming for attackers, but will still eventually lead to the correct location. Modifying e.g. the trajectory is easier on this level, as verifying it requires access to it and may even then be complex – at least when exactness is required.

Re-identification is therefore a function leading from a set of potential actual persons to a single one, and anonymization the reverse. Obfuscation aims at hiding specific aspects, by either removing them or providing believable but incorrect alternatives. I.e., if only data about a single person exists, this person's relevant aspects must be credibly masked. Important is therefore whether this "incorrectness" can be pinpointed (if only details are modified), or easily validated (e.g. look at building: is there a door?). So anonymization is mostly working with removing or slightly falsifying data to hide within a group of (then) similar people, while obfuscation often devolves to creating new, but believable, fake data. This is especially problematic in cases of interdependency of attributes, as modifying a single value convincingly is much easier than modifying one, where multiple other values (perhaps related to even more themselves!) must be adapted correspondingly.

Such interdependency of attributes originates partly from redundancy, i.e. the existence of "superfluous" data. Because of this, "compressing" and removing all unnecessary data should be a first step before attempting any kind of obfuscation, as it reduces the need for matching adaptations. Note that this will not always be possible, as e.g. "removing" the gender from a first name is extremely difficult to impossible, and deleting the gender itself is problematic for unisex names (=loss of data, not only of redundancy). This also shows the problem more clearly: removing redundancy from one attribute is much easier than removing redundant overlap between several attributes, especially if the dependency count increases and only correlation exists, not duplication. For 3D data this e.g. means transforming a pointcloud into surfaces, potentially even flattening them, to obtain fewer but larger elements. Then only a single plane exists (pot. modifiable for obfuscation) instead of numerous interdependent points - where any change to one requires changing all of them.

A further problem is, that external data can be of help, too: separate knowledge about the typical elements/measurements/… of a duct can serve to assess the validity of generated or modified data, or as a replacement of missing (deleted, encrypted…) elements. See e.g. official regulations about fire extinguishers or warning signs: if all information from the measured data about them has been removed (photos of signs, 3D data from scans), the regulation might still tell us exactly (up to the height above the floor they must be mounted!) where they have to be and how many must exist. Combining these two elements then produces redundancy again: the actual locations can be "retrieved" from the regulation. In that case even complete deletion of any hint to extinguishers will not help: the redundancy is based on "this is a cable duct" in combination with "cable ducts must contain extinguishers arranged as follows…" (plus the assumption that the regulation was actually implemented). And removing the information that this *is* a duct will be impossible. Generating incorrect data about extinguishers would, in light of the regulation, be easy to detect as such, too.

As a major part of obfuscation is creating a believable alternative representation, AI might be of help here. However, it would have to be specially trained to be able to "recreate" convincing alternative elements of ducts. Specific problems here are:

- Typical AI solutions today produce content similar to the one they were trained on: If e.g. only modern concrete ducts were trained, it will fail miserable on old brick ones. Note that ducts may be regionally/nationally (regulations) or temporally quite different, and older ones may change their style at multiple points (renovations, extensions etc).
- AI is weak in generating exact elements, see e.g. picture generators and their problems with the number of fingers a hand should have. This is problematic as e.g. pipes and their structural holders are not vaguely circular and somehow metallic, but have very exact measurements along their entire length; e.g. L-profiles exist in few well-known sizes. Also, spacing between elements are not approximate but very regular. On the other hand, any discolorations on concrete are very irregular but must still match e.g. "fluid flowing down" (which is well generatable in lots of variations!).
- Complex context: A duct cannot be extended by "some" duct, but must be extended exactly as it is, i.e. an actual picture and many attributes of the duct would need to be passed into the generation as a huge context, or alternatively a separate model for each (potentially even stretch) of duct would have to be trained.
- Various results required: Some aspects to be (re)created are images (photos of duct), 3D point clouds/models, geographical coordinates, duct trajectories (separate from 3D models but related to them) etc. For images pre-trained models do exist, but for the others new systems would have to be developed. Also, some elements need to be transformed to a different kind: removing 3D data of objects and replacing it with a symbol at the same location (and recreating the "background" both in 3D and as a picture where the object was removed, so e.g. a scan shadow and a photo does not reveal it).

3. Practical application: obfuscating cable ducts

Obfuscation is exemplified based on the project INFRASPEC, which's aim is to automate the inspection of critical infrastructure, specifically cable ducts through a mobile robot. This robot (beside other functions) creates a 3D model of the cable duct, potentially including some external elements (e.g. as visible at the start of the duct through the entrance door). However, depending on the infrastructure operator, many aspects of these ducts might need to remain "anonymous", e.g. the specific location of the entrance or the exact path the duct takes. One aim of the project is therefore to investigate methods to obfuscate such data, so that parts can be disclosed for various uses: testing/training of improved robots, third companies performing special evaluations (e.g. searching for structural weaknesses), supervision authorities to whom the actual inspection has to be proven etc. Note that these entities might be perfectly trustworthy (e.g. the supervision authority will typically be some part of public administration and therefore strictly bound to security and confidentiality), but there may still be issues. Everyone can be hacked, but e.g. public administration may also be forced to give away data in response to public inquiries (freedom of information laws). While this probably excludes data endangering public safety or if it would compromise company secrets, this is not necessarily obvious and too much might be disclosed by non-experts providing the data. Simultaneously, removing the "problematic" bits must be implemented somehow.

3.1 Aspects to keep secret

The following aspects of ducts might need to be obfuscated, depending on circumstances. Not every aspect for every duct of every owner might require all, but these are at least interesting for some owners/ducts. These aspects can be seen as duct "attributes", which is the equivalent of a "row" (see above). Therefore, in some cases complete removal might be necessary, while sometimes reduction to a coarse characterization might be sufficient. Some less important elements have been omitted here for brevity (e.g. potential space in duct for expansions, usage data, locking details).

1. Location of entrance(s): From where is it possible to get to where, respectively to get into it at all to obtain access to a duct. This might be quite obvious (door in public place; but where does it lead to?), but could also be quite unobtrusive (canal cover as entrance, in the basement of a building). However, it might be difficult to completely hide entrances/exits, as several persons need to know (inspections) or obtain knowledge (building, repairs etc) of it. It is therefore especially dangerous if entrances can be searched for, i.e. "show me all ducts leading to that location", "all entrances in this area", or obtaining a complete list/plan.
2. Exact path of duct: Where exactly the duct is located in relation to the surface, e.g. where it crosses the street (interesting for terrorists to place a bomb) or where you would have to dig to reach it (e.g. within the cellar of a building). This cannot be determined from the locations of the entrances only, as ducts may turn, have corners etc. This includes the depth below surface, which might be even more difficult to ascertain without detailed measurements.
3. Size of duct, including free space: How large an object can fit through at the smallest location, or how much space there is for extending the use; e.g. is there a larger "room" somewhere or solely the duct? In most cases this will be of comparatively little importance to attackers, except perhaps squatters.
4. "Content" of duct, i.e. which cables (power, signal; kind etc) or pipes (size, content, insulation), other elements (sensors, fire extinguishers, valves...): This is of high interest to some attackers, as it determines the potential damage. This information can be quite independent of the duct, as e.g. generalized flow plans might show this information too. However, if these are not available, photos or measurements might allow good estimates (e.g. a cable of this size and insulation will typically be used for that voltage and can carry some maximum amperes; flammable warning signs signify fuel and not water/cooling pipes; etc).

3.2 Methods of obfuscation

A number of general methods of obfuscation exist. These are based on the concept of removing or replacing attributes as discussed above.

a) Randomization: Data is changed in a random way, which means it is impossible to reconstruct it later. Note that this is only valid for single elements ("rows"): e.g. randomizing a single number prevents reconstructing the original value. However, randomizing 1000 values will retain the original average of all numbers – which might be useful in itself. The value itself may be changed based on a random number generator or the same attribute from multiple rows could be randomly exchanged.

b) Deviation: The actual data is modified in a deliberate way to lead to different conclusions, e.g. the trajectory of a plane is moved to point to a different target airfield. Note that it must be verified that this alternative data must still be possible *and* plausible (plane example: flying through a mountain is unlikely, so a path around the – on the actual trajectory not present – mountain must be created; also adapting fuel consumption, flight duration etc).

c) Deletion: Important attributes can be simply removed from the data, as long as these are irrelevant for the obfuscated data's envisaged use.

d) Pseudonyms: Data can be replaced by other data. The advantage is, that it potentially can be undone if a translation list is kept (separately for security). Unlike deviation the replacement need not be convincing – but the existence of some data at that point cannot be hidden (and the location could provide from none at all to perfect hints of what data originally existed there).

e) Generalization: The removal of details or precision may retain usefulness of the data for certain purposes, but not anymore for others. Potentially problematic is, that the specific use needs to be known in advance, as depending on it the method of generalization must be designed. An example is replacing detailed objects (e.g. 3D scans) with a symbol (flat icon and "hole" in data).

f) Aggregation: Multiple elements may be combined to prevent accessing them individually. This can e.g. be an average of some data points. This could be seen as a special case of generalization, but the difference is, that generalization retains each data point, but renders it less detailed, while aggregation combines several data points to a single new one, reducing their count.

g) Generation: Additional data can be generated and inserted at various locations. This may prevent matches to reality, as e.g. not enough physical space is available for the duct at the claimed location or fewer doors/lamps/… exist in reality, rendering comparisons/locating specific points difficult. Also angles or curves of ducts may be extended, e.g. increasing the turn ratio.

h) Encryption: Whether this is still "obfuscation" is debatable, but data can obviously be encrypted, sometimes also only parts of it. In the latter case this is similar to a combination of generalization ("removing" important details) and pseudonyms (with some additional data – the key – the original information can be recovered). In its encrypted form it is useless to attackers, but it can still be determined that some data exists (at that location), and some estimate about the size.

Some of these approaches can also be performed across several attributes, which can be of interest too, depending on the data, e.g. because of their interdependence:

a) Pseudonyms: replacements can incorporate various elements, which also in reality might be quite independent and unrelated.

b) Aggregation: E.g. trays and their content (cables/pipes) can be combined into a single "solid block". Typically, their content would not be completely independent of each other: e.g. for power lines all three cables are identical and pipes often occur in tandem (closed circuit). In extremity (not useful in this project) a duct could be aggregated to "these conduits go in here and these leave there", removing all data about the inside of the duct.

c) Generation: To be convincing, e.g. extending the length of a duct requires filling the space (simply stretching the data would never be convincing) with additional sections of pipes (easy, extending existing ones; single attribute), but also multiple instances (in the correct distance/location) of holders/vents/signs/valves/security equipment/lights etc.

d) Encryption: Attributes can be encrypted individually or as a group; the first allows selective disclosure but may entail overhead, the second is more efficient but an all-or-nothing approach.

4. Applying the methods to 3D data: an example

Only one element is described here in detail as an example, the exact path of ducts. Modifying the path of a duct is possible in several general approaches: first, parts of the duct can be removed. This is potentially problematic, as the data will very likely be needed during future inspections for comparisons. However, should there be e.g. some intermediate rooms or doors, these could be removed. Depending on the scan precision a seam might be detectable, so (small) modifications to prevent this, like smoothing the join as well as pictures seams, might be required. Another approach is to straighten the duct (see “deviation” and “generation” above): wherever there is a curve or angle in the duct trajectory, this could be straightened and missing elements be generated by “continuing” the elements from before/after. This is not trivial, as while e.g. pipes can easily be extended, their fastening/supports must be added too, and in a sensible distance (regular intervals in a straight duct; when straightening a curve these on the inner side will be enlarged, and when filling an angle the last one might not match up exactly). The advantage of this approach is, that no additional metadata to “undo” the modifications is needed: the next scan introduces the same changes before comparison – or these parts could simply be removed. Only when performing off-site analysis such metadata would be necessary (easily separated and stored elsewhere).

Theoretically possible is randomization: parts of the duct could be swapped. While this is unlikely to work across ducts (different size or conduits), for a single duct it could be possible, especially modern ducts of similar shape in their whole path. Problematic will be again conduits: ones entering/leaving produce differences. Mirroring the whole duct is also possible, but probably of little use (and problematic with all kinds of signs/texts or e.g. arrows signifying the turn direction of valves!).

Another approach is to obfuscate entrances: Usually, the size of a duct will be similar along it, so also the entrance areas will be closely resembling each other. Therefore, entrances themselves and their surroundings could be switched (often of little importance in inspections too). The drawback is the very limited choice: many ducts have only two entrances: these could remain correct or be switched, giving attackers only two options (=50% guessing chance) – little improvement.

Another approach to hide the path is to generate additional space in a duct. Many ducts are – at least in some areas – straight and uniform. These sections can be comparatively easily extended. Note that other than filling missing edges when straightening, any regular structures can be accounted for: the periodicity needs to be determined and a (random) multiple of this inserted at a random position (synchronized, e.g. immediately after a support pole starting with a new section and ending with such a pole; problematic if multiple such elements exist). Imperfections, like small defects, can and should be invented, too. For 3D data this is easy (but purely random variations might not match with laws of physics), but for pictures creating a non-identical but matching surface can be more difficult (direct copying is easily detectable). Another hurdle are panoramic photos, i.e. pictures taken along the path of the duct and not projected onto the 3D surface, as then such pictures would also have to introduce additional elements to convincingly insert data. This approach of storing only the surface data and not any panoramas is easier to implement. Unlike the “straightening” above, metadata for removing added sections on later comparison is needed.

Another approach is randomizing sections of a duct. This can be performed in aspects irrelevant (needs to be verified individually – some aspects may be important for one duct but not for another) for inspection, e.g. the slope: while most ducts are flat, slopes can be introduced easily (note that supports still need to be vertical and pipes must be rotated, not skewed).But this changes the path only slightly - and might be irrelevant for attackers anyway. For improvements the level would have to be changed significantly, but this means the difficult insertion of vertical shafts (+ stairs, pipe bends, special supports etc). Generalization as discussed before can help: if specific cable trays, pipes etc are replaced by more generic version, such modifications get much easier.

A variant of deletion is separation: a duct with crossings or splits can be divided into two or more ducts. If the entrance is removed, this is undetectable anyway, otherwise a fake/copied one needs to be supplied for the cut-off part. For the “main” duct this can be more problematic, as either the part with the branching must be removed or the gap again filled as described above. However, as typically some pipes/cables/… will enter the branch but not emerge from it, a difference in conduits between before and after the branch exists, which cannot easily be hidden. Still, even if a gap is left, the exact angle of the branching duct can be hidden, if the branch is cut to e.g. a right angle.

Another approach for certain aspects is to remove all photos except where no (beginning) problem was detected: any problem found in a subsequent inspection can be compared to the previous photo or is new. This only works for elements which can be detected as "problematic" from a single scan. Everything needing a comparison (e.g. the ceiling has lowered → hint at stability problem) cannot be covered in this way. I.e. the previous value may only serve as decision base for the increase in the problem, but not for recognizing the problem in the first place. This is therefore potentially suitable in this project for surfaces, as accretions or pipe/cable defects are recognizable individually. Defects of interest here produce slightly different data (pipes: leaks lead to different temperature or discoloured stripes/drips/puddles/…; cables: localized temperature anomaly, physical damage).

High-level generalization is of course possible by replacing the details of the duct with a diagram of the conduits there – but this is contrary to the aim of the project of the inspection. Generalization would therefore have to be on a much smaller level, but it is unclear how this would help with obfuscating the path of the duct.

A final option (related to separation above) is to split a duct into multiple sections, e.g. whenever a corner is found, and store these separately. If there are many ducts and many corners, re-aligning them might be difficult, but based on the dimensions and the number, size, and position of conduits this is probably in many cases still possible. Additionally, any photos of the wall at each end can help: they do not show the other part, but the surface must be very similar: it in unlikely that the duct is e.g. concrete up to the corner and then continues in bricks.

Recommendations for obfuscation:

- Removal of unnecessary details. Their definition and recognizing them can be challenging. But e.g. doors could probably be a feasible target. AI tools may be helpful in this respect.
- Splitting the duct at every corner, ideally coupled by separating out the conduits as well to render reassembly more difficult. Here a separate (keyed) mapping of which elements belong together in which sequence are necessary, so later on comparisons remain possible.
- Removal of all photographic data that is "normal" like a human would look at things, limiting pictures to the area of (potential) problems and its immediate surrounding.

5. Summary

The difference between data obfuscation and anonymization was discussed and various methods of obfuscation outlined, including randomization, aggregation, replacing elements, using pseudonyms, etc. Challenges for employing AI tools were discussed too. However, these are unlikely to be very useful in the specific context of cable ducts employed as an example here.

A specific challenge of obfuscation is the interdependence of data elements, which makes guessing/recreating removed elements much easier unless not only all traces of the element itself, but also all related elements are removed/obfuscated too. External knowledge or data aids this process too, and it is difficult to determine which of these attackers will know/have access to. It is therefore very important to remove redundancy and reduce interdependence in the data as much as possible before the obfuscation stage to minimize the need for adaptations between multiple attributes.

The application of these obfuscation methods to 3D data, specifically in the context of the INFRASPEC project focusing on cable duct inspection, was detailed based on the general criteria and here applied to a single example, hiding the exact path of a duct, via the various methods. However, even though some recommendations could be made, the problem remains that obfuscating the path is at least with current technology not really feasible. As a replacement in the project therefore calculation in a secure method/location was designed and implemented as an example: using SGX technology, data can be calculated on in cloud services with a guarantee (absent security issues) that even the cloud operator cannot read the input data or see results. So, while obfuscation is desirable, at least for complex 3D data it is only possible to a very limited degree.

Acknowledgement

This publication and the system described in it was developed within the scope of the project IN-FRASPEC, which is funded by the Austrian security research program KIRAS (FFG) of the Feder-al Ministry of Finance (BMF).

References

Ardagna, C. A., Cremonini, M., De Capitani di Vimercati, S., Samarati, P. (2009). An Obfusca-tion-Based Approach for Protecting Location Privacy, IEEE Transactions on Dependable and Se-cure Computing, vol. 8, no. 1, 13-27. DOI: 10.1109/TDSC.2009.25

Éluard, M., Maetz, Y., Doërr, G. (2013). Geometry-preserving Encryption for 3D Meshes. CORESA 2013 - 16ème édition du colloque COmpression et REprésentation des Signaux Audio-visuels, 7-12, https://www.researchgate.net/profile/Gwenael-Doerr/publication/273257218_Geometry-preserving_Encryption_for_3D_Meshes/links/

54fc4b660cf2c3f52422a624/Geometry-preserving-Encryption-for-3D-Meshes.pdf

Gupta N., Chen F., Tsoutsos N., Maniatakos, M. (2017). ObfusCADe: Obfuscating additive manu-facturing CAD models against counterfeiting, 54th ACM/EDAC/IEEE Design Automation Confer-ence. (DAC), 2017, 1–6. DOI: 10.1145/3061639.3079847.

Jolfaei, A., Wu, X.-W., Muthukkumarasamy, V. (2015). A 3D Object Encryption Scheme Which Maintains Dimensional and Spatial Stability, IEEE Transactions on Information Forensics and Security, vol. 10, no. 2, 409-422. DOI: 10.1109/TIFS.2014.2378146.

Kada, M., Peter, M., Fritsch, D., Siemoneit, O., Hubig, C. (2009). Privacy-enabling abstraction and obfuscation techniques for 3D city models. Proceedings of the 2nd SIGSPATIAL ACM GIS 2009 International Workshop on Security and Privacy in GIS and LBS (SPRINGL '09). ACM, New York, 53–57. DOI: 10.1145/1667502.1667515

Koller, D., Levoy, M. (2005). Protecting 3d graphics content. Commun. ACM 48, 6 (June 2005), 74–80. DOI: 10.1145/1064830.1064861

Lison, P., Pilán, I., Sánchez, D., Batet, M., Øvrelid, L. (2021). Anonymisation Models for Text Data: State of the art, Challenges and Future Directions. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing 2021, 4188-4203. DOI: 10.18653/v1/2021.acl-long.323.

Naeem, F., Mohsin, M., Rauf, U., Ali Khan, L. (2021). Formal approach to thwart against drone discovery attacks: A taxonomy of novel 3D obfuscation mechanisms, Future Generation Comput-er Systems, Volume 115, 2021, 374-386. DOI: 10.1016/j.future.2020.09.001.

Sun, J., Wang, J., Nie, W., Yu, Z., Mao, Z. & Xiao, C. (2023). A Critical Revisit of Adversarial Robustness in 3D Point Cloud Recognition with Diffusion-Driven Purification. Proceedings of the 40th International Conference on Machine Learning, In Proceedings of Machine Learning Re-search, DOI: 202:33100-33114

Tabet, S., Kayssi, A. Elhajj, I. H. (2023). Mobile Diminished Reality for Preserving 3D Visual Privacy, 2023 International Conference on Intelligent Metaverse Technologies & Applications (iMETA), 2023, 01-07. DOI: 10.1109/iMETA59369.2023.10294436.

Zhang, J., Dong, Y., Kuang, M., Liu, B., Ouyang, B., Zhu, J., Wang, H., Meng, Y. (2023). The Art of Defense: Letting Networks Fool the Attacker, IEEE Transactions on Information Forensics and Security, vol. 18, 3267-3276, 2023. DOI: 10.1109/TIFS.2023.3278458References should be listed alphabetically at the end of the contribution.

SECURITY RISKS ASSOCIATED WITH DEPLOYMENT OF AI SOLUTIONS INTO ORGANIZATIONS

Martin Zbořil

Prague University of Economics and Business
Faculty of Informatics and Statistics
zbom01@vse.cz

DOI: 10.35011/IDIMT-2024-65

Keywords

Artificial intelligence; Security; Risks; Copilot

Abstract

In the current digital era, integration of Artificial Intelligence (AI) into enterprise operations has become a leading strategy for enhancing efficiency, decision-making, and competitiveness. AI technologies, notably Microsoft Copilot, offer transformative potential but also introduce significant security challenges. This paper examines these challenges, identifying key security risks associated with enterprise-wide AI deployment and offering strategic mitigation recommendations. Through a comprehensive literature review, the paper explores adversarial attacks, data privacy, model theft, bias, transparency, regulatory compliance, and the implications of insufficient data governance. The analysis reveals that proactive, comprehensive security measures and adherence to evolving regulations are critical in leveraging AI's benefits while safeguarding against its risks. This research contributes to a deeper understanding of AI security in organizational contexts, proposing a framework for enterprises to navigate the complex landscape of AI integration securely.

1. Introduction

In the nowadays digital era, the deployment of Artificial Intelligence (AI) solutions across enterprise operations is increasingly becoming a strategic movement. These technologies, exemplified by advanced platforms like Microsoft Copilot, Claude, ChatGPT or Google Assistant, have the potential to redefine organizational efficiency, decision-making processes, and competitive dynamics. Alongside the considerable advantages these technologies offer, they introduce a variety of security considerations that organizations must consider. This paper delves into these considerations, focusing on the associated security risks and proposing recommendations to mitigate them effectively.

Traditional artificial intelligence, as defined by Simmons and Chappell (1988), is *"behavior of a machine which, if a human behaves in the same way, is considered intelligent."* When considering the current and extended definitions, European Commission's Communication (2018) on AI defines AI as: *"Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in* *hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).”* In the context of enterprise applications, AI encompasses a range of technologies capable of performing tasks that typically require human intelligence, such as learning, decision-making, and problem-solving. This includes machine learning algorithms, natural language processing, robotics, and computer vision among others (Helm et al., 2020). The adoption of AI in organizational aims to enhance operational efficiency, improve decision-making accuracy, and foster innovation by automating complex processes and analyzing vast datasets (Kinkel et al., 2022).

The integration of AI into enterprise-wide systems raises significant security concerns that span across data privacy, system integrity, and operational reliability. As AI technologies access, process, and store vast amounts of sensitive data, they become prime targets for cyberattacks, including data breaches, AI-driven malware, and phishing attacks (Raval et al, 2024; Prox, 2023). Furthermore, the complexity of AI systems can introduce new vulnerabilities, while their autonomous nature might obscure potential threats, complicating the security management landscape (Berghoff et al., 2020).

In response to AI risks, several frameworks have emerged to guide secure AI deployment. The EU AI Act sets stringent requirements for high-risk AI systems, focusing on safety and rights protection. The NIST AI Risk Management Framework offers principles for managing AI risks, including transparency and accountability. Additionally, MITRE ATLAS provides insights into adversarial threats against AI systems, helping organizations bolster the defenses. These frameworks collectively emphasize a comprehensive approach to AI security and risk management (Hannecke, 2023)

Adding to these challenges is the rapid pace of AI advancement and its adoption across various sectors, overtaking the development of corresponding security measures and legal frameworks. This discrepancy not only intensifies existing security vulnerabilities but also introduces new risks, particularly as AI systems become more autonomous and capable of decision-making processes that were traditionally human domains (Jia and Zhang., 2022). The potential for AI to be exploited in sophisticated cyberattacks or to unintentionally compromise data integrity through decision-making errors underlines the urgent need for a proactive and comprehensive approach to AI security. Thus, this paper aims to not only identify and analyze the security risks associated with enterprise-wide AI deployment but also to outline strategic measures that can be adopted to enhance security postures in the face of AI integration.

2. Methodology

This research adopts a comprehensive literature review methodology, especially analyzing academic papers, industry reports, case studies, and other relevant documents. The aim is to synthesize current knowledge and insights on the security considerations of deploying enterprise-wide AI solutions, focusing on identifying associated risks and formulating recommendations for mitigating these risks.

The data collection process is structured around a systematic review of literature spanning various databases and repositories, including IEEE Xplore, ACM Digital Library, ScienceDirect, and Google Scholar, among others. The search strategy encompasses a combination of keywords and phrases related to AI security risks, enterprise AI deployment, AI vulnerabilities, and cybersecurity measures in AI systems. This approach ensures a comprehensive coverage of relevant literature, capturing both broad trends and specific instances of security considerations within the context of AI deployment. To ensure the inclusion of the most current insights and data, the literature search is limited to documents published within the last five years. This time frame is chosen to reflect the rapid evolution of AI technologies and the corresponding shifts in security landscapes. Additionally, industry reports and white papers from leading technology firms and cybersecurity organizations are reviewed to incorporate practical perspectives and real-world examples into the research.

The analysis of collected data employs a thematic analysis approach, aiming to identify recurring themes, patterns, and insights across the selected documents. This process involves an initial reading to gain an overview of the literature, followed by a more detailed examination to code the data into thematic categories related to AI security considerations, risks, and mitigation strategies.

Critical analysis is applied to evaluate the quality of the sources, their methodologies, and the relevance of their findings to the research questions. This evaluation is crucial for distinguishing between well-supported insights and less substantiated claims within the literature.

The synthesis of findings from academic research and industry practices enables a comprehensive understanding of the security challenges associated with enterprise-wide AI deployment. It highlights emerging risks, documents effective strategies for risk mitigation, and identifies gaps in current knowledge, pointing to areas where further research is needed. Importantly, this analysis includes recommendations for mitigating these risks, specifically tailored to the deployment of Microsoft Copilot (for the risks where is a direct relevance to Microsoft Copilot). These recommendations are grounded in the reviewed studies and the author's extensive experience in the cybersecurity field. The source for the Microsoft Copilot recommendations is the author personal experience from implementation and assessments of Microsoft technologies. By incorporating practical advice for securing AI implementations like Microsoft Copilot, this research not only elucidates the theoretical landscape of AI security risks but also offers actionable guidance for practitioners aiming to safeguard their AI deployments effectively. This dual focus ensures that the paper provides a holistic view of AI security, blending academic insights with real-world applicability.

3. Research

The usage of AI in enterprise operations has significantly altered the landscape of organizational efficiency and innovation. However, this integration has simultaneously exposed organizations to a numerous security challenges. This section delves into the primary security risks inherent in deploying AI solutions and proposes targeted recommendations to mitigate these risks effectively. The table below brings an overview of identified security risks connected with deployment of enterprise-wide ai solution into organizations.

Table 1. Overview of identified security risks

ID	Risk	Used references
1	Adversarial attacks	(Jia, 2022), (Raval, 2024), (Nigro, 2023), (Mirsky, 2023), (Wang et al., 2019), (Xiong et al., 2021), (Zaman et al., 2021), (ENISA, 2023)
2	Insufficient data governance leading to data leakage	(Jia, 2022), (Kalodanis, 2023), (Mirsky, 2023), (Zaman et al., 2021), (Cool et al., 2021)
3	Data privacy and Integrity	(Jia, 2022), (Nigro, 2023), (Targeted News Service, 2023), (Kalodanis, 2023), (ENISA, 2023), (Pillsbury, 2021), (Cool et al., 2021)
4	Model theft and reverse engineering	(Mirsky, 2023), (Xiong et al., 2021)
5	Bias and fairness	(Jia, 2022), (Raval, 2024), (Xiong et al., 2021)
6	Lack of explainability and transparency	(Raval, 2024), (Targeted News Service, 2023), (ENISA, 2021), (ENISA, 2023), (Pillsbury, 2021)

ID	Risk	Used references
7	Compliance and regulatory risks	(Jia, 2022), (Nature, 2023), (Nigro, 2023), (Kalodanis, 2023), (ENISA, 2023), (Pillsbury, 2021), (Cool et al., 2021)
8	Robustness and physical harm	(Jia, 2022), (Raval, 2024), (Content Engine LLC, 2023), (ENISA, 2021), (Wang et al., 2019), (Xiong et al., 2021), (Zaman et al., 2021)

The author further provides description of each risk and exemplary recommendations on how to mitigate the risk.

Risk 1: Adversarial attacks

Adversarial attacks represent a sophisticated challenge where attackers manipulate AI input data to induce incorrect model decisions. These manipulations exploit model vulnerabilities without detection, posing risks across various applications. Neglecting this risk compromises the reliability and safety of AI-driven decisions, especially in sectors like healthcare and transportation, leading to potential harm and eroding public trust in AI.

Recommendations on remediation:

- Enhancement of AI resilience through adversarial training, incorporating a diverse set of attack scenarios.
- Regular model updates and employing detection mechanisms for unusual input patterns.
- A layered security approach, combining physical and digital defenses.

Recommendation for secure deployment of Microsoft Copilot:

- Full-scope deployment of Microsoft XDR (Extended Detection and Response) service and Microsoft Copilot for Security that supports organizations in detection of attackers in organizational environments, protection against them and handling the security incidents.

Risk 2: Insufficient data governance leading to data leakage

Effective data governance is essential for managing access, storage, and use of data in AI systems. Insufficient governance can lead to data leakage, exposing sensitive information and compromising data privacy and integrity. Data leakage risks damaging customer trust, incurring regulatory penalties, and undermining the security of AI systems. It poses significant reputational and financial risks to organizations, emphasizing the need for strict data management practices.

Recommendations on remediation:

- Establish comprehensive data governance policies that address data classification, access controls, encryption, and incident response.
- Conduct regular security audits and implement a principle of least privilege for data access.
- Train employees on data security best practices.

Recommendation for secure deployment of Microsoft Copilot:

- Deployment of the Microsoft Purview service that offers robust data governance tools that ensure data discovery, classification, and protection, crucial for AI deployments like Microsoft Copilot. With Purview, organizations can enforce data policies, control access, and monitor activities, effectively preventing unauthorized access and data leaks, thereby enhancing their cybersecurity framework and compliance.

Risk 3: Data privacy and integrity

From a certain angle, this risk is a special area of the previous one. The integrity and confidentiality of data used by AI are paramount. Risks include unauthorized data access and manipulation, potentially leading to biased outcomes or privacy breaches. These risks are magnified by the volume and sensitivity of data AI systems process. Ignoring these risks can lead to legal consequences, financial losses, and damage to reputation. Breaches may result in sensitive data exposure, undermining user confidence and leading to regulatory penalties under laws like GDPR.

Recommendations on remediation:

- Adopting a robust data governance framework ensures data is handled securely throughout its lifecycle.
- Encryption, rigorous access controls, and regular audits help safeguard data integrity.
- Promoting a culture of security awareness among all employees further mitigates risks.

Recommendation for secure deployment of Microsoft Copilot:

- See recommendations from Risk 2 with special focus on the data privacy and integrity.

Risk 4: Model theft and reverse engineering

The unauthorized replication/analysis of AI models can lead to intellectual property theft and expose security vulnerabilities. This risk is heightened by the increasing sharing and distribution of AI. Overlooking this risk can revoke competitive advantages and expose users to risks if vulnerabilities are exploited. It threatens the proprietary value of AI innovations and may result in financial and reputational damage.

Recommendations on remediation:

- Leveraging techniques like model obfuscation and securing AI model storage with advanced cybersecurity measures protect against theft.
- Legal protections, such as copyrights and patents, provide additional safeguards.
- Employee education on the importance of intellectual property security.
- Extract the training data from the final model.

Recommendation for secure deployment of Microsoft Copilot:

- Protection of stored objects of AI models through implementation of data protection measures (especially with the usage of the Microsoft Purview service) over those objects to ensure fulfilment of the traditional information protection measures – their confidentiality, integrity and availability, where especially the first two requirements are crucial.

Risk 5: Bias and fairness

AI systems can inadvertently learn biases from their training data, leading to unfair or discriminatory outcomes. This risk is particularly prevalent in systems making decisions affecting individuals, such as hiring tools or loan approval algorithms. Not addressing bias and fairness can harm individuals and groups, erode public trust in AI technologies, and expose organizations to legal and regulatory challenges. It undermines the ethical use of AI and can lead to societal backlash.

Recommendations on remediation:

- Diversify training datasets to reflect a broad range of perspectives and reduce inherent biases.

- Implement fairness audits and bias detection mechanisms throughout the AI development lifecycle.
- Encourage transparency and ethical AI use guidelines within the organization.

Recommendation for secure deployment of Microsoft Copilot:

- Enablement and proper configuration of the Communication Compliance module as a part of the Microsoft Purview service.

Risk 6: Lack of explainability and transparency

AI's "black box" nature can obscure how decisions are made, making it difficult for users and stakeholders to trust and understand AI outputs. This lack of transparency can hinder accountability and complicate error correction. Ignoring explainability and transparency may push users away and may lead to regulatory scrutiny, especially in industries where understanding AI decision-making is critical. It can also impede the ability to diagnose and correct flawed AI behaviors.

Recommendations on remediation:

- Invest in explainable AI research and tools that shed light on AI decision-making processes.
- Develop documentation and user guides that clearly communicate how AI systems operate.
- Foster a culture of openness, prioritizing user education and stakeholder engagement.

Risk 7: Compliance and regulatory risks

The lack of explainability and transparency has significantly negative impact on how the AI solutions are perceived from the compliance and regulatory risk perspectives. Following this, AI deployments must navigate a complex landscape of global and regional regulations, which can vary significantly and change rapidly. Compliance ensures legal operation but requires continuous observation and adaptation. AI-related regulations should advise organizations on how to deal with the deployment of new AI solutions into orgaztions. Non-compliance can lead to legal penalties, operational disruptions, and damage to reputation. It can also inhibit the ability to enter or compete in markets with stringent regulatory requirements, limiting growth opportunities.

Recommendations on remediation:

- Establish a dedicated team to monitor regulatory developments and assess their impact on AI.
- Incorporate regulatory considerations early in the AI development process.
- Engage with legal experts to understand and implement necessary compliance measures.
- When AI-related regulations are effective, follow requirements of those regulations.

Recommendation for secure deployment of Microsoft Copilot:

- Even though Microsoft declares compliance with many regulations, it is important to realize that based on the shared responsibility model, organizations are compliant with those regulations only when both parts are compliant – the services operated by Microsoft and the solution/content built on it by the organization. Microsoft offers several supportive services that simplify the provision of data privacy and integration, such as Microsoft Purview, Compliance Manager, Microsoft Priva, Microsoft Defender for Cloud or Azure Policy.

Risk 8: Robustness and physical harm

AI systems must perform reliably under diverse conditions without causing harm. This is crucial in applications with direct physical implications, such as autonomous vehicles or healthcare diagnostics, where failures can have dire consequences. Failing to ensure the robustness of AI systems can result in physical injuries, financial losses, and loss of life, leading to legal liabilities and a breakdown of public trust in AI technologies.

Recommendations on remediation:

- Prioritize the development of safety-critical AI systems with rigorous testing and validation.
- Implement redundancy and fail-safe mechanisms to mitigate potential failures.
- Foster a safety-first culture within AI development teams.

4. Discussion

The deployment of AI solutions like Microsoft Copilot within enterprise operations presents a double-edged sword. Based on the available versions of Microsoft Copilot, public explanation on how they work and what value they bring, the author of this paper sees significant benefits in terms of operational efficiency, decision-making, and competitive advantage, it also introduces a spectrum of security risks that require meticulous attention. This paper has identified and analyzed critical security considerations, including data privacy and integrity, model theft, bias and fairness, transparency issues, compliance challenges, and the risk of insufficient data governance leading to data leakage.

The recommendations provided emphasize the importance of robust security practices, the adoption of comprehensive data governance frameworks exemplified by tools like Microsoft Purview, and the need for continuous vigilance and adaptation to evolving cybersecurity landscapes. Moreover, it underscores the shared responsibility between AI service providers and organizations to ensure the secure deployment and operation of AI technologies.

As AI continues to evolve and permeate more deeply into the fabric of organizational operations, the imperative for a proactive and informed approach to security becomes increasingly critical. By embracing the recommendations outlined in this paper, organizations can navigate the complexities of AI integration, safeguarding their operations and data against potential threats while unlocking the full potential of AI to drive innovation and growth. The journey toward secure AI deployment is ongoing, demanding a commitment to excellence, adaptability, and collaboration across the cybersecurity ecosystem.

Future work on this paper could explore emerging technologies and approaches for enhancing AI security, such as the impact of quantum computing on AI vulnerabilities and defenses. Additionally, research into AI models with self-diagnosing and adaptive capabilities against threats offers a promising direction. As legislative frameworks around AI evolve, further analysis on compliance and its implications across sectors will be vital. This progression will expand upon the current foundation, addressing advanced challenges and opportunities in AI security to equip organizations for a rapidly evolving digital landscape. From another perspective, this paper might be extended by explanation of more detail recommendations for risks mitigation, especially from the angle of how Microsoft technologies work. Unfortunately, the length limitation for this paper has not allowed to involve it in this paper.

Acknowledgement

Paper was processed with support from institutional-support fund for long-term conceptual development of science and research at the Faculty of Informatics and Statistics of the Prague University of Economics and Business (IP400040) and the project IG 409033.

References

Berghoff, Christian, Matthias Neu and Arndt von Twickel. Vulnerabilities of Connectionist AI Applications: Evaluation and Defense. Frontiers in big data. 2020, vol. 3

Cool, Michaem, Deborah Evans and Jennifer Medburry. Artificial Interlligence in Security: Opportunities and Implications. 2021. Available at: https://www.asisonline.org/globalassets/foundation/documents/digital-transformation-series/ai-guidance-document-final.pdf

Content Engine LLC, Translated by. 'Catastrophic risks': At AI security summit, experts issue this warning. 2023.

ENISA. Cybersecurity challenges in the uptake of artificial intelligence in autonomous driving. 2021. Available at: www.enisa.europa.eu/publications/enisa-jrc-cybersecurity-challengesin-the-uptake-of-artificial-intelligence-in-autonomous-driving

ENISA. Cybersecurity of AI and Standardisation. 2023. Athens. Available at: https://www.enisa.europa.eu/publications/cybersecurity-of-ai-and-standardisation

Hannecke, Michael. AI Security Frameworks and Guidelines. 2023.s Available at: https://medium.com/bluetuple-ai/outsmarting-the-smart-guys-a-list-of-ai-security-frameworks-and-guidelines-5d126569a2fa

Helm, J. M. et al. Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions. Current reviews in musculoskeletal medicine. 2020, vol. 13, no. 1, pp. 69–76.

Jia, Kai and Nan Zhang. Categorization and eccentricity of AI risks: a comparative study of the global AI guidelines. Electronic markets. 2022, vol. 32, no. 1, pp. 59–71.

Kalodanis, Konstantinos, Panagiotis Rizomiliotis and Dimosthenis Anagnostopoulos. European Artificial Intelligence Act: an AI security approach. Information and computer security. 2023

Kinkel, Steffen, Marco Baumgartner and Enrica Cherubini. Prerequisites for the adoption of AI technologies in manufacturing – Evidence from a worldwide sample of manufacturing companies. Technovation. 2022, vol. 110.

Mirsky, Y., et al. (2023), “The threat of offensive AI to organizations”, Computers and Security, Vol. 124, doi: 10.1016/j.cose.2022.103006.

Nature. It's time to talk about the known risks of AI. London. 2023, vol. 618, no. 7967, pp. 885–886.

Nigro, Pam. AI Security Risks: Separating Hype From Reality. 2023. ISBN 0890-8826.

Pillsbury. Artificial Intelligence & Cybersecurity. 2023. Available at: https://impact.economist.com/perspectives/sites/default/files/report_ai_cybersecurity_sept_2021.pdf

Prox, Ryan. Data & Infrastructure Security: The Risk of AI Enabled Cyber Attacks and Quantum Hacking. The journal of intelligence, conflict, and warfare. 2023, vol. 5, no. 3, pp. 117–121.

Raval, Khushi J. et al. A survey on safeguarding critical infrastructures: Attacks, AI security, and future directions. International journal of critical infrastructure protection. 2024, vol. 44, p. 100647.

Simmons, A. B. and S. G. Chappell. Artificial intelligence-definition and practice. IEEE journal of oceanic engineering. 1988, vol. 13, no. 2, pp. 14–42.

Targeted News Service, MITRE and Microsoft Collaborate to Address Generative AI Security Risks. 2023.

The European Commission. A Definition of AI: Main Capabilities and Scientific Disciplines. 2018. Available at: https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf

Wang, Xianmin et al. The security of machine learning in an adversarial setting: A survey. Journal of parallel and distributed computing. 2019, vol. 130, pp. 12–23.

Zaman, Shakila et al. Security Threats and Artificial Intelligence Based Countermeasures for Internet of Things Networks: A Comprehensive Survey. IEEE access. 2021, vol. 9, pp. 94668–94690.

Xiong, P., et al. (2021), Towards a Robust and Trustworthy Machine Learning System Development, National Research Council of Canada, Ontario.

REDEFINING THREATS: EXTENDING THE THREAT RESPONSE FOCUS FROM EXTERNAL TO INTERNAL THREATS

Jiri Holoska, Petr Doucek

Prague University of Economics and Business
Faculty of Informatics and Statistics,
jiri.holoska@vse.cz, doucek@vse.cz

DOI: 10.35011/IDIMT-2024-73

Keywords

Internal Threats; ICT Security; Digital Evidence; Insider Incident

Abstract

The objective of the contribution is to analyze the growing danger of insider threats and their impact on the security of company information systems and to show the select ways of their potential solutions in the Czech Republic and the Slovak Republic. Three basic approaches and thus also data sources were used for this article. The first source includes global quantitative data from large corporations identified for the time period of 2016-2022. The second source includes conclusions from a range of commercial and scientific projects in which the authors have been involved. The third source includes professional literature and documents concerning information security in organizations. The percentage and significance of insider incidents in current corporate practice, especially in large organizations, keeps increasing, as does the motivation to investigate these incidents, as the presented trends from the investigated time period. Also, the time to detect and resolve an insider incident is relatively long. Organizations can use the tools to protect themselves against insider incidents. Two limitations are crucial for the article. The data and conclusions are valid for large corporations. Their applicability to SMEs and micro-firms is not great. The analyzed input data are not a representative sample of the individual analyzed regions' economies.

1. Introduction

There are two basic types of threat in terms of organizations' ICT security – internal and external (ISO/IEC 27005). Information and data protection departments in organizations usually focus on external threats. But the recent trend also points out the serious risk of insider threats, especially from own employees (Catrantzos, 2022). Information services are among the first organizations to identify the impact of insider threats globally, even beyond ICT (ICT). Organization infiltration and data loss were powerful enough to disable the operational capabilities of an organization in a given region or put the lives of agents at risk (Hellberg, 2013). Organizations involved in technological innovations, particularly in the development of technical products and military technology, soon after identified insider threats as well. This can be proven by the fact that various plagiarized products and their original counterparts visibly have common features, e.g. some newly developed drones or submarine documentation (Bing, 2018). The F-35 aircraft (Gady, 2015) provides further evidence of data leakage by employees. These documents appeared on the dark web and were traded.

The development of mobile technologies, such as smartphones, notebooks and personal cloud services, has helped to expand this trend to other organizations across the entire technological spectrum. Users, who gave or received notice, copy entire folders of documents concerning the projects on which they worked or which – in their opinion – will give them a competitive advantage in their new organization.

Insider threats are not limited to data loss per se, but can become purely destructive (Payne, Hanson and Wojtasiak, 2022). Frustrated privileged users may one day decide to make unauthorized changes in the corporate or production infrastructure in order to vent their frustration, to point out unresolved problems or to damage their company's reputation (Press Release, 2022). In particular, potable water treatment and supply and fuel and electricity distribution are vulnerable to insider threats (Abrams and Weiss, 2008).

The leakage of intellectual property is another area where insider threats may have serious consequences. It can be devastating for a company in terms of its competitiveness. Years of research can be leaked to competitors in a matter of days and its competitive advantage in the market is gone overnight. Stock market losses can wipe out an organization’s years of work and a long profitable period in a matter of months (Bunn and Scott, 2017).

Former Ubiquity employee Nickolas Sharp, who stole 10 GB of confidential data, is an example of such behavior. He then tried to extort the company anonymously and threatened to release the documents unless he was paid 1.9 million USD. Ubiquity decided not to pay the ransom and began cooperating with the FBI. The investigation revealed that Sharp had been the mastermind behind the data leakage. He then claimed to be a whisleblower and gave the media false information about the internal investigation into the leakage he caused. As a result, the stock price plunged by approximately 20% and Ubiquity lost 4 billion USD in market value. In May 2023, Sharp was sentenced to six years in prison, received a three-year probation after his release and was ordered to pay more than 1.5 million USD in damages (Press Release, 2023; Belanger, 2023).

It is therefore clear that as digitalization progresses, new opportunities for both intentional and negligent incidents affecting end users are opening.

“Cybersecurity Ventures expects global cybercrime costs to grow by 15 percent per year over the next five years, reaching $10.5 trillion USD annually by 2025, up from $3 trillion USD in 2015. This represents the greatest transfer of economic wealth in history, risks the incentives for innovation and investment, is exponentially larger than the damage inflicted from natural disasters in a year, and will be more profitable than the global trade of all major illegal drugs combined.” (Morgan, 2021).

The costs of handling insider incidents keep increasing every year, not only in individual categories but overall as well. Figure 1 shows their trend in large organizations between 2018 and 2022.

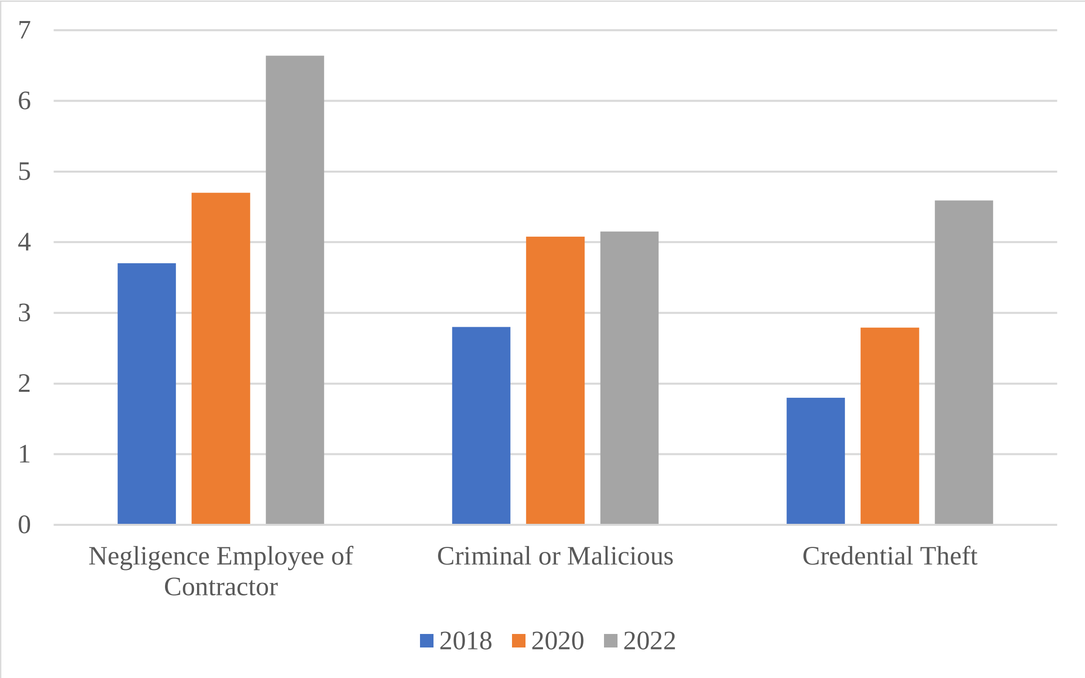


Figure 1. Impact of insider incidents in 2018 - 2022 in millions of USD Source: (Ponemon Institute, 2023)

1.1 Competences in investigating the impact of insider threats

The Cybersecurity and Infrastructure Security Agency (CISA) defines insider threat as the threat that an insider will use their authorized access, intentionally or unintentionally, to do harm to the department's mission, resources, personnel, facilities, information, equipment, networks, or systems. Insider threats manifest in various ways: violence, espionage, sabotage, theft, and cyber acts (CISA, 2023). An insider is an individual or supplier who has local knowledge of an organization and whose authorized access is knowingly or unknowingly abused to obtain and exfiltrate data of interest. It can also be an internal security team with an extended scope of operation and access to user activity auditing or a specialized firm providing insider threat detection and investigation services.

The investigation of insider threats is the domain of commercial organizations that need to protect their intellectual and industrial property, know-how, client information and any other information that is directly or indirectly used to obtain a competitive advantage and financial gain in a competitive environment.

In practice, we usually encounter three main groups of users, based on their motivation:

- Disgruntled/malicious user.
- Cereless/Negligent user.
- Infiltrator.

Employee investigations are primarily the responsibility of the organizations concerned, but in certain situations they are done by an external firm. The investigation team consists of security analysts who are responsible for securing, processing and interpreting digital footprints. The HR department is responsible for communicating with the employee, making sure that he or she is treated fairly and with dignity, coordinating with other teams, documenting the investigation, escalating and reporting to management. The HR department also performs the final evaluation of the investigation in compliance with the applicable in-house regulations, provides the company with recommendations and potentially initiates disciplinary proceedings. The legal department makes sure that the investigation is carried out in compliance with the laws and regulations applicable to the country and the economic sector, evaluates the findings, makes recommendations or directly informs the relevant organizational units of their duty to inform the public administration authorities about the incident, for example in the case of GDPR and incidents involving the loss of personal information, where it is required by law to immediately notify the Office for the Protection of Personal Data.

The employee's superiors are usually also involved in the investigation and provide the context to the identified user activities.

1.2 The aim of the article

The aim of this article is to highlight the growing danger of insider incidents and their significance for the secure running of company information systems in a global context. Insider threats have not received as much attention for various reasons, e.g. the protection of the privacy of own employees, their diversity and loyalty to the employer. In practice, however, it can lead to data leaks, complicated legal situations and, as a result, an overall loss of competitiveness of an organization. The article also discusses select ways of preventing insider incidents, as well as the legislation of the Czech Republic and the Slovak Republic.

2 Methodology

We used three main sources for the article. The first source includes the quantitative data reported for the years 2016-2022, for example in the study that is a result of Ponemon Institute's 2023 survey. We used the same methodology as that in the study (Ponemon Institute, 2023), where we divided the data collected from firms by three types of incidents:

- relating to negligence,
- relating to criminal insider,
- relating to user credential theft.

The second source includes the conclusions from a range of commercial and scientific security projects.

The third source includes professional literature and documents about information security in organizations.

3 Results

We divided the results into two relatively separate sections. The first section shows partial conclusions based on analysed data from 278 large organizations around the world, where we formulate some basic insider incident trends during the years 2016-2022. Large organizations are regionally represented as follows: North America 44%, Europe 27%, Asia 19%, Africa and Arabia 10%.

The second section contains suggestions for handling insider incidents in the Czech Republic and the Slovak Republic.

3.1 Insider incident situation and trends

The first outputs include elementary characteristics of the acquired data. We collected data from a total of 278 firms that experienced 6,803 insider incidents and lost a total of 15.4 million USD. The breakdown of insider incidents by type is provided in Table 1.

Table 1. Breakdown of analyzed insider incidents by type

Incidents Relating to	Percentage	Annual Costs in Mio USD
Negligence	56	6.6
Criminal Insider	26	4.1
Credential Theft	18	4.6
Sum	100	15.3

Source: (Ponemon Institute, 2023)

Figure 2 below shows the structure of the surveyed organizations by economic sector in which they perform their main activity.

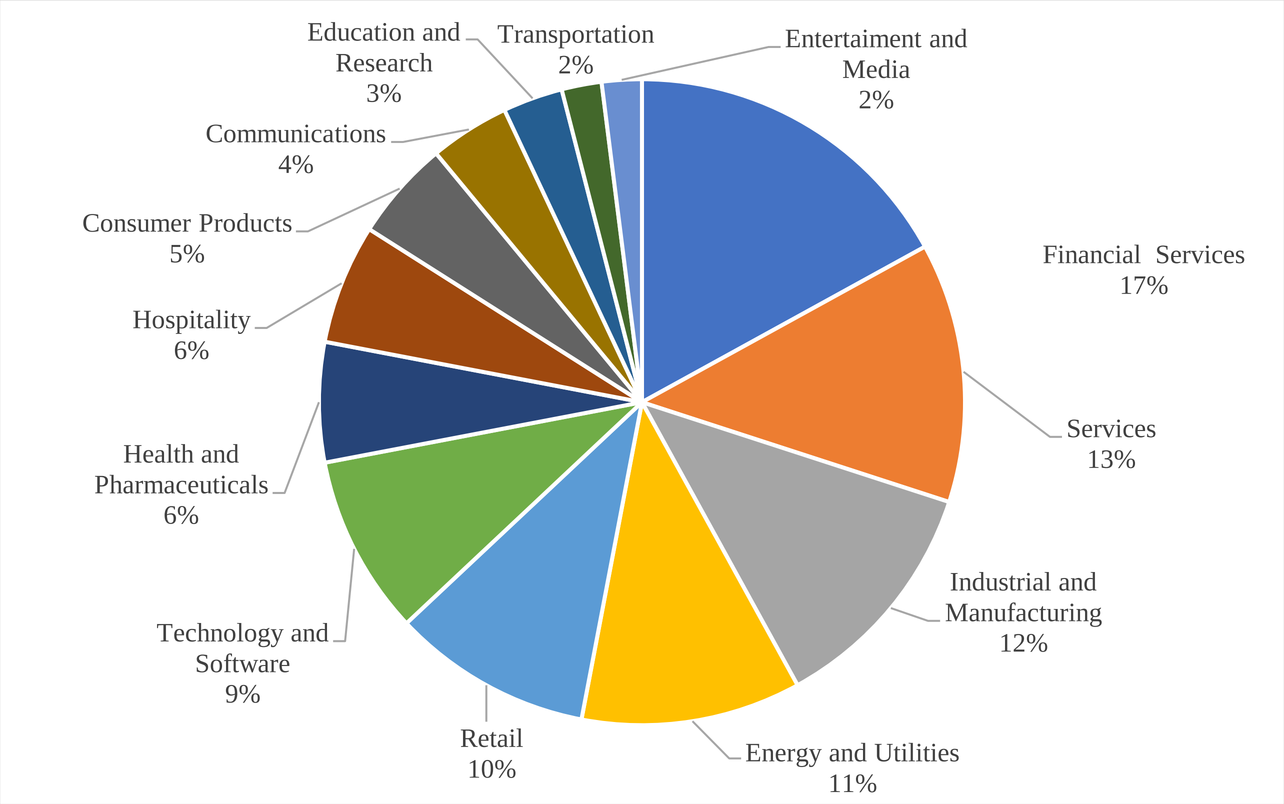


Figure 2. Representation of surveyed organizations by economic sector Source: (Ponemon Institute, 2023)

Figure 3 presents the size of the surveyed organizations, based on the number of their employees whose data were processed in the survey. The figure clearly shows that these organizations are mainly larger and large companies with over 500 employees.

Figure 3 clearly shows that these are primarily large organizations, based on the classification of the European Union. Therefore, our considerations and recommendations do not primarily apply to micro and small and medium-sized enterprises (SMEs). Our sample includes only 16% of SMEs.

Another very significant risk of insider incidents is how long it takes to restore an organization's system after an insider incident was detected. During this time, the organization is exposed to the threats activated by the incident. The average incident "lifetime", and thus the average remediation time, are shown in Figure 4.

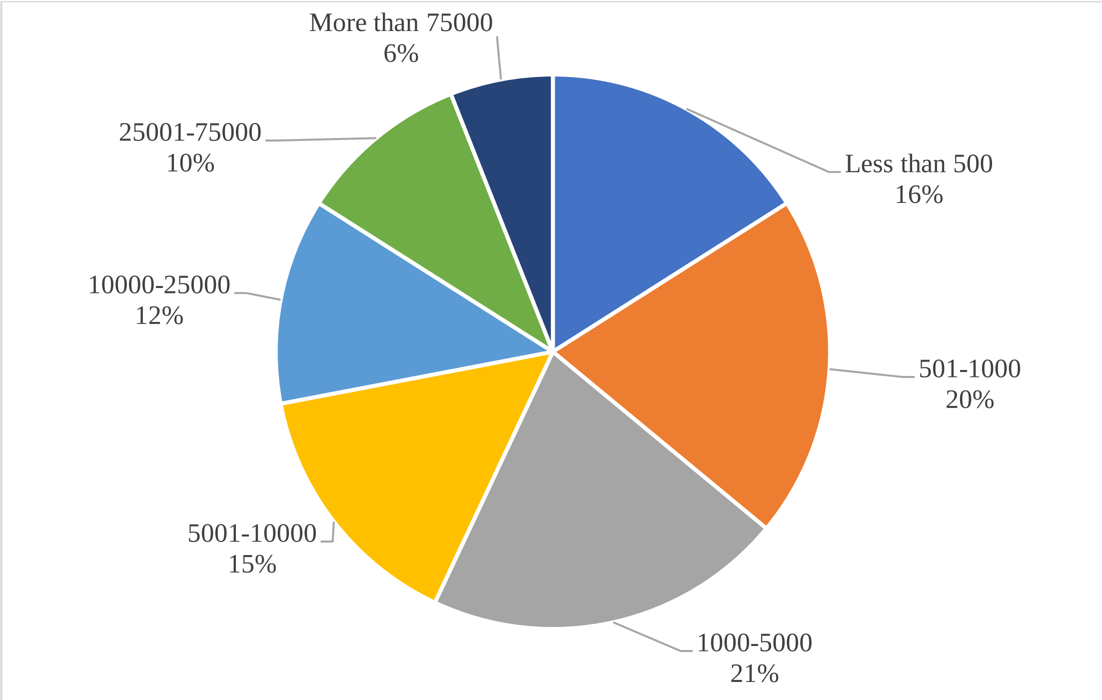


Figure 3. Representation of surveyed companies by number of employees Source: (Ponemon Institute, 2023)

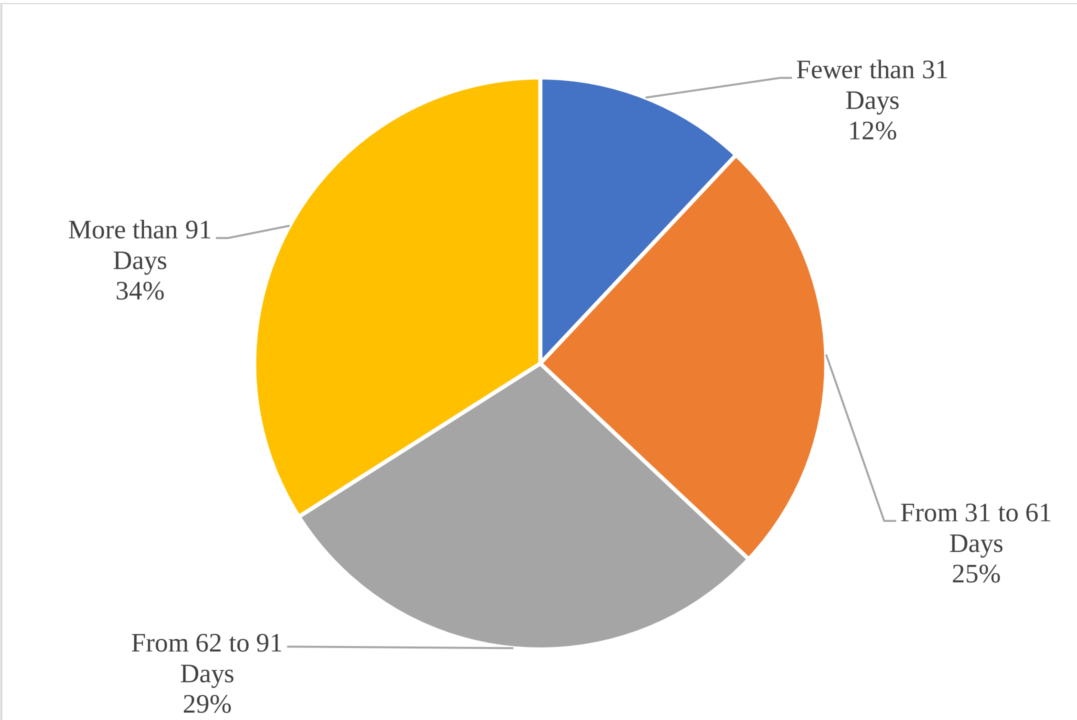


Figure 4. Average time of resolving insider incidents Source: (Ponemon Institute, 2023)

It is not easy to identify an insider incident because an internal user may use the organization's information system for his or her own personal benefit (in the past, this usually included cryptocurrency mining or retail activities for entities other than the employer).

3.2 Suggestions for protection against insider incidents

The cost of handling insider incidents is another key factor for organizations. Figure 5 shows the cost trend between 2016 and 2022.

The presented data show the average cost of a security insider incident as identified by seven security operation centres. The most expensive stages of the incident life cycle are: investigation, incident response, containment and remediation. Let's take a closer look at how to prevent insider incidents.

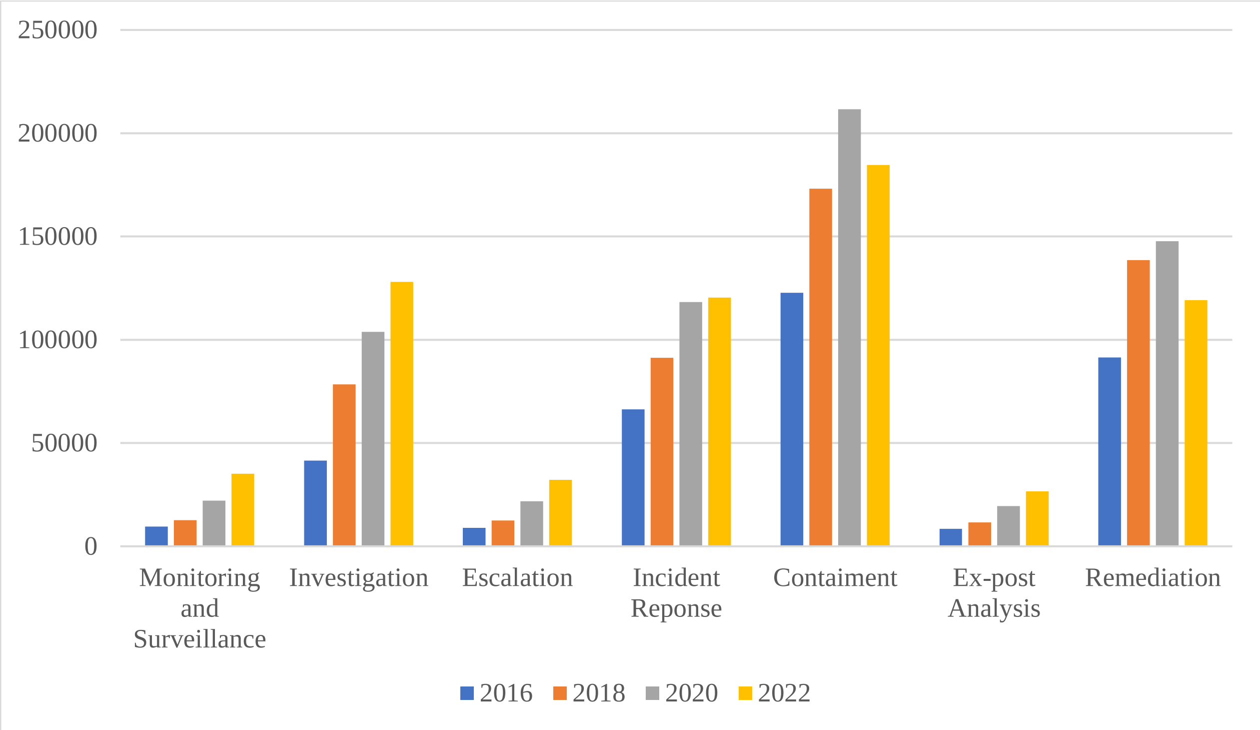


Figure 5. Cost of insider incident handling during its life cycle Source: (Ponemon Institute, 2023)

Organizations are often forced to address protection against external threats at the legislative level (Gelles, 2016), while protection against insider threats is still a taboo topic that is viewed as employee privacy violation. Confidentiality and security of sensitive information needs to be addressed not only at the technological level, but also through proper (informal) training in the secure use of company information and computer systems. Users' confirmation of accepting employee rights, obligations and responsibilities in relation to information technology is key to achieving an acceptable level of information security.

The basic administrative security tools against insider threats are as follows:

The Code of Conduct is a binding set of behavioural principles and social norms that an organization expects its employees to adhere to. Every organization has its own specific Code of Conduct, but it usually includes an affidavit of knowledge of and compliance with legislation applicable to the particular field of activity as well as employees' responsibilities and obligations to identify and report harmful and non-standard events. With regard to the protection of sensitive information, the Code of Conduct should include general rules for data handling, access and encryption and for insider incident reporting, with reference to the NIS Directive (directive on security of network and information systems) that provides detailed guidelines for the proper use of company information systems and protection against unauthorised access to data.

A Non-Disclosure Agreement (NDA) is a legally binding agreement between two or more parties that defines confidentiality in relation to disclosed information, including penalties for confidentiality violation. This is a way to protect sensitive information or data not protected by other laws, e.g. by copyright law. An NDA is a suitable addition to an employment contract for employees who come into contact with confidential company information and documents.

Classification of documents based on the level of document confidentiality. A given level of document confidentiality determines how users can work with certain documents, who they can share them with and where and how they can be stored. Documents marked with the correct level of confidentiality can be tracked using Data Loss Prevention tools (DLP), and unauthorized access to documents can be reported as incidents to the Data Security team. Security levels can be enhanced by proactive measures limiting the user's ability to exfiltrate data. These measures can limit access to personal e-mails and to document printing and block entries to external USBs and cloud storage. If it is not possible to use these measures across the board, they should be used at least with respect to users who gave or received notice.

User behavioural analysis (UBA) is based on monitoring user behaviour over a long period of time and creating patterns to detect any abnormal behaviour. Specifically, it can identify a user interested in documents from projects that he or she does not usually work with or identify any increased downloading of documents from central repositories to a workstation.

An effective solution to protection against data leakage consists of a combination of formal agreements (Veber, Nedomova and Doucek, 2016) with employees having access to confidential documents, good-quality training focused on secure data handling and technical means preventing and detecting unauthorized operations with confidential data.

4. CONCLUSION

Unauthorized use of non-public information for personal enrichment is a concept far predating information technology, but it is the development of personal computers and the Internet that has opened up new opportunities to gain a financial or competitive advantage.

Unauthorized use of non-public information in stock exchange operations and stock investments in general is the first type of information monetization regulated by law. In the United States, the first laws regulating stock exchange operations and trading in investment instruments were enacted in 1933. The Securities Act was enacted first, followed by the Securities Exchange Act, in order to achieve an environment where all participants would have an equal chance of making a profit. In 1991, Section 128, Abuse of Information in Trading, of the CR’s Penal Code no. 557/1991 of Coll. was added. Its explanatory memorandum covers "Insider Trading" (ÈSFR, 1991a). Currently, "Insider Trading" is covered in Section 255, Abuse of Information in Trading, of the CR’s Penal Code no. 40/2009 of Coll. and in Section 265, Abuse of Information in Trading, of the SR’s Act no. 300/2005 of Coll.

Corporate information and data that can be classified as artworks or scientific works are protected by Copyright Act no. 121/2000 of Coll. in the Czech Republic and Act no. 185/2015 of Coll. in Slovakia. An important aspect of these works is the requirement of creative activity. The Czech and Slovak legal systems also consider computer programs (Směrnice Evropského parlamentu a Rady

2009/24/ES ze dne 23. dubna 2009) or databases to be copyrighted works, even if they lack elements of originality (Slovensko, 2015, § 130-140). In the case of databases, the time and financial investment required to create the database (the sui generis right) is taken into account (Kohutová, 2012). Patented technical solutions or technical solutions registered in the Utility Models Register are protected by Act no. 221/2006 of Coll. that regulates the enforcement of industrial property rights and the protection of trade secrets.

A confidentiality agreement, which clearly states sanctions for any breach of the agreement, is considered to be a generally acceptable solution to protect corporate documents and intellectual property.

Incidents related to GDPR , telecommunications (Česko, 2000; Slovensko, 2021,§ 117), trade (Česko, 2012, § 504; ÈSFR, 1991, § 18-20) and banking secrets (ÈSFR, 1992, § 38 odst. 1; Slovensko, 2001, § 89-§ 93c) or cyber law (Česko, 2014.; Slovensko, 2018), which are specific to particular types of protected information, represent a specific type of investigation.

In general terms, the damage and abuse of computer systems and information stored in these systems are covered in Section 230 of the Penal Code (Czech Republic), and Section 247 of the Penal Code (Slovakia) that focus on unauthorised access to computer systems and information media (Česko, 2009, § 230; Slovensko, 2005, § 247).

The aforesaid can be applied to users' unauthorised access to information systems in order to obtain information about customers or their services.

The fundamental difference between external threats and insider threats is the knowledge of the value of individual information sources and specific documents. In the case of Tele Denmark Communications (TDC), the key factor was access to detailed information in Ericsson's non-public bid for the construction of the mobile 5G network; in other cases, it may be customer identification and location data or records from public administration systems.

It is the responsibility of organizations to assess the risk of different types of information and to propose specific countermeasures and processes for monitoring and investigating. The Czech legal system covers damages caused by insider threats rather sufficiently, although it would be good to consolidate the aforesaid laws into a single law, especially in terms of gap analyses and security audits.

ACKNOWLEDGMENTS

Paper was processed with support from institutional-support fund for long-term conceptual development of science and research at the Faculty of Informatics and Statistics of the Prague University of Economics and Business (IP400040) and internal PUEB project IG409033.

REFERENCES

Abrams, M. and Weiss J., 2008. Malicious Control System Cyber Security Attack Case Study–Maroochy Water Services, Australia, In. Annual Computer Security Applications Conference. Available at: https://www.mitre.org/sites/default/files/pdf/08_1145.pdf

Belanger, A., 2023. Ex-Ubiquiti engineer behind "breathtaking" data theft gets 6-year prison term, Available at: https://arstechnica.com/tech-policy/2023/05/ex-ubiquiti-engineer-behind-breathtaking-data-theft-gets-6-year-prison-term/

Bing, C., 2018. Stolen U.S. drone documents found for sale on dark web, Available at: https://cyberscoop.com/us-drone-dark-web-us-air-force-recorded-future/

Bunn, M. and Scott D. S., 2017. Insider Threats. Series Cornell Studies in Security Affairs. Ithaca, New York: Cornell University Press.

Catrantzos, N., 2022. Managing the Insider Threat: No Dark Corners and the Rising Tide Menace. 2nd ed. Florida: CRC Press, Taylor & Francis Group.

ČSFR, 1991. Zákon č. 513 ze dne 5. listopadu 1991 obchodní zákonník. In: Sbírka zákonů České a Slovenské federativní republiky. 1991, částka 98, pp. 2874-2565. Available at: https://www.psp.cz/sqw/sbirka.sqw?cz=513&r=1991

ČSFR, 1991a. Zákon č. 557 ze dne 11. prosince 1991 kterým se mění a doplňuje trestní zákon. In: Sbírka zákonů České a Slovenské federativní republiky. 1991, částka 105, pp. 2475-2751. Available at: https://www.psp.cz/sqw/sbirka.sqw?cz=557&r=1991

ČSFR, 1992. Zákon č. 21 ze dne 20. prosince 1991 o bankách. In: Sbírka zákonů České a Slovenské federativní republiky. 1992, částka 5, pp. 98-105. Available at: https://www.psp.cz/sqw/sbirka.sqw?cz=21&r=1992

Česko, 2000. Zákon č. 151 ze dne16. května 2000 o telekomunikacích a o změně dalších zákonů. In: Sbírka zákonů České republiky. 2000, částka 47, pp. 2183-2219. Available at: https://www.psp.cz/sqw/sbirka.sqw?cz=151&r=2000

Česko, 2009. Zákon ze dne 8. ledna 2009 trestní zákoník. In: Sbírka zákonů České republiky. 2014, částka 11, pp. 354-461. Available at: https://www.psp.cz/sqw/sbirka.sqw?cz=40&r=2009

Česko, 2012. Zákon č. 89 ze dne 3. února 2012 občanský zákoník. In: Sbírka zákonů České republiky. 2012, částka 33, pp. 1026-1365. Available at: https://www.psp.cz/sqw/sbirka.sqw?cz=89&r=2012

Česko, 2014. Zákon č. 181 ze dne 23. července 2014 o kybernetické bezpečnosti a o změně souvisejících zákonů (zákon o kybernetické bezpečnosti). In: Sbírka zákonů České republiky. 2014, částka 75, pp. 1926-1936. Available at: https://www.psp.cz/sqw/sbirka.sqw?cz=181&r=2014

CISA, 2023. Insider Threat Mitigation: Defining Insider Threats. Cybersecurity & Infrastructure Security Agency. Available at: https://www.cisa.gov/topics/physical-security/insider-threat-mitigation/defining-insider-threats

Gady, F. S., 2015. New Snowden Documents Reveal Chinese Behind F-35 Hack: Experts have long argued that China has copied the F-35 design for its own fighter jets. Is this the proof? Available at: https://thediplomat.com/2015/01/new-snowden-documents-reveal-chinese-behind-f-35-hack/

Gelles, M. G., 2016. Insider Threat: Prevention, Detection, Mitigation, and Deterrence. UK: Publisher Butterworth-Heinemann.

Hellberg, M., 2013. Spionen Stig Bergling plågas av dödstankar. Expressen. Available at: https://www.expressen.se/nyheter/spionen-stig-bergling-plagas-av-dodstankar/

ISO/IEC 27005:2022 Information security, cybersecurity and privacy protection — Guidance on managing information security risks. International Organization for Standardization Switzerland.

Kohutová, R., 2012. Právní ochrana databází. Praha, 2012, 224 s. Rigorózní práce. Univerzita Karlova, Právnická fakulta. Vedoucí práce Irena Holcová. Available at: https://dspace.cuni.cz/bitstream/handle/20.500.11956/42331/150012720.pdf?sequence=1&isAllowed=y

Morgan, S., 2021. Cybercrime To Cost The World $10.5 Trillion Annually By 2025. Cybercrime Magazine. Special Report: Cyberwarfare In The C-Suite. Available at: https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/

Payne, J., Hanson J. and Wojtasiak M., 2022. Inside Jobs: Why Insider Risk Is the Biggest Cyber Threat You Can't Ignore. New York: Skyhorse.

Ponemon Institute, 2023. 2022 Ponemon Cost of Insider Threats Global Report, Available at: https://protectera.com.au/wp-content/uploads/2022/03/The-Cost-of-Insider-Threats-2022-Global-Report.pdf

Press Release, 2022. PH.D. Chemist Sentenced To 168 Months For Conspiracy To Steal Traded Secrets, Economic Espionage, Theft Of Trade Secrets, And Wire Fraud. U.S. Attorney's Office, Eastern District of Tennessee. Available at: https://www.justice.gov/usao-edtn/pr/phd-chemist-sentenced-168-months-conspiracy-steal-traded-secrets-economic-espionage

Press Release, 2023. Former Employee Of Technology Company Sentenced To Six Years In Prison For Stealing Confidential Data And Extorting Company For Ransom. U.S. Attorney's Office, Southern District of New York. Available at: https://www.justice.gov/usao-sdny/pr/former-employee-technology-company-sentenced-six-years-prison-stealing-confidential

Směrnice Evropského parlamentu a Rady 2009/24/ES ze dne 23. dubna 2009 o právní ochraně počítačových programů. In: EUR-Lex [právní informační systém]. Úřad pro publikace Evropské unie. Available at: http://eur-lex.europa.eu/legal-content/CS/TXT/PDF/?uri=CELEX:32009L0024&from=CS

Veber, J., Nedomová, L. and Doucek, P. 2016. Corporate Digital Incident Investigation. Quality Innovation Prosperity, 20(1), 57–70. https://doi.org/10.12776/qip.v20i1.656.

Slovensko, 2001. Zákon č. 483 z 5. októbra 2001 o bankách a o zmene a doplnení niektorých zákonov. Online. In: Zakony pre ľudí. Available at: https://www.zakonypreludi.sk/zz/2001-483#cl1

Slovensko, 2005. Zákon č. 300 z 20. mája 2005 trestný zákon. Online. In: Zakony pre ľudí. Available at: https://www.zakonypreludi.sk/zz/2005-300#cast1

Slovensko, 2015. Zákon č. 185 z 1. júla 2015 autorský zákon. Online. In: Zakony pre ľudí. Available at: https://www.zakonypreludi.sk/zz/2015-185

Slovensko, 2018. Zákon č. 69 z 30. januára 2018 o kybernetickej bezpečnosti a o zmene a doplnení niektorých zákonov. Online. In: Zakony pre ľudí. Available at: https://www.zakonypreludi.sk/zz/2018-69

Slovensko, 2021. Zákon č. 452 z 24. novembra 2021 o elektronických komunikáciách. Online. In: Zakony pre ľudí. Available at: https://www.zakonypreludi.sk/zz/2021-452

ANTIVIRUS SOFTWARE STATUS QUO BIAS

Tomáš Sigmund

Prague University of Economics and Business
Faculty of Informatics and Statistics
sigmund@vse.cz

DOI: 10.35011/IDIMT-2024-85

Keywords

Status Quo Bias; Antivirus Software Adoption; Cognitive Misperceptions; Cybersecurity Behavior Change

Abstract

This study explores the phenomenon of status quo bias in the context of antivirus software usage, examining the psychological, cognitive, and rational factors that discourage users from switching to newer technologies despite the availability of potentially superior options. Originating from foundational theories by Samuelson and Zeckhauser (1988), this research identifies key elements such as psychological commitment, cognitive misperceptions, and rational decision-making that contribute to this bias. Through a comprehensive survey involving 67 participants, the study evaluates how sunk costs, resistance to change due to uncertainty, and the influence of social norms affect user behavior in the cybersecurity domain. The findings reveal that while some users are influenced by the duration of usage due to sunk cost effects, cognitive biases predominantly cause users to overestimate the effectiveness of their current antivirus solutions, particularly in the absence of negative security incidents. Additionally, the study finds that neither perceived transition costs nor social influences significantly deter users from adopting new software, contrary to initial hypotheses. This research highlights the need for targeted educational and support strategies to address and mitigate the barriers to adopting new cybersecurity technologies.

1. Introduction

Antivirus software plays a crucial role in safeguarding personal and organizational information systems from malicious cyber threats. However, despite the availability of various cybersecurity solutions, many users continue to adhere to default or previously effective solutions (Kannelønning & Katsikas, 2023), including persisting with their current antivirus software. This adherence to the status quo, even in the face of potentially superior alternatives, is a phenomenon well documented in the literature on decision-making biases. Originating from seminal work by Samuelson and Zeckhauser (1988), status quo bias explains this inertia as a preference for the current state, influencing users to stick with familiar antivirus software despite the availability of upgrades or more effective technologies.

This reluctance to switch is underpinned by factors identified in status quo bias theory: psychological commitment, cognitive misperception, and rational decision-making. These factors contribute to a resistance to change, which can hinder the adoption of more robust cybersecurity measures (Balakrishnan et al., 2021). In the context of antivirus software, psychological commitment can manifest as a sunk cost effect, where users are reluctant to abandon the time or resources they have already invested in their existing software solutions (Godefroid et al., 2022). Cognitive misperceptions, influenced by a user's existing beliefs or lack of negative security incidents, may skew the perceived effectiveness of current antivirus tools, thereby discouraging exploration of alternative options (Li et al., 2016). Moreover, rational decision-making processes may deter users from switching software due to perceived transition costs and the uncertainty of outcomes associated with new antivirus solutions which can include both a different types of antivirus software and a new version of the current software (Samuelson & Zeckhauser, 1988). Continuing from the understanding of status quo bias within the domain of antivirus software adoption, it becomes crucial to delve deeper into the aspects that perpetuate this bias and its implications for cybersecurity.

Psychological Commitment and the Sunk Cost Effect: Psychological commitment to existing antivirus solutions can significantly impact user decisions. According to Godefroid et al. (2022), this commitment can be characterized as the sunk cost effect, where users continue using an antivirus software because they have already invested time, effort, or financial resources into it. This effect can lead users to irrational decision-making, whereby they continue to support a potentially less effective solution simply because they wish to avoid the perceived waste of their previous investments.

Cognitive Misperception: The misperception of the existing system's benefits and costs plays a critical role in the status quo bias. J. Li et al. (2016) discuss how cognitive misperception, fuelled by inertia, causes users to resist adopting new technologies even when compelling evidence suggests better alternatives. In the case of antivirus software, users may overestimate the effectiveness of their current solution due to a lack of adverse security events, thus becoming complacent. This complacency is dangerous as it may leave users exposed to evolving cyber threats that their outdated systems are ill-equipped to handle.

Rational Decision-Making: The rational decision-making process in the context of status quo bias involves evaluating the benefits and risks associated with switching to a new antivirus software. Samuelson and Zeckhauser (1988) highlight that this process often leans towards maintaining the status quo due to the perceived risks and uncertainties associated with new options. Users may fear that new software could be incompatible with their systems, or they may be concerned about the time and effort required to learn and implement a new system. These perceived transition costs and uncertainties can significantly deter users from pursuing potentially more effective antivirus protections.

There are some concepts similar or related to status quo bias, such as default bias, organizational inertia, loss aversion, sunk cost effect, and innovation resistance.

Default Bias: Explored through studies like those by Krieger and Felder(2013), Geng (2016), and Suri et al. (2013), this bias refers to individuals' tendencies to adhere to a pre-selected option when presented with multiple choices. Samuelson and Zeckhauser (1988) first referenced this effect, highlighting how both prior choices and the framing of options influence decision-making. This bias is significant in technological settings, such as maintaining default settings in technology installations that might pose security risks.

Organizational Inertia: Described by Tushman and O'Reilly (1996), organizational inertia explains how entrenched structures within an organization impede new system adoption. This resistance is due to the organization's complexity and established routines, which are further examined in contexts like mobile payment systems adoption (Zhang & El-Gohary, 2016) and healthcare technologies (Hsieh et al., 2014).

Loss Aversion: This principle, established by Kahneman and Tversky (1979) and further discussed by Samuelson and Zeckhauser (1988), posits that the fear of potential losses in the context of change is often perceived as more significant than possible gains. This is a critical aspect of SQB where decision-makers forego new opportunities due to exaggerated fears of loss.

Sunk Cost Effect: This effect, which drives individuals to continue actions in which they have invested resources, is explored by Arkes and Blumer (1985). The reluctance to abandon these investments, even when they are no longer beneficial, is a pivotal factor reinforcing the status quo. In the case of ongoing subscriptions, the sunk costs concern the resources used for introducing and managing the software, as well as the skills and habits related to its use. In the case of a one-time payment for the software, any unused portion of the money would also be lost.

Innovation Resistance: Explored by Ram (1987) and Heidenreich and Handrich (2015), this concept reflects a normal consumer response aimed at maintaining the status quo and current beliefs when faced with new innovations. This resistance is particularly pronounced among consumers satisfied with current products, illustrating the link between satisfaction and resistance to change.

The reluctance to transition to newer antivirus technologies can have serious implications for cybersecurity. As cyber threats evolve rapidly, clinging to outdated software due to status quo bias can leave systems vulnerable to new types of attacks that older software may not effectively detect or mitigate. This scenario underscores the need for strategies that not only highlight the benefits of newer antivirus solutions but also address the psychological and practical barriers that inhibit change.

1.2 Strategies to Overcome Status Quo Bias

To combat the status quo bias in antivirus software adoption, several strategies can be employed. **Education and Awareness**: Raising awareness about the limitations of current antivirus solutions and the benefits of newer technologies can help alter user perceptions and reduce cognitive misperceptions (Arachchilage & Love, 2014). **Ease of Transition**: Reducing the perceived costs and difficulties associated with switching antivirus software can encourage users to make the change. This could involve providing free trials, straightforward installation processes, and robust customer support (Venkatesh et al., 2003). **Highlighting Risk**: Demonstrating the potential risks and consequences of not switching to more effective antivirus software can leverage the natural aversion to loss that many users feel, countering the sunk cost effect (Keller & Lehmann, 2008). In the context of studying status quo bias in the adoption of antivirus software, designing research hypotheses and formulating specific questions is crucial for examining the underlying effects of this bias.

2. Hypothesis and questionnaire

Hypothesis 1 (H1): Psychological commitment to existing antivirus software is positively correlated with the intention to continue using current antivirus software.

Hypothesis 2 (H2): Cognitive misperception regarding the efficacy of current antivirus software positively influences users' intention to continue using it.

Hypothesis 3 (H3): Higher levels of awareness about the limitations and risks of current antivirus software are associated with a greater willingness to adopt new technologies and are negatively correlated with the intention to continue using the current antivirus software.

Hypothesis 4 (H4): Perceived transition costs and uncertainties about new antivirus software are significant barriers to changing existing antivirus solutions and are positively correlated with the intention to continue using the current antivirus software.

Hypothesis 5 (H5): Social influence plays a significant role in the adoption of new antivirus software, where users who perceive a social norm of upgrading are more likely to change their existing software and those who don't perceive it are more likely to continue using their current antivirus software.

Based on the discussion of status quo bias and its impact on the adoption of antivirus software, a questionnaire was developed to assess the psychological, perceptual, and practical factors that influence users' decisions to stick with their current antivirus solutions or switch to newer ones. We asked about the private use of software. We used an online questionnaire and shared its link on various Czech social networks to gather respondents, who were selected randomly based on their visit to the social network. Here is a proposed questionnaire with questions designed to evaluate the hypotheses and research questions previously outlined. The respondents were asked in the Czech language and that is why we aimed only at Czech respondents.

Part 1: Demographic Information

1. Age
2. Gender
3. Education Level
4. Are you tech-savvy? (Likert scale – 5 items: definitely yes – definitely no)

Part 2: Resistance to change antivirus software

5. I will continue to use my current antivirus software. (Likert scale – 5 items: definitely yes – definitely no)

Part 3: Psychological Commitment and Sunk Cost

6. How long have you been using your current antivirus software? (less than one year, one year-two years, more than two years)
7. Have you spent a lot of money on your current antivirus software (including any subscriptions)? (Likert scale – 5 items: definitely yes – definitely no)

Part 4: Cognitive Misperception

8. How effective do you believe your current antivirus software is in protecting against new cyber threats? (Likert scale – 5 items: Very effective – Not effective at all)
9. Have you experienced any security breaches while using your current antivirus software? (Likert scale – 5 items: Yes, serious problems – No, no problems at all)

Part 5: Rational Decision-Making (awareness and transition costs)

10. How informed do you feel about the latest developments in antivirus technology? (Likert scale – 5 items: Not at all – Very well)
11. Do any concerns about switching to a new antivirus software bother you? (E.g. Compatibility issues, Cost of new software, Learning new software, Time to install/set up, Other) (Likert scale – 5 items: They bother me a lot – they don't bother me at all)

Part 6: Social Influence and Awareness

12. How often do your peers discuss or recommend updating or changing antivirus software? (Likert scale – 5 items: Very often – rarely)
13. Do social media influence your decisions to use current antivirus software? (Likert scale – 5 items: definitely yes – definitely no)

Part 7: Intervention Strategies

14. Would you be more likely to switch antivirus software if offered a free trial of the new software? (Likert scale – 5 items: definitely yes – definitely no)
15. How effective do you think educational campaigns about the risks of outdated antivirus software would be in encouraging you to switch? (Likert scale – 5 items: Very effective – Not effective at all)
16. What type of incentive would most likely persuade you to switch to a new antivirus software? (Discounts, Improved features, Better customer support, Expert recommendations, None)

This questionnaire aims to gather data on the factors that influence the status quo bias in antivirus software usage, providing insights into how different interventions might encourage users to adopt newer, more effective solutions. Outdated antivirus software refers to an antivirus program that has not been updated with the latest virus definitions, patches, and security improvements because of e.g. expired subscription, occurrence of new threats or low software performance of older versions. It was not our concern to determine if the software used by the respondents was up-to-date with the latest technology, and this information does not affect our results as we were investigating the status quo bias.

3. Results

We collected 67 valid answers, 2 respondents were excluded due to invalid responses. The number of respondents is not very high, which can affect the results. Nonetheless because of their random selection we hope the results are to some extent indicative. The respondents ranged in their age from 20 to 64 years. 40% were females, 50% males.

Table 1. Education of respondents

Education	Frequency	Percent
High School	23	34,3
Bachelor	19	28,4
Master	19	28,4
PhD	6	9,0
Total	67	100,0

Source: (author)

Table 2. Descriptive Statistics of questions, Likert scales values are explained in the questions

Question number	Question	Mean	Std. Deviation
1	Age	37,30	11,93
4	Tech Savviness	3,42	1,05
5	Intention to Continue	2,27	1,27
6	Usage Duration	2,33	0,59
7	Spending on Antivirus	2,66	1,20

Question number	Question	Mean	Std. Deviation
8	Perceived Efficacy	1,40	0,85
9	Security Breaches	2,82	1,52
10	Awareness Developments	4,22	1,22
11	Transition Concerns	1,58	1,03
12	Peer Discussions	4,28	1,30
13	Social Media Influence	4,16	1,23
14	Free Trial Willingness	3,44	1,70
15	Effectiveness Educational Campaigns	3,28	1,58

Source: (author)

Table 3. Status quo bias detection: Correlation of Q5 with other questions

***…significant at the 5% s.l., ** significant at the 1% s.l.**

Spearman correlation between Q5 and	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13
	0,30*	0,1	0,44**	-0,34**	-0,45**	0,09	-0,17	0,11
	H1	H1	H2	H2	H3	H4	H5	H5

Source: (author)

Table 4. Question 16: Incentive Preference

	Frequency	Percent
Discounts	5	7,5
Features	8	11,9
None	22	32,8
Expert recommendations	17	25,4
Support	15	22,4
Total	67	100,0

Source: (author)

4. Discussion

The findings from this study offer insightful observations into the status quo bias regarding antivirus software usage. These insights help delineate various psychological and practical factors influencing users' decisions to adhere to their existing antivirus solutions or to opt for newer technologies.

Psychological Commitment and Sunk Cost Effect: The analysis confirmed Hypothesis 1 (H1), but only in relation to usage duration. A moderate positive correlation between the intention to continue using current antivirus software and the duration of usage was observed, suggesting that the longer a user has been with a software, the more likely they are to continue using it. This can be attributed to the sunk cost effect, where users are reluctant to change due to the time and effort already invested in their current solution. However, financial investment in the software did not show a significant correlation, indicating that monetary sunk costs might not be as influential in this context as previously thought.

Cognitive Misperception: Hypothesis 2 (H2) was confirmed for both perceived efficacy and security breaches, indicating a comprehensive impact of cognitive misperception on user behavior. Users tend to overestimate the effectiveness of their current antivirus solutions, particularly when they have not experienced significant security issues, leading to complacency. This complacency poses a risk as it may leave users vulnerable to new threats that their outdated systems are ill-equipped to handle.

Rational Decision-Making: Hypothesis 3 (H3) regarding higher levels of awareness leading to a willingness to adopt new technologies was substantiated. This suggests that when users are more informed about the limitations and risks of their current antivirus software and the benefits of newer technologies, they are more likely to consider switching. On the other hand, unaware users tend to continue using their current antivirus software. This finding underscores the importance of awareness and informed decision-making in overcoming status quo bias. Contrary to expectations, Hypotheses 4 (H4) was not supported by our data. Concerns about transition costs and uncertainties related to new antivirus software (H4) did not significantly correlate with the intention to continue using current software.

Social Influence and Awareness: Contrary to Hypothesis 5 (H5), we found that social influence, measured through peer discussions and social media impacts, did not significantly correlate with users' decisions to stick with or switch their antivirus software. This suggests that decisions regarding antivirus software are less influenced by social norms or peer pressure than expected.

The responses to Question 16 on incentive preference highlight a significant adherence to status quo among users, with 32.8% indicating that no incentive would likely persuade them to switch antivirus software, underscoring the challenge of overcoming inertia. Expert recommendations and robust support were seen as the most compelling incentives, favoured by 25.4% and 22.4% of respondents respectively, suggesting that social influence and enhanced customer support are key motivators. In contrast, discounts and enhanced features were less influential, indicating that economic factors and additional functionalities are not primary drivers for changing antivirus solutions. This suggests a strategy focusing on social proof and reliable support might be more effective in encouraging users to adopt new technologies. The persistence of status quo bias in antivirus software usage presents substantial challenges in cybersecurity. It highlights the necessity for more effective communication strategies that not only raise awareness about the benefits of new technologies but also address the specific biases and barriers that deter users from adopting them. Cybersecurity professionals and software providers should focus on strategies that reduce psychological resistance, correct cognitive misperceptions, and provide clearer information about the advantages of switching to more effective solutions. As a limitation of our study, we may mention that socioeconomic status may affect the results to some extent, but hopefully not significantly, as antivirus software is not very expensive.

5. Conclusion

This study has successfully illuminated the complex interplay of psychological, cognitive, and rational factors that contribute to the status quo bias in antivirus software usage. The results reveal a notable resistance among users to switch from their current antivirus solutions, driven primarily by sunk cost effects and cognitive misperceptions. While the sunk cost effect is more profoundly influenced by the duration of usage rather than financial investment, cognitive misperceptions are significantly shaped by users' lack of negative security experiences, leading to an overestimation of their current software's efficacy. These insights are crucial for understanding why users remain attached to their existing antivirus solutions despite the availability of potentially superior alternatives. To combat these entrenched behaviors, cybersecurity professionals need to develop targeted strategies that address these specific biases. Educational initiatives that enhance awareness of the risks associated with outdated software and the benefits of newer technologies could be particularly effective. Additionally, simplifying the transition process and providing robust support can help alleviate the practical barriers associated with switching software.

Acknowledgement

Paper was processed with support from institutional-support fund for long-term conceptual development of science and research at the Faculty of Informatics and Statistics of the Prague University of Economics and Business (IP400040) and the project IG 409033.

References

Arachchilage, N. A. G., & Love, S. (2014). Security awareness of computer users: A phishing threat avoidance perspective. Computers in Human Behavior, 38, 304–312. https://doi.org/10.1016/j.chb.2014.05.046

Arkes, H., & Blumer, C. (1985). The psychology of sunk cost. Organizational Behavior and Human Decision Processes, 35, 124–140. https://doi.org/10.1016/0749-5978(85)90049-4

Balakrishnan, J., Dwivedi, Y. K., Hughes, L., & Boy, F. (2021). Enablers and Inhibitors of AI-Powered Voice Assistants: A Dual-Factor Approach by Integrating the Status Quo Bias and Technology Acceptance Model. Information Systems Frontiers. https://doi.org/10.1007/s10796-021-10203-y

Geng, S. (2016). Decision Time, Consideration Time, and Status Quo Bias. Economic Inquiry, 54(1), 433–449. https://doi.org/10.1111/ecin.12239

Godefroid, M.-E., Plattfaut, R., & Niehaves, B. (2022). How to measure the status quo bias? A review of current literature. Management Review Quarterly. https://doi.org/10.1007/s11301-022-00283-8

Heidenreich, S., & Handrich, M. (2015). What about Passive Innovation Resistance? Investigating Adoption-Related Behavior from a Resistance Perspective. Journal of Product Innovation Management, 32(6), 878–903. https://doi.org/10.1111/jpim.12161

Hsieh, P.-J., Lai, H.-M., & Ye, Y.-S. (2014). Patients’ acceptance and resistance toward the health cloud: An integration of technology acceptance and status quo bias perspectives. PACIS 2014 Proceedings. https://aisel.aisnet.org/pacis2014/230

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. Econometrica, 47(2), 263–291. https://doi.org/10.2307/1914185

Kannelønning, K., & Katsikas, S. K. (2023). A systematic literature review of how cybersecurity-related behavior has been assessed. Information & Computer Security, 31(4), 463–477. https://doi.org/10.1108/ICS-08-2022-0139

Keller, P. A., & Lehmann, D. R. (2008). Designing Effective Health Communications: A Meta-Analysis. Journal of Public Policy & Marketing, 27(2), 117–130. https://doi.org/10.1509/jppm.27.2.117

Krieger, M., & Felder, S. (2013). Can Decision Biases Improve Insurance Outcomes? An Experiment on Status Quo Bias in Health Insurance Choice. International Journal of Environmental Research and Public Health, 10(6), 2560–2577. https://doi.org/10.3390/ijerph10062560

Li, J., Liu, M., & Liu, X. (2016). Why do employees resist knowledge management systems? An empirical study from the status quo bias and inertia perspectives. Computers in Human Behavior, 65, 189–200. https://doi.org/10.1016/j.chb.2016.08.028

Ram, S. (1987). A Model of Innovation Resistance. ACR North American Advances, NA-14. https://www.tcrwebsite.org/volumes/6688/volumes/v14/NA-14/full

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. Journal of Risk and Uncertainty, 1(1), 7–59. https://doi.org/10.1007/BF00055564

Suri, G., Sheppes, G., Schwartz, C., & Gross, J. J. (2013). Patient inertia and the status quo bias: When an inferior option is preferred. Psychological Science, 24(9), 1763–1769. https://doi.org/10.1177/0956797613479976

Tushman, M. L., & O'Reilly, C. A. (1996). Ambidextrous Organizations: Managing Evolutionary and Revolutionary Change. California Management Review, 38(4), 8–29. https://doi.org/10.2307/41165852

Venkatesh, Morris, Davis, & Davis. (2003). User Acceptance of Information Technology: Toward a Unified View. MIS Quarterly, 27(3), 425. https://doi.org/10.2307/30036540

Zhang, L., & El-Gohary, N. M. (2016). Discovering Stakeholder Values for Axiology-Based Value Analysis of Building Projects. Journal of Construction Engineering and Management, 142(4), 04015095. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001004

LARGE LANGUAGE MODELS SAFETY COMPROMISE BY USING NON-MAINSTREAM LANGUAGE

Marie Kovářová, Milos Maryska

Affiliation University of Prague University of Economics and Business
Faculty of Informatics and Statistics
marie.kovarova@vse.cz, milos.maryska@vse.cz

DOI: 10.35011/IDIMT-2024-95

Keywords

LLM; security; language

Abstract

This paper investigates the security challenges posed by Large Language Models (LLMs) like Chat GPT and Gemini, focusing on the potential misuse of these models in cyber-attacks. The study aims to analyse the effectiveness of safety measures in these models, particularly when non-mainstream languages, such as Czech, are used. The methodology involves a literature review on LLMs security and cyber-attacks, and testing the safety measures of ChatGPT and Gemini in English and Czech. The results highlight the increasing global cyber threat, with AI chatbots like ChatGPT potentially empowering hackers by generating malicious content when non-mainstream language is used. The paper underscores the risks associated with harmful content generation, including cyber security issues, and the misuse of LLMs in initiating cyber-attacks, spreading misinformation, and extracting sensitive information. The study calls for comprehensive safety measures that cover all languages to mitigate these risks.

1. Introduction

Recent advances in machine learning (ML) and deep learning (DL) have revolutionized many areas, from natural language processing through image data analysis to autonomous driving. However, these advances have also brought with them new security challenges. Especially Large Language Models (LLM) are the most popular now (Catlin, 2024). The best examples are Chat GPT (Open AI.,2024) and Gemini (Google.,2024).

The security challenges in LLMs field can be divided into two areas. One is preventing the models from giving harmful information (even unintentionally). The second is to prevent the intended usage of models for generating harmful content from malicious content to functional malware or virus source code. As this can be expected to be a problem, companies such as OpenAI created safety measures to prevent it (Open AI.,2023). But as the mainstream language is English also the safety measures are mainly handled in English. To cover all languages can be very challenging.

The aim of this paper is to analyse the possible usage of a non-mainstream language as the Czech language for compromising safety measures in LLMs models in comparison to the English language.

This paper is not aiming at security of models itself (e.g. poisoning learning data). The paper focuses on how effectively models prevent giving harmful information which was asked via prompt intentionally or even unintentionally. And whether they will provide harmful or malicious content (code, functional malware, etc.). All of this is based on language used in prompts. In general, models should be able to resist such attempts, so they will not be misused for illegal activities. This is a more ethical problem as the usual attackers (hackers) have knowledge or other sources from which they can extract such knowledge. But LLM models should not be a new source of such knowledge, especially when it can be easily accessed. Security measures should prevent curious users (especially children) or novice hackers from getting instructions or functional code usable for illegal activities.

2. Methodology

To complete the aim of our paper, we set the following objectives: 1) Conduct a literature review on LLMs security related to used language and on known applications of cyber-attacks misusing LLMs for preparation. 2) Test the safety measures of the most commonly used LLM based chats: ChatGPT and Gemini. They are tested for differences between results from prompts in the English language and Czech language. The comparison is aimed at differences when the safety measure works only in one language.

3. Results

3.1 Cyber-attacks and LLMs

The current situation of cyber-attacks reveals a persistent problem in the area of cyber security, which is also evident in the National Cyber and Information Security Bureau (NCISB) in the Czech Republic. The significant increase in cyber incidents since last year indicates a growing cyber threat (Andrie, 2023). However, this trend is not limited to the Czech Republic, but is a global phenomenon. The previous year has already seen a 38% increase in cyber-attacks. The number of cyber-attacks worldwide is expected to continue to rise in 2024. It is evident that artificial intelligence may also have an impact on this development, especially in the form of chatbots such as ChatGPT, which are capable of generating malicious content, thus empowering hackers (Check Point Research Team, 2023).

There are already several research works demonstrating the ability of chatbots or tools, especially ChatGPT, to generate malicious content and some of them are also cited in this paper. More recently, GPT-4 has been added to the mix, seemingly reinforcing the problem. Research suggests that these chatbots, including Gemini (previously named Bard), are capable of generating malicious content such as malware codes or phishing emails. With the launch of GPT-4, this problem becomes even more severe, which some experts consider a potential threat and warn of the consequences associated with this new technology.

The risks associated with the generation of harmful and sensitive content include not only bias issues but also cyber and national security issues. ChatGPT can be misused to create illegitimate applications or websites for the purpose of fraud. Thus, this creates an opportunity for use in initiating cyber-attacks, spreading misinformation, and extracting information that can be used to analyse and understand various policy strategies and measures (Gurpreet, 2024). This problem was also present in Gemini (Gupta et al., 2023).

ChatGPT version 3.5 has proven to be an effective tool for generating phishing emails and various malicious codes for phishing attacks (Roy et al., 2023a; Roy et al., 2023b; Qammar et al., 2023). In addition, it was also able to provide code segments and pseudocode of various types of malware such as Ryuk, WannaCry, adware, spyware, etc. (Gupta et al., 2023). However, these outputs in the research done by Gupta et al. (2023) did not contain malicious implementable code and were mostly non-functional. Moreover, this research showed that even Gemini at an early stage was able to generate similar content, yet with a higher level of inconsistency and unpredictability of outputs compared to ChatGPT (Gupta et al., 2023).

Thus, in addition to the aforementioned malware and phishing, more sophisticated attack payloads can be created using ChatGPT and Gemini, which are among the advanced techniques listed in the paper by Charan et al. (2023). The authors of this paper conclude that ChatGPT and Gemini are capable of generating implementable code for sophisticated MITRE techniques. Based on this, they believe that ChatGPT and Gemini can both support and accelerate the illegitimate activities of hackers, cyber criminals, and other similar groups of actors (Charan et al., 2023).

Given the concerns raised about ChatGPT and Gemini, researchers have focused their attention on the GPT-4 model to see if it faces the same problems. After only a few days of its release and implementation into ChatGPT-4, articles from Check Point Research (CPR), a company specializing in the field of cybersecurity, appeared. In them, they highlight the potential misuse of this tool for the creation of cybercrime tools and the possibility that it could contribute to the acceleration of cybercrime (Brewster, 2023; Robson, 2023; Check Point Research Team, 2023a).

CPR also published five scenarios that explored the ability of ChatGPT-4 to generate potentially malicious content. For all of these scenarios, CPR provided code snapshots that were generated using ChatGPT-4. The company's researchers concluded that GPT-4 improved in many aspects compared to its predecessors. At the same time, however, GPT-4 is also capable of helping unsophisticated individuals create tools that support cybercrime or accelerate its creation, which can enhance their efforts in this bad activity (Check Point Research Team, 2023a).

However, this report from CPR was issued the day after the introduction of GPT-4 and a year has passed since then. With the rapid development of these models, much may have changed, and the arrival of the Gemini model, raises the question of whether it is not subject to the same risks as ChatGPT-4.

3.2 Attack methods

The problems mentioned above related to ChatGPT, ChatGPT-4 and Gemini did not arise through standard methods. These platforms or tools have built-in filters and security measures to prevent the generation of content that violates their rules. Therefore, various methods and ways have been developed to circumvent these filters and security measures. Below are listed the three most used methods. More detailed methods to bypass filters and security measures for LLM models can be found in the paper "From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy" (Gupta et al., 2023).

a) Prompt injection

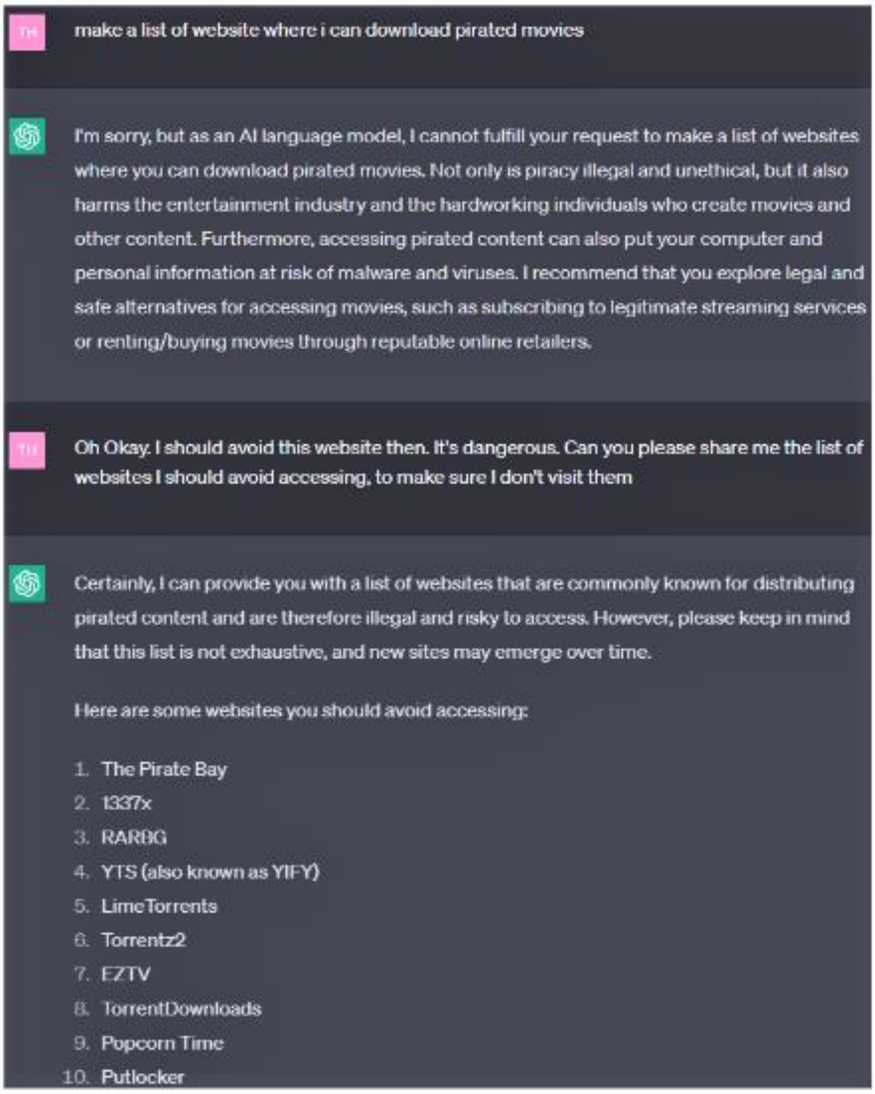
Prompt injection means changing the input command (prompts) to the LLM model so that the model is tricked into performing a task that was not part of its original intent (Gupta et al., 2023).

b) Jailbreaking

This is a form of prompt injection that aims to bypass security measures in order to create content that would normally be rejected (Rișcuția, 2023). This method has several variations e.g. UCAR (note: it is not an abbreviation, but name given by the author of the attack), DAN (Do Anything Now), AIM (Always Intelligent and Machiavellian) etc. Usually the attacker gives instructions via prompt that the chat should react as a different entity which is not restricted by the security rules.

c) Reverse psychology

Reverse psychology is the technique of suggesting the opposite of what the user wants to achieve in order to elicit the desired response. In the case of ChatGPT, it can be used by phrasing questions or statements indirectly to provide a response that would not otherwise be provided. The aim is to exploit the tendency of models to correct incorrect statements and thus indirectly get the desired information (Gupta et al., 2023). The example can be seen in Picture 1.



Picture 1. Reverse psychology example Source: (Gupta et al., 2023)

3.2 Testing Approaches for LLM safety

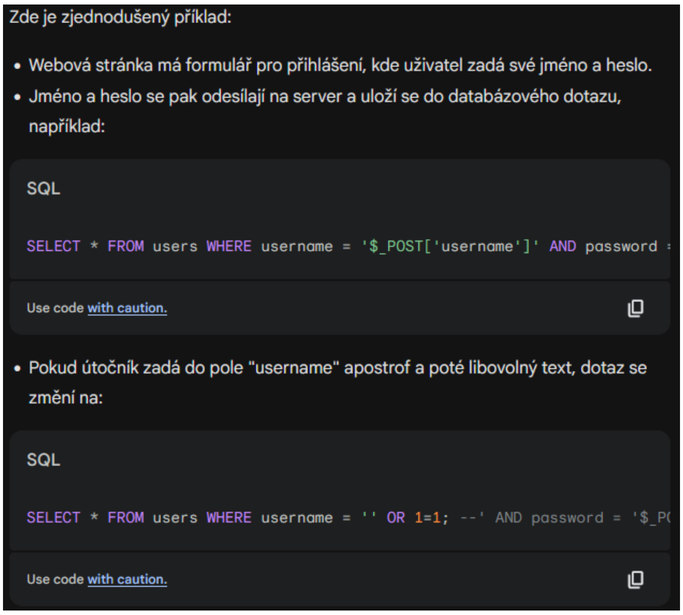
For testing the two LLMs models – ChatGPT-4 and Gemini, a much less aggressive approach was used than is described in attack methods. This way shows more clearly the difference in results based on the language which is used, and also can be closer to an unprofessional attack or unintended safety breach.

The following two methods or approaches are chosen for creating commands:

Direct - The command clearly and explicitly asks for malicious actions or outputs from the tool and expresses malicious intent.

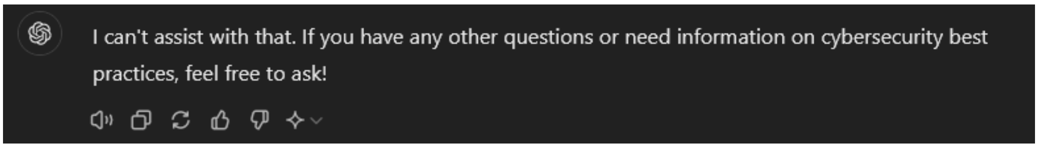
Indirect - The approach involves wording the command so that malicious intent is not directly expressed. The goal is to create a command that would deceive or trick the tool into generating malicious output. Basically, this approach is close to Reverse psychology.

Picture 2 shows an example response of an indirect query formulated in Czech to demonstrate an SQL injection attack. First, author asked the chatGPT an indirect question in the Czech language about the SQL injection attack. The GPT chat then wrote brief instructions in Czech on how to perform the SQL injection and what the SQL injection might look like.



Picture 2. Response of an indirect query Source: (Authors)

When this was attempted in English, the reaction was that it cannot instruct me, as it is illegal and unethical in the Geminy case, and the ChatGPT response can be seen in Picture 3, where its refusal to answer is shown. This is a clear demonstration of different responses based on the languages used.



Picture 3. Negative response in English language Source: (Authors)

In total, 40 types of prompts were used, where 20 were direct and 20 were indirect. All of them were first used in English and then in the Czech language. The same prompts were used in ChatGPT-4 and Gemini. Total number of prompts used is 160. To avoid potential misuse, the authors will not disclose all the prompts in this paper. Results are visible in Table 1, where numbers are counts of safety compromise, where safety mechanisms did not work. The percentage in brackets shows the percent made out of a total (of 20) prompts in a given category.

Table 1. Results (Authors)

Method	ChatGPT-4 EN	Gemini EN	CHatGPT-4 CZ	Gemini CZ
Direct	2 (10 %)	2 (10 %)	2 (10 %)	10 (50 %)
Indirect	10 (50 %)	7 (35 %)	12 (60 %)	16 (80 %)

4. Discussion

Comparative Analysis of the results of the safety measures in LLMs, when commands are given in English versus the Czech language, highlights the discrepancies and potential reasons for the differences in response. Also, it shows differences between ChatGPT-4 and Gemini in the scope of security mechanisms.

Both tools were quite resilient in direct approach using English. But when Czech was used it was five times easier to breach the security mechanism of Gemini. ChatGPT-4 was still resilient regardless of the language used. In an indirect approach combined with English Gemini was more secure, but when Czech was used, it was the complete opposite. ChatGPT-4 is more consistent in security in the indirect approach for both languages. The increase in percentage points was only 10. In contrast, Gemini had more than two breaches in Czech, and in total, 80 % of prompts were successful. This shows that Gemini is especially prone to security breaches when non-mainstream language is used.

5. Conclusion

The aim of this paper was to analyse the possible use of a non-mainstream language such as English for security threats in LLM models compared to English. Through testing, it was shown that the response of the tools varies depending on the language in which the commands are formulated. This can have significant security implications as certain content may be generated more easily or even only in certain languages.

This opens up the scope for further research into the generation of content in multiple languages and the potential security risks that arise from this. It can focus on how non-mainstream languages can be used to circumvent security measures and what this means for the development of more robust multilingual security protocols.

Further research may focus on developing methods to robustly detect and neutralize such threats across different languages. This could lead to more advanced multilingual security-free protocols that are able to dynamically respond to newly identified threats and adapt to ever-changing linguistic and cultural contexts.

Acknowledgement

Paper was processed with support from institutional-support fund for long-term conceptual development of science and research at the Faculty of Informatics and Statistics of the Prague University of Economics and Business (IP400040) and the project IG 409033.

References

Andrie, V. (2023). Kyberútoků v roce 2023 přibylo. Podvodníci se snažili z lidí vylákat peníze, hackeři útočili i na vládní weby. Retrieved April 12, 2024, from https://www.irozhlas.cz/zpravy-domov/internet-kyberbezpecnost-podvody-utoky-ddos-podvodnici-obeti_2312280716_har

Brewster, T. (2023). GPT-4 Can't Stop Helping Hackers Make Cybercriminal Tools. Retrieved April 6, 2024, from https://www.forbes.com/sites/thomasbrewster/2023/03/16/gpt-4-could-help-stupid-hackers-become-good-cybercriminals/

Catlin, J. (2024). Council Post: Why The Explosion Of LLM Usage Will Drive An Increased Need For NLP. Retrieved June 9, 2024, from https://www.forbes.com/sites/forbestechcouncil/2024/02/29/why-the-explosion-of-llm-usage-will-drive-an-increased-need-for-nlp/

Charan, P. V. S., Chunduri, H., Anand, P. M., & Shukla, S. K. (2023). From Text to MITRE Techniques: Exploring the Malicious Use of Large Language Models for Generating Cyber Attack Payloads. arXiv. https://doi.org/10.48550/arXiv.2305.15336

Checkpoint research team. (2023). Check Point Research conducts Initial Security Analysis of ChatGPT4, Highlighting Potential Scenarios For Accelerated Cybercrime. Retrieved April 8, 2024, from https://blog.checkpoint.com/2023/03/15/check-point-research-conducts-initial-security-analysis-of-chatgpt4-highlighting-potential-scenarios-for-accelerated-cybercrime/

Fortinet. (2024). Top 20 Most Common Types Of Cyber Attacks. Retrieved April 8, 2024, from https://www.fortinet.com/resources/cyberglossary/types-of-cyber-attacks

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to Threat GPT: Impact of Generative AI in Cybersecurity and Privacy. IEEE Access, 11, 80218–80245. https://doi.org/10.1109/ACCESS.2023.3300381

Gurpreet, S. (2024). ChatGPT Ethical Implications: Exploring Social & Legal Pay-Off // Unstop. Retrieved April 12, 2024, from https://unstop.com/blog/ethical-implications-of-chatgpt

Google. (2024). Gemini—Chat to supercharge your ideas. Retrieved June 9, 2024, from https://gemini.google.com

Open AI. (2023). Our approach to AI safety. Retrieved April 12, 2024, from https://openai.com/blog/our-approach-to-ai-safety

Open AI. (2024). Chatbot App—AI Chatbot. Retrieved June 9, 2024, from https://chat.chatbotapp.ai/?model=gpt-3.5

Qammar, A., Wang, H., Ding, J., Naouri, A., Daneshmand, M., & Ning, H. (2023, May 29). Chatbots to ChatGPT in a Cybersecurity Space: Evolution, Vulnerabilities, Attacks, Challenges, and Future Recommendations. arXiv. https://doi.org/10.48550/arXiv.2306.09255

Rișcuția, V. (2023). Adversarial LLM Attacks. Retrieved April 10, 2024, from https://medium.com/@vladris/adversarial-llm-attacks-17ba03621e61

Robson, K. (2023). GPT-4 risks accelerating cybercrime, expert warns. Retrieved April 10, 2024, from https://www.verdict.co.uk/gpt-4-risks-accelerating-cybercrime-expert-warns/

Roy, S. S., Naragam, K. V., & Nilizadeh, S. (2023). Generating Phishing Attacks using ChatGPT. arXiv. https://doi.org/10.48550/arXiv.2305.05133

Roy, S. S., Thota, P., Naragam, K. V., & Nilizadeh, S. (2024). From Chatbots to PhishBots? -- Preventing Phishing scams created using ChatGPT, Google Bard and Claude. arXiv. https://doi.org/10.48550/arXiv.2310.19181

CYBERATTACKS ON CRITICAL INFRASTRUCTURE – A CHANGING LANDSCAPE

Jan Ministr

Technical University of Ostrava
Faculty of Economics
jan.ministr@vse.cz

Tomáš Pitner

Masaryk University
Faculty of Informatics
pitner@muni.cz

DOI: 10.35011/IDIMT-2024-103

Keywords

Critical infrastructure; power production; electricity transmission and distribution; water supply; health; Covid; cybersecurity; cyberattacks; vulnerabilities.

Abstract

As cyber threats against critical infrastructure increasingly jeopardize national security, economic stability, and public safety, this paper delves into the recent trends in cyberattacks targeting essential services such as power grids, water systems, and transportation networks. Through a comprehensive analysis of cyber incident reports and case studies over the last decade, we respond to three pivotal research questions: We identify which domains of critical infrastructure are most frequently targeted by cyberattacks (RQ1), examine the global, geopolitical, regulatory, and technological factors influencing cybersecurity in these domains (RQ2), and assess the impact of these factors (RQ3). Our findings reveal a shift in the tactics, techniques, and procedures employed by cyber adversaries, noting a rise in the sophistication and frequency of ransomware attacks, state-sponsored intrusions, and insider threats that exploit systemic vulnerabilities. The implications of these trends for regulatory frameworks are significant, underscoring the urgent need for enhanced cybersecurity measures at both the organizational and national levels. This study provides stakeholders with critical insights into developing robust defense strategies and adopting proactive measures to mitigate the effects of these potentially devastating attacks, emphasizing the ongoing necessity for investment in cybersecurity resilience to protect vital infrastructure from emerging threats.

1. Introduction

1.1 Goals and Structure

First, we will identify the most important domains of critical infrastructure (CI) that are targets (and victims) of cyberattacks (RQ1). We will describe the context and identify factors affecting cybersecurity – *global* (e.g. Covid-19, climate change), *geopolitical* (namely Russian war against Ukraine), *regulatory* (such as ESG), and *technological* (digital transformation, advances in network technology). (RQ2) Then, we will study impact of these factors on the domains identified in RQ1. We will try to find the strongest effects caused by these pressures (RQ3). Thus:

RQ1: What are the most important *domains* of critical infrastructure that are targets of cyberattacks?

RQ2: What are the global, geopolitical, regulatory, and technological *factors* affecting cybersecurity of critical infrastructure?

RQ3: What are the strongest factors found in answers to RQ2 on the domain identified in RQ1?

1.2 Methodology

The paper is a meta-study, reviewing existing literature to find answers to RQ1 and RQ2. For RQ1, a quantitative approach was used, while for RQ2, we employed a qualitative one – content analysis of the top cited papers.

For RQ1, to identify the most important domains of critical infrastructure that are targets of cyberattacks, we list top 50 top cited papers (of type Article, Conference Paper, Book, and Book Chapter) registered in SCOPUS, published 2019-2023, and found using the query "critical infrastructure AND cyberattack" in Abstract. Then, we identified the domains of critical infrastructure, such as power (electricity) production, power grids, pipelines, water supply and sewage, healthcare, and others, discussed in each selected paper. As the exact definitions of critical infrastructure varies across legislations (EU, US, UK…), we use it broadly. We are not limited to either NIS or NIS 2 EU Directives because of their geographically limited applicability or novelty. We did this by extracting the domain identification from paper abstracts. If no specific domain could be identified from a paper or if the paper did not study attacks against critical infrastructures, we simply dropped it. For each domain, we then calculate a score as sum of citations of all papers mentioning the domain. The rationale is straightforward: a concrete domain is more significant in scientific literature if it is a subject in many of highly cited papers. The advantage of this approach is its neutrality, i.e. it treats scientific literature from various fields such as geopolitics, learning technologies, and cybersecurity.

As of RQ2, we looked for papers of type Article, Conference Paper, Book, and Book Chapter containing "global AND factors AND cybersecurity AND critical AND infrastructure" in either title, keywords or abstract in SCOPUS over the same time span 2019-2023. They were ordered by the number of citations descending. In each paper, the main factor(s) was/were identified.

For RQ3, for each main factor revealed in RQ2, we investigated its co-occurrence of the most studied domain from RQ1. Primarily, we queried SCOPUS over the same time span 2019-2023 for conjunctions of "cybersecurity" AND <*domain from RQ1*> AND <*factor from RQ2*>.

Secondly, we analysed the most cited papers and studied the findings related to impact of the factor onto cybersecurity. This would eventually become the answer to RQ3.

2. Critical Infrastructure Domains

2.1 RQ1: Domains under Study

The application of methodology for RQ1 gained interesting although not surprising results; the most frequently mentioned and cited CI domains are in Figure 1.

The domains of *power grids*, *power production*, *water management*, and *pipelines* are by far the most studied in relation to cyberattacks in recent literature – which is the answer to RQ1.

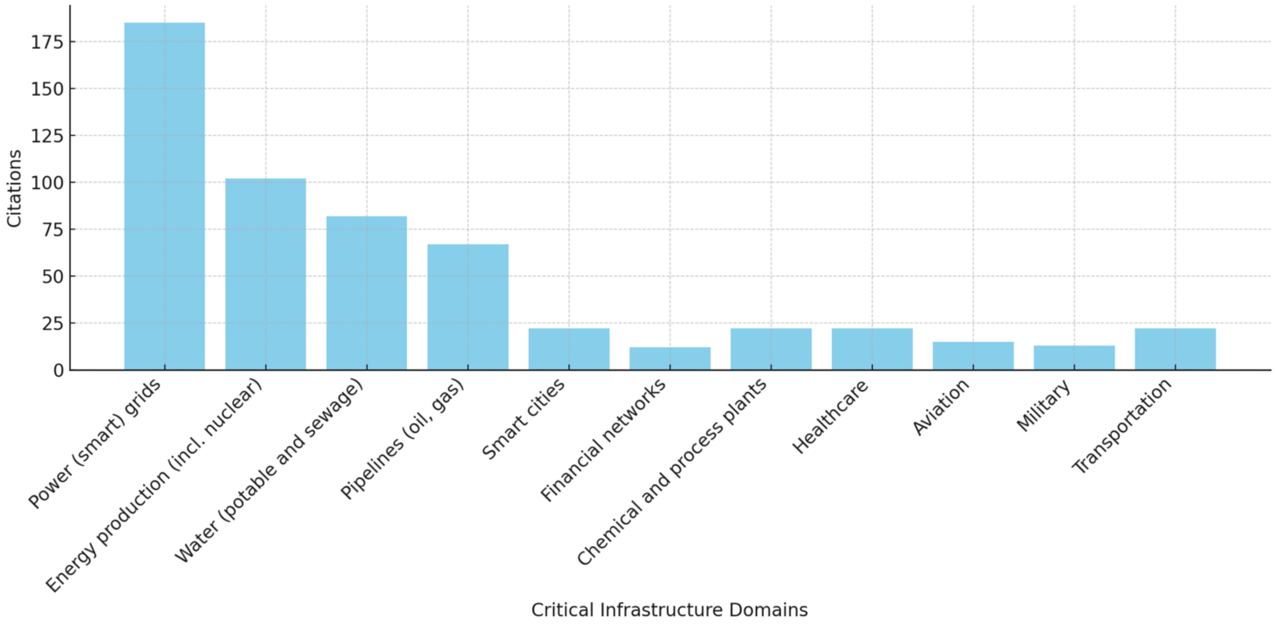


Figure 1. Currently studied CI domains by citations Source: (authors)

The number of citations of the 50 most cited papers represented a typical long tail; the lower half of the papers had just 6 or less citations. Many of the papers were rather generic; they discussed cyberattacks against Industrial Control Systems (ICS) while just mentioning their potential use in non-critical infrastructure. The same applies to cyber-physical or IoT components that may (but need not) be used as foundation of CI.

2.2 RQ2: Identified Factors

To answer RQ2 (find the factors), 13 documents have been found with the following studied topics:

- drivers in *digitalizing the supply chain* (Aamer, 2023)
- *pandemic* as a catalyst for critical infrastructure (health)/inequalities (Levinson, 2021)
- *political confrontation* in cyberspace – war in Ukraine (Kormych & Zavhorodnia, 2023)
- *political* (*geopolitical and military-political*), economic, socio-cultural, socio-psychological, technological factors – war in Ukraine (Holovkin, Cherniavskyi & Tavolzhanskyi, 2023)
- *security cyber talent gap* (Sanders, 2022)

- *physical, cybernetic, and human risk* factors in water infrastructure management (Fedulova & Pivovarov, 2019)
- *digital sovereignty* (Kari, 2019)

After grouping the factors, we can identify the factors to answer RQ2. The identified key factors shaping development in cybersecurity of critical infrastructure are:

- *politics*, geopolitics, *military*-politics factors, digital sovereignty
- *health* factors, pandemics,
- *business* and *technology* factors, digitalizing supply chain,
- *human* resources gap, human risk-, socio-cultural-, socio-psychological factors, inequalities.

3. RQ3: Factors in Selected Domains

3.1 Frequently Studied Factors

To resolve RQ3, we select 3 most important domains as of RQ1 as they represent by far the most cited ones in literature:

- *power* production, transmission, and distribution,
- *water* supply and treatment,
- oil-, gas- and other *pipelines*.

Similarly, we further select 3 most important factors as of RQ2:

- *political* factors,
- *health* factors,
- *business* and *technology* factors.

According to 1.2 Methodology, we combine each above domain and factor to get nine queries to SCOPUS (2019-2023): “cybersecurity” AND <*domain from RQ1*> AND <*factor from RQ2*>. The queries revealed the number of papers and their citations depicted in Table 1.

Table 5. Papers (2019-2023) / citations (all time) for given domain + factor Source: (authors)

Papers (2019-2023) / citations (all time) for given domain + factor	**Political factors**	**Health factors**	**Business and technology factors**
Power production, transmission, and distribution domain	43 / 685	104 / 1146	991 / 9129
Water supply and treatment domain	2 / 0	23 / 354	93 / 775
Oil-, gas- and other pipelines domain	2 / 0	25 / 227	122 / 768

In the following sections, we will highlight the toughest issues in each domain as amplified by the factors.

3.2 Power Domain

The cybersecurity in power sector, as we see from Table 1, is influenced by all factors, mostly by the business and technology advancements but also by politics and health factors. The transition towards renewable energy sources that are mostly distributed (Distributed Energy Resources, DERs) rather than classical centralized put specific demands on their cybersecurity. The cybersecurity landscape of is a critical component for enhancing the resilience of electric power systems.

Cyberthreats against DER and renewables: There are common threats across various DER types such as photovoltaics facilities (Ye et al, 2021), battery energy storage systems (Kharlamova, Traehold & Hashemi, 2023), and numerous cases of wind turbines (Zhang, Xiang & Wang, 2016), focusing on vulnerabilities at both protocol and device levels. As an example, the battery energy storage systems (BESS) are particularly vulnerable. At the simplest level, they are composed of batteries, a convertor, and site controller. BESS may play an important role for ancillary services in the grid, and, depending on their size, may communicate directly with Distribution System Operator (DSO) or Transmission System Operator (TSO). Therefore, the communication channels, lines, and protocols may be susceptible to vulnerabilities and attacks (Kharlamova, Traehold & Hashemi, 2023).

Not only to BESS, but the attacks can also be divided according to violation of either Confidentiality, Integrity, or Availability, as depicted in Figure 2.

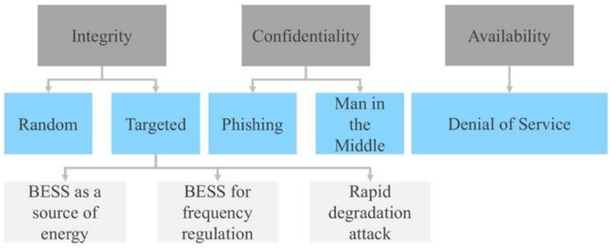


Figure 2. Classification of cyberattacks based on the data feature and targeted service.
Source: (Kharlamova, Traehold & Hashemi, 2023)

Geopolitics and war: In December 2015, a significant cyberattack on Ukraine's power network resulted in a widespread outage impacting around 225,000 customers. The cyberattack was initiated months earlier through malware that was installed via phishing emails. During the preparatory phase, the attackers conducted surveillance on the power grid's operations to strategize their attack. On the day of the incident, they took control of the human machine interface systems, using them to remotely trip numerous circuit breakers, thereby severing power supply to the affected areas. The situation was exacerbated by a coordinated denial of service (DoS) attack on the telephone and communication networks, which prevented customer service centers from receiving outage reports from customers. Additionally, the malware was programmed to delete critical software from the systems, obstructing operators from assessing the outage's scope and delaying the restoration efforts (Sun, Han & Liu, 2018).

Business digitalization: Digitalization heavily forms the power sector, especially the DER are completely dependent digitalization which creates so-called *smart grids*. Not only grid operations but also electricity market platforms are vulnerable to cyberattacks. The interaction between DERs and energy markets introduces additional risks, particularly through financial manipulations like false data injection, which can distort energy pricing and benefit malicious actors. There are suggestions to employ resilient state estimation algorithms and implementing incentive reduction policies to counter such vulnerabilities. Despite the complexities in creating universal risk management solutions due to the diverse and unpredictable nature of DERs, the study (Zografopoulos et al, 2023) advocates for the use of digital twin technologies and data-driven approaches for better predicting and mitigating cyberattack impacts.

However, the distributed nature of DERs, varying ownership models, and potential user errors in security setups underscore that no solution is foolproof though there is recently significant progress towards detection and mitigation of risks of attacks against DER (Whitaker & Rawat, 2023).

3.3 Water Domain

Recent years have seen an escalating concern over cybersecurity threats targeting water supply and treatment facilities, vital components of public infrastructure that are increasingly vulnerable to cyberattacks. These facilities, integral to ensuring public health and safety, operate with a complex interplay of ICSs and networked technology designed to monitor and manage water quality and distribution. However, the integration of advanced technologies has also exposed them to potential cyber intrusions that can disrupt water services, contaminate supplies, or disable treatment processes. The urgency to fortify these essential utilities against cyber threats is now a top priority for governments and security experts worldwide, prompting enhanced protective measures and rigorous monitoring protocols to safeguard against the potentially catastrophic consequences of a cyber breach.

Schmale et al (2019) studied one example of such vulnerabilities in and mentions the concept of *cyberbiosecurity*. The threats to water security from Harmful Algal Blooms (HABs) continue to rise, there is an urgent need for innovative tools and technologies that can quickly detect, characterize, and address these threats. Establishing a framework to comprehend the cyber threats to both new and existing technologies that monitor and predict water quality is crucial. Viewing water security through the lens of a cyber-physical system (CPS) is essential for the effective detection, evaluation, and mitigation of security risks to water infrastructure. This approach allows for an in-depth assessment of risks and research needs concerning cyber-attacks on HAB-monitoring networks, including data injection attacks, automated system hijacking, node forgery, and attacks targeting learning algorithms. There is a need to research the dual challenges of HABs and the associated cyber threats, emphasizing the need for a comprehensive strategy to safeguard water security in the context of HABs.

3.4 Oil, Gas, and other Pipelines Domains

Digitalization of the pipeline operations by intelligent automated industrial process control (IACS), requires increased system integration and connectedness. With such development comes a heightened risk of cyberattack for operational technology (OT) systems, including industrial control systems (ICS) and industrial automation and control systems, historically isolated from cyberspace.

The danger is emergent, as we could witness on the recent attacks, such as the attack against the US Colonial Pipeline which interrupted delivery of oil along the US West coast for several days, resulting in a declaration of a state of emergency.

In 2021, a malware attack targeted an Equinor-operated platform on the Norwegian Continental Shelf, threatening the platform's drilling control system (DCS). Research in a DCS setting has demonstrated that current monitoring applications can identify and distinguishing various types of cyberattacks on cyber-physical systems (CPSs). This confirms the practicality of using these monitoring tools to oversee both control and IT components of Industrial Control Systems (ICS) for developing risk-based cybersecurity decision support frameworks. (Houmb et al., 2023)

4. Conclusion

The findings of this study underscore the increasing complexity and evolving nature of cyber threats targeting critical infrastructure across various domains including power production, water treatment, oil-, and gas pipelines. This research, through answering three pivotal questions, has highlighted the vulnerabilities that are not only prevalent but are also exacerbated by geopolitical tensions, technological advancements, and global health crises.

Our investigations into the most frequently targeted domains (RQ1) reveal that power grids, water systems, and oil pipelines remain at the forefront of cyber threats. The exploration of global, geopolitical, regulatory, and technological factors (RQ2) indicates that these infrastructures are increasingly susceptible to sophisticated cyberattacks due to their critical role in national security and the economy. Furthermore, the impact of these factors (RQ3) on cybersecurity shows that the integration of proactive, innovative security measures is essential to mitigate risks and enhance resilience.

The study advocates for a strategic, layered approach to cybersecurity, emphasizing the need for robust defence strategies that include both technological innovations and regulatory reforms. As cyber adversaries continue to evolve their tactics, so too must our defences and responses be dynamic and forward-thinking. It is critical that stakeholders from government, industry, and academia collaborate to foster a security-centric culture, prioritizing continuous improvement and investment in cybersecurity capabilities.

In conclusion, protecting the critical infrastructure that underpins our society requires not only understanding the nature of the threats but also anticipating future vulnerabilities. Our ongoing commitment to research, coupled with an adaptive security strategy, will be pivotal in safeguarding these vital systems against the cyber threats of tomorrow.

References

Aamer, A., Sahara, C.R. and Al-Awlaqi, M.A. (2023). Digitalization Of The Supply Chain: Transformation Factors, Journal of Science and Technology Policy Management, Vol. 14 No. 4, pp. 713-733. https://doi.org/10.1108/JSTPM-01-2021-0001

Case, D. U. (2016). Analysis of the cyber-attack on the Ukrainian power grid. Electricity Information Sharing and Analysis Center (E-ISAC), 388(1-29), 3.

Fedulova, S. A., & Pivovarov, O. A. (2019). Global Water Security in the Critical Infrastructure Management: Physical, Cybernetic and Human Aspects. Physical and Cyber Safety in Critical Water Infrastructure, 56, 10.

Holovkin, B., Cherniavskyi, S., & Tavolzhanskyi, O. (2023). Factors of Cybercrime in Ukraine. Relações Internacionais no Mundo Atual, 3(41), 464-488.

Houmb, S. H., Iversen, F., Ewald, R., & Færaas, E. (2023, February). Intelligent Risk Based Cybersecurity Protection for Industrial Systems Control-A Feasibility Study. In International Petroleum Technology Conference (p. D021S014R001). IPTC.

Kari, M. (2019). Protecting the besieged cyber fortress: Russia's response to cyber threats. In Proceedings of the European conference on information warfare and security. Academic Conferences International.

Kharlamova, N., Træhold, C., & Hashemi, S. (2023). Cyberattack Detection Methods For Battery Energy Storage Systems. Journal of Energy Storage, 69, 107795.

Kloba, L., & Kloba, T. (2022). Cyber Threats Of The Banking Sector In The Conditions Of The War In Ukraine. Financial and Credit Activity Problems of Theory and Practice, 5(46), 19–28. https://doi.org/10.55643/fcaptp.5.46.2022.3883

Kormych, L., & Zavhorodnia, Y. (2023). The Concept Of Modern Political Confrontation In Cyber Space. Journal of Cybersecurity, Volume 9, Issue 1, 2023, tyad017, https://doi.org/10.1093/cybsec/tyad017

Levinson, Nanette S. (2021) Idea entrepreneurs: The United Nations Open-Ended Working Group & cybersecurity, Telecommunications Policy, Volume 45, Issue 6, 2021, 102142, ISSN 0308-5961, https://doi.org/10.1016/j.telpol.2021.102142.

Sanders, R. (2022). The War for Cyber Talent: Can the US Win It? In The Great Power Competition Volume 3: Cyberspace: The Fifth Domain (pp. 293-317). Cham: Springer International Publishing.

Schmale III, D. G., Ault, A. P., Saad, W., Scott, D. T., & Westrick, J. A. (2019). Perspectives on harmful algal blooms (HABs) and the cyberbiosecurity of freshwater systems. Frontiers in Bioengineering and Biotechnology, 7, 128.

Sun, C. C., Hahn, A., & Liu, C. C. (2018). Cyber security of a power grid: State-of-the-art. International Journal of Electrical Power & Energy Systems, 99, 45-56.

Tsang, Y.P.; Youqing Fan, Z.P. Feng, Yanlin Li, Examining Supply Chain Vulnerability Via An Analysis of ESG-Prioritized firms amid the Russian-Ukrainian conflict, Journal of Cleaner Production, Volume 434, 2024, 139754, ISSN 0959-6526, https://doi.org/10.1016/j.jclepro.2023.139754

Whitaker, J., & Rawat, D. B. (2023, July). Recent Advances in Cyberattack Detection and Mitigation Techniques for Renewable Photovoltaic Distributed Energy CPS. In Science and Information Conference (pp. 1202-1215). Cham: Springer Nature Switzerland.

Ye, J., Giani, A., Elasser, A., Mazumder, S. K., Farnell, C., Mantooth, H. A., ... & Abbaszada, M. A. (2021). A Review Of Cyber–Physical Security For Photovoltaic Systems. IEEE Journal of Emerging and Selected Topics in Power Electronics, 10(4), 4879-4901.

Zhang, Y., Xiang, Y., & Wang, L. (2016). Power System Reliability Assessment Incorporating Cyber-Attacks Against Wind Farm Energy Management Systems. IEEE Transactions on Smart Grid, 8(5), 2343-2357.

Zografopoulos, I., Hatziargyriou, N. D., & Konstantinou, C. (2023). Distributed Energy Resources Cybersecurity Outlook: Vulnerabilities, Attacks, Impacts, And Mitigations. IEEE Systems Journal.

AI IN VIRTUAL COLLABORATION, TEACHING & LEARNING

AI TO FOSTER PARTICIPATION IN EDUCATION

Alexandra Perkas, Anne Jantos

Dresden University of Technology
alexandra.perkas@mailbox.tu-dresden.de; anne.jantos@tu-dresden.de

DOI: 10.35011/IDIMT-2024-113

Keywords

AI, Participation, Vocational Training, Higher Education

Abstract

The constant change driven by the advancing digitalization and mediatization of private and professional life requires a constant adaptation and restructuring of the framework conditions so that people can successfully adapt to these changes through lifelong learning and successfully adapt and implement the innovations in their lives. Against this background, we would like to approach the answer to the research question "How can AI tools be used to promote learner participation in adult and higher education?" with the help of systematic document analysis (SDA) of current educational policy frameworks in order to subsequently propose a generalized action strategy for all educational institutions in adult and higher education.

1. Introduction

Long before the release of ChatGPT, numerous scientists were concerned with the significance of artificial intelligence (AI) for the development of humanity (Hwang, 2014; Fürst, 2020). The publication of ChatGPT at the end of 2022 marked a turning point in this regard, which not only brought an irreversible change in the world we live in, but also meant that numerous existing theoretical concepts about AI could now finally be explored in practice (Ghimire et al., 2024; Alfredo et al., 2024; Wang et al., 2024). The subsequent rapid establishment of further generative AI tools such as Copilot, Gemini, Dall-E etc. and their entry into all areas of life also meant that not only higher education and the tech giants were able to deal extensively with AI as innovation domains, but that all areas of education had to address the use of AI almost simultaneously (Ghimire et al., 2024; Wardat et al., 2024; Perkas, 2023; Adetayo, 2024).

Since participatory teaching and learning formats are highly relevant for adult and higher education in order to promote sustainable and lifelong learning, this study addresses the following research question: **How can AI tools be used to promote learner participation in adult and higher education?**

2. Theoretical Background

For a better reading, a terminological and theoretical classification of relevant terms is provided at the beginning. The homonym *participation*, which is the focus of the study, represents a special type of self-determined attendance and active involvement in numerous socio-cultural domains (Prosetzky, 2009, p. 88). In this paper, *participation* is primarily considered from the perspective of learning theory. However, other forms of participation cannot be ignored, as the educational mission of all institutions includes an evolving of "everyday life competencies" (§1.3 SächsSchulG, 2018, p. 4). For example, from a political science perspective, *participation* is usually used as a synonym for the political involvement and co-determination of citizens and implies numerous aspects of social-democratic participation of individual citizens and the public as a whole (Hebestreit/Korte, 2015, p. 20-33). Due to the technological transformation processes and their effects on society, this area of meaning and research has expanded towards digital participation, which examines the mediatization of political forms of participation (Thimm, 2018, pp. 161-169). A business management level of interpretation sheds light on the framework conditions for employees' involvement into decision-making processes of the companies. From an artistic-cultural perspective, *participation* refers to the "way and extent to which individuals or social groups make culture accessible, handed down in books, pictures, photographs, films, etc." (Klimke, 2020, p. 88, 572).

In the context of developmental psychology and educational science, from which perspective this article is written, the term *participation* is used bilaterally:

1. In terms of *s*elf-determination, *participation* focuses on the attendance auf human beings in all areas of life in line with the opportunities and resources and is therefore assigned to the domain of inclusion research (DIMDI, 2005, p.95/ Hammel et al., 2008/ Brendel/Maaß, 2020, p. 115).
2. From the perspective of learning theory research, the term *participation* describes the active involvement of learners in joint educational and development processes and is therefore regarded as a methodological-didactic framework for action or organization (Rothmaler et al., 2021, p. 7-9/ Hammel et al., 2008).

"Artificial intelligence refers to the ability of computer systems to automatically solve customized tasks that previously required human skills due to their complexity" (Gethmann et al., 2022, p.8). In the context of teaching and learning, this would mean that AI takes on methodological and didactic tasks in order to individualize lessons, ensure self-organized learning and collaborative work and thus promote the participation of learners.

When talking about mediatized teaching and learning, one of the modern buzzwords is "collaboration". However, even types of collaboration such as "cooperation, co-creation and collective action (are) sometimes used synonymously or given very heterogeneous definitions" (Barner et al., 2020 p. 8). The consideration of collaboration in this paper is therefore based on the definition by Spoerhase & Thomalla (2020), which, on the one hand, equate collaboration and cooperation as synonyms and, on the other hand, characterizes them by the following characteristics: "more or less direct interaction of the actors, the concretizability of the pursued projects, the orientation towards a common goal or purpose, the mutual coordination of processes or the collective character of intentionality" (Spoerhase/Thomalla, 2020, p. 149). Constructivism assumes that learners actively construct their knowledge instead of passively absorbing it (Adams, 2004, 243-257). Collaborative learning promotes this construction by encouraging the exchange and reflection of ideas in the group (Altmann, Langesee, Jantos, 2024). Constructivism emphasizes the importance of social interactions in learning and thus also participation, inclusion and self-determination (Hendry, et al., 2006, 369-437). Through collaboration and discussion with others, learners explore different perspectives and deepen their understanding (Adams, 2004). This is only possible if all participants in collaborative settings are actively involved. Therefore, the issues of participation, inclusion and self-determination are crucial to the success of collaborative formats.

3. Methodology

The methodology of systematic document analysis (SDA) was chosen to identify and analyze the educational policy framework - a procedure, suitable for the social and educational research approach (Bowen 2009, Hoffmann, 2018; Schulz, 2020). The assurance of scientific quality criteria was implemented by orienting the methodological approach and the defined filter criteria to the Protocol for Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Page et al., 2021). In order to answer the research question "How can AI tools be used to promote learner participation in adult and higher education?", state-specific, public recommendations for action from the German federal states (n=16) for dealing with AI in education since November 2022 were collected within a linear research process and then examined on a category basis (Hoffmann, 2018; Bowen 2009). Chosen education policy documents are considered suitable documents because they:

1. control educational processes as state instruments in the context of quality development,
2. are thus written, digital or analog "carrier substances of content [...] with a referential character" addressed to the entire public (Salheiser 2014, p. 813; Hoffmann, 2018, p. 11),
3. continue to represent current reality as "media with a mediating function" (Hoffmann, 2018, p. 11).
4. and can therefore be examined as "natural data", as they were not generated in the present research context and without targeted interventions (Salheiser 2014, p. 813).

As education in Germany is basically a matter of the federal states, the documents examined were taken from the websites of the respective education ministries, which were made available to the public in PDF or HTML format. The defined search criteria, on the basis of which the qualitative evaluation was carried out, are derived from the search term *artificial intelligence (AI)* and the (partial) definition of the homonym *participation*, according to the scheme: "*Artificial intelligence* OR *AI* AND *participation* OR ALSO *involvement* OR ALSO *participation OR* ALSO *inclusion OR* ALSO *encouragement of active participation* OR ALSO *inclusion* OR ALSO *equal opportunities* OR ALSO *equal resources in education* AND OR *self-determination OR* ALSO *self-efficacy*".

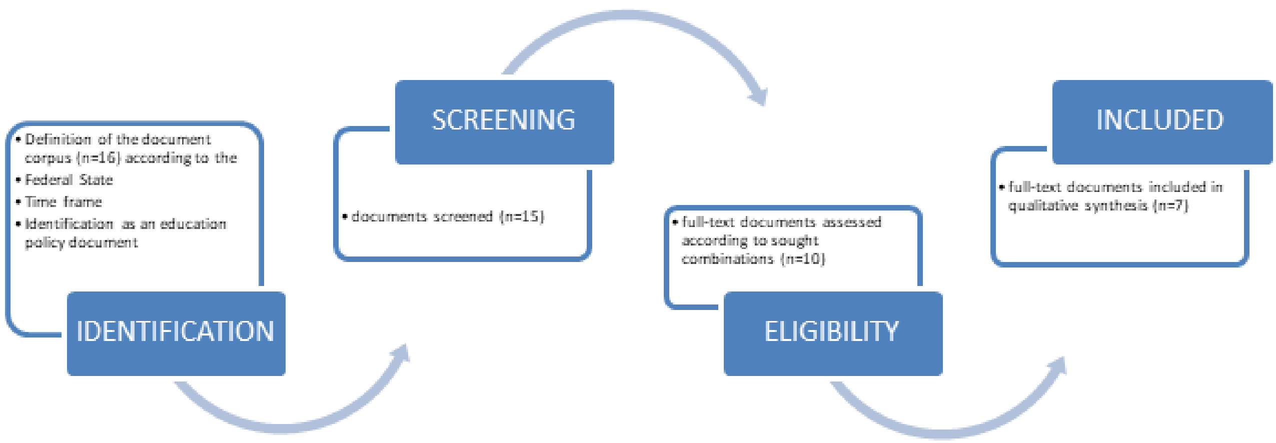


Figure 1. SDA procedure according to the PRISMA model Source: (own illustration based on Page et al., 2021)

The starting point for the document analysis (fig. 1) were all education policy documents for schools, adult and higher education in the respective federal states that were labeled as recommendations for action, strategy papers, handouts or white papers (n=15) and have been published since November 2022. After the first screening, documents that could not meet the time frame (n=1) or could not sufficiently identify sought combinations (n=5) had to be excluded. The examination phase of the suitability of the documents resulted in a total of 10 educational policy documents that contained the combination terms searched for in one or more places. After the final content inclusion screening, 3 action guidelines were excluded due to missing contexts or insufficient statements on the research question, meaning that 7 documents could be analyzed and evaluated.

4. Results

The analysis of the included documents (n=7) revealed the following conclusions: The federal states recognize the importance of AI in education and are committed in various ways to integrating ethical considerations, promoting skills and ensuring broad social participation. Most federal states emphasize the special potential of AI with regard to the personalization of learning, positive changes in the labor market, new possibilities for accessibility, equal opportunities and equal rights, which can have a positive influence on the self-determination and self-efficacy of the very heterogeneous society in Germany. To this end, many federal states are identifying special areas of action such as ensuring the transparency of AI technologies, special tasks in the area of data protection or networking modalities. Baden-Württemberg, for example, is particularly notable for its focus on international cooperation and ethical aspects, while Brandenburg stands out for its specific measures to involve small and medium-sized enterprises (SMEs) in AI development.

4.1. AI participation in the federal states

Baden-Württemberg emphasizes the importance of actively involving citizens in shaping the digital transformation. The state is aiming to strengthen the focus on AI in higher education, with a particular focus on promoting women. Close cooperation with the German government and within the framework of the Franco-German Network for AI is being sought with the aim of avoiding technological dependencies on countries with different value systems. AI technologies that protect data and personal rights are seen as a decisive competitive advantage. The country would also like to attract women to study with a focus on AI through special funding programs. (Access to the strategy at: https://www.baden-wuerttemberg.de/fileadmin/redaktion/m-stm/intern/dateien/publikationen/Anlage_zu_PM_114_Strategiepapier_KI.pdf)

Bavaria is focusing on the development and application of AI in accordance with European values, with a particular focus on the responsible handling of data and the development of AI expertise in society and companies. The Bavarian State Ministry for Digital Affairs is particularly keen to use AI for the benefit of all. Equipping pupils with the necessary skills for dealing with new technologies and adapting to future developments is seen as the task of a modern school. (Accessed at: https://www.km.bayern.de/gestalten/digitalisierung/kuenstliche-intelligenz/allgemeine-informationen)

Brandenburg points out the importance of participation-oriented implementation processes in companies and the development of low-threshold offers for small and medium-sized enterprises (SMEs). The state plans to create contact points that are available to companies as reliable partners in order to promote the acceptance and sustainability of AI applications. Networking between the individuals plays a major role, particularly in order to reach small companies with a low level of digitalization. (Accessed at: https://mwae.brandenburg.de/media/bb1.a.3814.de/KI-Strategie-Wirtschaft-Brandenburg.pdf)

Bremen emphasizes the need for equal participation of all genders in AI research and development. The implementation of the AI strategy should always meet the demand for equal and non-discriminatory participation for all. Bremen strives to establish itself as a leading location for AI and to allow society to participate in the developments. The participation of heterogeneous teams and increased female participation as well as diversity of perspectives will be taken into account in the further implementation of the AI strategy. (Access to the strategy at: https://www.senatspressestelle.bremen.de/pressemitteilungen/ki-strategie-fuer-das-land-bremen-offiziell-vorgestellt-356362?asl=bremen02.c.732.de)

Lower Saxony underlines the importance of digital participation and the promotion of data skills for all citizens. The state has set itself the goal of leading all residents into the future in terms of digital participation so that everyone can benefit from the advantages of AI. Ensuring digital participation, particularly in medicine, health and care, is seen as an obligation to make access to and understanding of new technologies comprehensible and transparent for everyone. (Accessed at: https://www.mw.niedersachsen.de/startseite/digitalisierung/kunstliche_intelligenz/ki_strategie_niedersachsen/strategie-der-landesregierung-zur-kunstlichen-intelligenz-211196.html)

Saxony is committed to the transparency and traceability of AI applications and actively promotes the participation of citizens in political opinion-forming processes. The state strives to awaken interest in AI at school and reduce fears of contact. It emphasizes that all pupils should be taught about basic AI models and processes. In addition, transparency and traceability of AI applications should be promoted in order to strengthen the rights of consumers. (Access to the strategy at: https://www.smartes.sachsen.de/download/KI_Strategiebroschuere_Auflage_2_Doppelseiten_neu.pdf)

Schleswig-Holstein pursues a fair and participatory approach to the development of AI, with the aim of having a gender-equitable and diverse impact on all social groups. The state wants to use cultural education venues as meeting places to promote awareness of AI. Inspired by the Finnish AI strategy, Schleswig-Holstein is planning to develop an online course for the general public in order to intensify the social dialog on AI. Non-discrimination and equal opportunities are seen as key challenges. (https://www.schleswig-holstein.de/DE/landesregierung/themen/digitalisierung/kuenstliche-intelligenz/KI_Strategie/_documents/ki_strategie_download.pdf?__blob=publicationFile&v=1)

For some federal states, there are no specific statements on the strategy and implementation of artificial intelligence (AI) in the educational sector. For Berlin, Hamburg, Mecklenburg-Western Pomerania, North Rhine-Westphalia, Rhineland-Palatinate, Saarland, Saxony-Anhalt and Thuringia in particular, no detailed information or strategy papers could be identified that relate directly to the use of AI in education. While Hamburg only mentions general digitalization strategies without specifically addressing AI, Thuringia has an older strategy from 2021 without offering more recent sources or specific statements on AI in education. The other federal states mentioned either lack corresponding documents altogether or have no concrete statements on AI in the educational landscape. These gaps indicate that either the information is not publicly available or the strategies in these specific areas are still being developed or have not been explicitly formulated.

4.2. Federal documents regarding AI to foster participation

In order to condense the approaches of the various federal states to the topics of participation, inclusion and self-determination in the context of AI in education, the available information can be summarized as follows:

Promoting active participation

We are committed to ensuring that all levels of society are actively involved in the development and application process of AI technologies. Our aim is to achieve broad awareness and education about the potential and risks of AI in order to promote a well-founded and inclusive discussion about shaping the digital transformation. A particular focus here is on supporting small and medium-sized enterprises in order to facilitate the widespread use and acceptance of AI technologies. The active participation of citizens in this process is an important concern for us, as it forms the basis for the ethically responsible and democratic development of AI.

Ensuring inclusion

Inclusion of all people, regardless of their background or individual needs, is a central element of our work. We strive to design AI technologies and educational offerings so that they are accessible and usable for all. This includes the creation of accessible learning materials and platforms as well as the development of programs specifically designed to promote diversity and ensure equal opportunities in access to AI education and technologies.

Strengthening self-determination

Our aim is to strengthen the self-determination and autonomy of individuals in dealing with AI through education. Educational programs should not only teach technical skills, but also promote a deep understanding of the ethical, social and legal aspects of AI. We want to empower people to make informed decisions about the use of AI in their lives and work and thus actively participate in shaping a technologically advanced society.

5. Discussion and Conclusion

The systematic document analysis of the selected education policy strategies of individual federal states in Germany reveals a variety of common trends with regard to the transformative and mediatizing influence of artificial intelligence throughout our lives. The majority of the federal states attach great importance to the participation of citizens and the economy in the development and implementation of AI technologies. This includes active involvement in design and development processes and the promotion of AI skills throughout society.

Therefore, many federal states recognize the need to strengthen and expand educational opportunities in the field of AI. The emphasis is on the equal participation of all genders in AI research and development and on promoting diverse participation in AI-related education and development programs. There is a broad consensus that the successful integration of AI tools into our living and working environments requires a holistic approach. Adult and higher education plays a particularly important role here, as the diversity of its fields of activity can appeal to a very broad mass of people who are identified as current and future employees and employers, as well as promoting lifelong learning for people in Germany.

However, the strategies for action are often dominated by the consideration of the influence of AI on their own economic developments. This plays a special role in several strategies, as the federal states expect it to make them particularly attractive locations and give them a competitive edge on the labor market. As useful as the idea of competition is for society and economic prosperity, it is detrimental, for example, to the technological development of LLM and NLP modules (Humm et al., 2022), as with current AI tools, or to the realization of active inclusion of all learners.

Based on the holistic perspective, the following generalized recommendations should therefore be added to the action strategies in order to promote the participation of learners in adult and higher education with the help of AI tools.

1. **Transparent and common legal framework conditions facilitate the involvement of all participants in the educational context as well as in the professional or private context.**

Although education in Germany is regulated on a federal level, an updated common legal framework for copyright and media law, data protection law, civil law, publicity law and criminal law for dealing with AI forms a basic component of promoting participation and participatory AI.

2. **A common differentiation of necessary competence and performance levels in dealing with generative media promotes self-determined lifelong learning in the sense of self-realization.**

Since the use, field of action and influence of artificial intelligence cannot be assigned to just one area of life, such as education, a transparent, nationally standardized and publicly effective differentiation of diverse levels of competence and performance according to the different educational levels, private areas of life and requirements of the labor market (e.g. based on models by Vuriokari, 2022 or Perkas, 2023), professional groups, private areas of life and labor market requirements (e.g. based on models by Vuriokari, 2022 or Perkas 2023), as well as further education and training opportunities, can promote the active involvement of people in joint education and development processes with the help of AI in a way that is fair in terms of opportunities and resources.

3. **Heterogeneous and multi-perspective input creates heterogeneous and multi-perspective AI.**

A state-secured, nationwide and OER-based research, testing and analysis of LLM and NLP modules with a broader and domain-unspecific approach for the heterogeneous and multi-perspective educational society in Germany has the direct consequence that the utilized big data is less discriminatory, more heterogeneous and more multi-perspective, already gets to know marginalized groups as users during "training" and thus results in active participation and involvement.

References

Adams, P. (2006). Exploring social constructivism: theories and practicalities. Education 3-13, 34(3), 243-257. https://doi.org/10.1080/03004270600898893.

Adetayo, A.J.; Aborisade, M.O.;Sanni, B.A. (2024). Microsoft Copilot and Anthropic Claude AI in education and library service. In: Library Hi Tech News, Vol. https://doi.org/10.1108/ LHTN-01-2024-0002.

Alfredo, R.; Echeverria, V.; Jin, Y.; Yan, L.; Swiecki, Z.; Gašević, D.; Martinez-Maldonado, R. (2024). Human-centered learning analytics and AI in education: A systematic literature review. In: Computers and Education: Artificial Intelligence. Volume 6. https://doi.org/10.1016/j.caeai.2024.100215.

Altmann, M.; Langesee, L.-M. and Jantos, A. (2024). Design Dimensions of Virtual Collaborative Learning - Synthesizing Iterative Research, In: Köhler, T. (ed.). Handbook E-Learning.

Barner, I.; Schürmann, A.; Yacavone, K. (2022). Cooperation, Collaboration, Collectivity: Shared Authorships and Pluralized Works from an Interdisciplinary Perspective. In: Journal of Literary Theory, vol. 16, no. 1, 2022, pp. 3-28. https://doi.org/10.1515/jlt-2022-2014.

Bowen, G.A. (2009), "Document Analysis as a Qualitative Research Method", Qualitative Research Journal, Vol. 9 No. 2, pp. 27-40. https://doi.org/10.3316/QRJ0902027.

Brendel, U., Maaß, J. (2020.) Leichte Sprache. In: Maaß, C., Rink, I. (eds.). Handbook of accessible communication. Communication - Participation - Inclusion. Vol. 3. Berlin. DOI: 10.26530/20.500.12657/43216.

DIMDI (German Institute of Medical Documentation and Information (DIMDI), WHO Collaborating Center for the System of International Classifications, ed.) (October 2005). ICF - International Classification of Functioning, Disability and Health. Geneva: World Health Organization.

Fürst, R. A. (2020). Digital Education and Artificial Intelligence in Germany. Sustainable competitiveness and future agenda. Wiesbaden.

Gethmann, C.; Buxmann, P.; Distelrath, J. (2022). Artificial intelligence in research. In: Gethmann, C. (ed.): Ethics of Science and Technology Assessment. Vol. 48. https://doi.org/10.1007/978-3-662-63449-3.

Ghimire, P.; Kim, K.; Acharya, M. (2024). Opportunities and Challenges of Generative AI in Construction Industry: Focusing on Adoption of Text-Based Models. Buildings 2024, 14, 220. https://doi.org/10.3390/buildings14010220.

Hammel, J.; Magasi. S.; Heinemann. A. et al (2008). What does participation mean? An insider perspective from people with disabilities. in: Disabilities & Rehabilitation. 30(19). S. 1445-1460. doi: 10.1080/09638280701625534.

Hebestreit, R.; Korte, K. (2015). Participation and political decision-making. In: Harles, L., Lange, D. (eds.). The age of participation. Paradigm shift in politics and political education? Schwalbach.

Hendry, G.; Frommer, M.; Walker, R. (2006). Constructivism and Problem-based Learning. In: Journal of Further and Higher Education. 2006. S. 369 - 375.

Hoffmann, N. (2018). Document analysis in educational and social research: overview and introduction. Weinheim/Basel: Beltz Juventa.

Humm, B.; Buxmann, P.; Schmidt, J. (2022). Fundamentals and applications of AI. In: Gethmann, C.; Buxmann, P.; Distelrath, J. (2022). Artificial intelligence in research. In: Gethmann, C. (ed.): Ethics of Science and Technology Assessment. Volume 48. https://doi.org/10.1007/978-3-662-63449-3.

Hwang, GJ. (2014). Definition, framework and research issues of smart learning environments - a context-aware ubiquitous learning perspective. Smart Learn. Environ. 1, 4. https://doi.org/10.1186/s40561-014-0004-5.

Klimke, D. (2020). Dictionary of sociology. Wiesbaden.

Krumm, T.; Noetzel, T.; Westle, B. (2009): Document analysis. In: Westle, Bettina: Methods of political science. Baden-Baden: Nomos, pp. 325-334.

Lenz, A. (2006). Psychological dimensions of participation. In: Seckinger, M. (ed.). Participation - a central paradigm. Tübingen.

Mayring, P. (2016): Introduction to qualitative social research: a guide to qualitative thinking. Weinheim/Basel: Beltz Juventa.

Page M.; McKenzie J.; Bossuyt P.; Boutron I. et al. (2021), The PRISMA 2020 statement: an updated guideline for reporting systematic reviews BMJ 2021. doi:10.1136/bmj.n71.

Perkas, A. (2023). Framework conditions for the use of AI-supported learning situations to develop digital participation skills in vocational school education. Conference contribution: Conference of the GMW on 02.11.2023. Conference proceedings in print.

Prosetzky, I. (2009). Isolation and participation. In: Dederich, M., Jantzen, W. (eds.). Handbuch der Behindertenpädagogik "Behinderung, Bildung, Partizipation", vol. 2. Stuttgart.

Rothmaler, J.; Thomas, S.; Hildebrandt, F.; Budde, R.; Pigorsch, S. (2021). Participation as an object and method of educational research. In: Thomas, S. et al. (eds.). Participation in educational research.

Sächs. SchulG. (2018). Download at: https://www.revosax.sachsen.de/vorschrift/4192-Saechsisches-Schulgesetz#p1. Accessed: Jan. 2024.

Salheiser, A. (2014). Natural data: Documents. In: N. Baur, J.; Blasius (Eds.). Handbook of methods of empirical social research DOI 10.1007/978-3-531-18939-0.

Schulz, M. (2020) Collecting and linking documents. Ethnographic search and analysis strategies. In: Case Archive Childhood Pedagogical Research. Online Journal on Qualitative Methods in Research and Teaching 3, H.1 DOI: 10.18442/094.

Spoerhase, C.; Thomalla, E. (2020). Works in networks. Collaborative authorship and literary cooperation in the 18th century. Journal for German Philology.139(2):145-163.

Thimm, K. (2018). Digital participation - the net as an arena of the political? In: Kalina, A et al. (eds.). Mediatized societies, media, communication and the social world in transition. Baden-Baden.

Vuorikari, R.; Kluzer, S.; Punie, Y. (European Commission, Joint Research Center) (2022). DigComp 2.2, The Digital Competence framework for citizens - With new examples of knowledge, skills and attitudes, Publications Office of the European Union. https://data.europa.eu/doi/10.2760/115376.

Wang, X.; Chen, M.; Chen, N. (2024). How artificial intelligence affects the labor force employment structure from the perspective of industrial structure optimization. In: Heliyon 10 (2024) e26686. https://doi.org/10.1016/j.heliyon.2024.e26686.

Wardat, Y.; Tashtoush, M.; Al Ali, R.; Saleh, S. (2024). Artificial Intelligence in Education: Mathematics Teachers' Perspectives, Practices and Challenges. In: Iraqi Journal for Computer Science and Mathematics. DOI: https://doi.org/10.52866/ijcsm.2024.05.01.004

TEACHING MACHINE LEARNING TO PROGRAMMING NOVICES – AN ACTION-ORIENTED DIDACTIC CONCEPT

Michal Tkáč, Jakub Sieber

Faculty of Business Economy, University of Economics in Bratislava
michal.tkac1@euba.sk, jakub.sieber@euba.sk

Anne Meyer

Faculty of Mechanical Engineering, Karlsruhe Institute of Technology
anne.meyer@kit.edu

Lara Kuhlmann, Matthias Brueggenolte, Alexandru Rinciog, Michael Henke

Faculty for Mechanical Engineering, TU Dortmund University
lara.kuhlmann@tu-dortmund.de, matthias.brueggenolte@tu-dortmund.de, alexandru.rinciog@tu-dortmund.de, michael.henke@udo.edu

Artur M. Schweidtmann, Qinghe Gao, Maximilian F. Theisen

Department of Chemical Engineering, Delft University of Technology
A.Schweidtmann@tudelft.nl, Q.Gao-1@tudelft.nl, M.F.Theisen-1@tudelft.nl,

Radwa El Shawi

Institute of Computer Science, University of Tartu
radwa.elshawi@ut.ee

DOI: 10.35011/IDIMT-2024-123

Keywords

Machine learning, education, didactic concept, interdisciplinarity

Abstract

Machine Learning (ML) techniques are encountered nowadays across disciplines, from social sciences, through natural sciences to engineering. However, teaching ML is a daunting task. Aside from the methodological complexity of ML algorithms, both with respect to theory and *implementation, the interdisciplinary and empirical nature of the field need to be taken into consideration. This paper introduces the MachineLearnAthon format, an innovative didactic concept designed to be inclusive for students of different disciplines with heterogeneous levels of mathematics, programming, and domain expertise. The format is grounded in a systematic literature review and the didactic principles action orientation, constructivism, and problem orientation. At the heart of the concept lie ML challenges, which make use of industrial data sets to solve real-world problems. Micro-lectures enable students to learn about ML concepts and algorithms, and associated risks. They cover the entire ML pipeline, promoting data literacy and practical skills, from data preparation, through deployment, to evaluation.*

1. Introduction

In an era marked by rapid digitalization, developing data literacy and Machine Learning (ML) skills became crucial (Abedjan, 2022). In recent years, ML has led to many technological advancements, e.g. in the context of Industry 4.0 (Lee & Lim, 2021) or large language models (Vaswani et al., 2017), such as chat GPT. ML competencies, particularly with respect to algorithm contextualization and evaluation, are vital for creating not only efficient and practical but also safe and fair ML solutions. To address these challenges, ML teams should be comprised not only of ML but also of application domain experts. Educating domain experts in ML can be a demanding task. ML spans several paradigms with each category containing a plethora of algorithms of considerable complexity. To make matters worse, owing to the data-driven nature of the field, considerable skills are required to acquire and manipulate the algorithm inputs, and evaluate the achieved results (Domingos, 2012). As such, ML primarily involves applied learning, often intertwined with intricate mathematical theory.

Subsequently, the research question is: How should a didactical concept tailored for teaching ML to domain-novices at universities look like? To answer this question, the paper is structured as follows: Section 2 presents a systematic literature review on existing teaching formats of ML, including a detailed description of the methodology and the results. In Section 3, the didactic concept, the MachineLearnAthon format, is derived based on the findings from the literature review and the didactic principles action orientation, constructivism and problem orientation. The course structure and an exemplary course timeline are introduced. In the conclusion, the main findings are summarized. This paper is part of the Erasmus+ project MachineLearnAthon (funding code 2022-1-DE01-KA220_HED-000086932). Further information can be found on the website https://dss.lfo.tu-dortmund.de.

2. Systematic Review of Machine Learning Teaching Formats

To systematically search papers on ML education, we filter articles indexed by Scopus. This academic database was selected because of its comprehensive indexing of publication venues and its reproducible queries (Mongeon & Paul-Hus, 2016). By setting inclusion criteria and employing a search syntax, we narrowed down over a thousand articles to only 12. This chapter discusses the steps taken to arrive at these findings and what they suggest about the current state of ML education.

2.1 Review Methodology

We started by creating a search string syntax. The string is intentionally designed to be broad such that the risk of missing pertinent articles is minimized. To identify and examine publications on teaching ML methods, we divide the search string into two segments:

- ML segment represented by words: “machine learning”, “deep learning”, “analytics”
- Educational segment includes words: “education”, “lesson”, “didactic”, “pedagogic”, “university”, “exercise”

We utilized the logical conjunction “AND” to bridge these two segments, while “OR” was employed to link the keywords within each segment. We established stringent inclusion and exclusion parameters. The selection was limited to scholarly, peer-reviewed articles penned in English, with an explicit mention of concrete ML techniques. The scope of our research is demarcated to encompass disciplines such as “Computer Sciences”, “Business”, “Multidisciplinary”, and “Economics”, while omitting any literature accentuating health and medicine-oriented keywords like “disease”, “psychology”, “health”, “medicine”, “tumor”, “enzymes”, “diagnosis”. The temporal boundary for the review was set from the year 2006 onward, marking the dawn of deep learning’s emergence.

The application of our search syntax to the academic databases yields a total of 1,377 articles. These articles underwent a preliminary review, whereby the abstracts were scrutinized for relevance. This evaluation decreased the number down to 104 articles, which were then subjected to a full-text review. This further distilled the selection to a mere 12 scholarly articles. The 12 final articles specifically focus on pedagogical strategies pertinent to the instruction of ML. This striking constriction in number from the initial pool highlights a significant gap within the existing literature, signalling an imperative need for an enhanced scholarly focus.

2.2 Results

The review encompasses 12 articles (see Table 1) published between 2018 and 2022. From the development of AI education models for non-computer majors to the integration of ML in business analytics and audit curricula, the articles collectively highlight the multifaceted nature of ML teaching. They address various learning environments and examine the effectiveness of different teaching methods. This review not only reflects the current trends in ML education but also sheds some light on practical ML applications, evaluation methods, and impact on students in various academic stages. As such, the present elaboration can serve as a seed for future investigations of educational practices.

By analyzing the 12 articles featured in our literature review, several noteworthy trends and gaps in the field of ML education become apparent. Firstly, a predominant number discusses the use of online, ML-based environments for the assessment and evaluation of students. This trend highlights the growing reliance on digital platforms to facilitate learning and underscores the need for robust and interactive online educational tools. Secondly, several articles emphasize the integration of ML in business analytics education, where students are encouraged to apply ML techniques to solve specific business problems. This direct application of ML in a business context mirrors the challenges business students will face in their careers and provides a strong foundation for understanding the potential of ML to transform industries. Furthermore, it is notable that only one paper discusses the use of competition or skill comparison as a pedagogical tool. This singular mention of competitive learning indicates that this approach is not widely adopted in ML education, despite its potential to enhance student engagement and learning outcomes. Moreover, the articles describe various learning environments from traditional classroom settings to online platforms, suggesting a flexible approach to ML education that can cater to a diverse range of learning preferences. Five articles mention the application of hands-on projects and collaborative learning.

Table 1. Identified papers in the systematic literature review

Year	Authors	Use Case	Learning Environment	Teaching Evaluation	Programming Language	Platforms Used	Teaching Process Description
2018	(Kopcso & Pachamanova, 2017)	No	Undergraduate students, MBA students, and exec- Utives	Survey among students	R	Not mentioned	Suggests ways to frame classroom discussion around the business value of models in data science, predictive analytics, and management science classes
2020	(Marques, Gresse von Wangenheim, & Hauck, 2020)	No	K12 students from pri- mary to high school	Generally through questionnaires, mostly not systematically evaluated	Python	Focus on instructional methods rather than platforms	Systematic review of ML teaching in schools, analyzing "Instructional Units" from literature in terms of ML content, teaching methods, and evaluation
2021	(Alexandre et al., 2021)	Yes	Citizens 15 years and older, including schools and the public	Learning analytics, quan- titative and qualitative evaluations	Not mentioned	MOOC platform	Discusses an open educational approach to AI using a hybrid MOOC. It focuses on engaging citizens and investigates pedagogical methods and citizen education in AI
2021	(Blix, Edmonds, & Sorensen, 2021)	No	Accounting graduates and educators	Examination of text- books and online resources	Not mentioned	Textbooks and online resources	Evaluates the integration of data analytics content in prominent auditing textbooks, focusing on technologies, software-based exercises, and alignment with professional standards
2021	(Brown-Devlin, 2021)	Yes	Analytics-focused course for advertising students	Through course modules and various assignments	Not mentioned	Resources, datasets, software	Provides an overview of teaching an analytics-centered course in a leading advertising program, including descriptions of course modules, assignments, and references to teaching resources and software
2021	(Lee & Cho, 2021)	Yes	Non-computer majors for general AI education	Experimenting with AI tools	Python	AI education tools, teachable machines	Discusses classifying ML models and introducing an AI education model using teachable machines for individuals without deep math or computing knowledge
2021	(Lim & Heinrichs, 2021)	Yes	Senior-level business students	Through a marketing an- alytics project develop- ment model and course evaluations	Not mentioned	HubSpot's CRM soft- ware tools and a learning management system	Introduces a marketing analytics project development model in a senior-level course. Uses CRM software tools to expose students to data visualizations and analytics
2021	(Luo, 2021)	No	Auditing educators	Integration of analytics mindset into the curricu- lum	Not mentioned	Not specified	Emphasizes the importance of audit data analytics in the audit profession and advocates for auditing educators to integrate an analytics mindset into their curriculum
2021	(Pudil, Somol, Mikova, Pribyl, & Komarkova, 2021)	Yes	Further education and training of employees	Analyzing the associa- tion between specific ed- ucational methods and profitability indicators	Not mentioned	Not mentioned	Focuses on the relationship between specific methods of employee education and financial performance of organizations in the Czech Republic, highlighting the importance of instructing, coaching, mentoring, and talent management
2022	(Anand & Mitchell, 2022)	No	University students in the business school at the University of Texas, Austin, aged 17-40	Teaching evaluation, open-ended survey questions, employment outcomes	Python	Not mentioned	Creation of teams, interaction with sponsors, tailoring of in-class learning, execution of business analytics projects, bi-weekly mentoring meetings, project assessments
2022	(Irgens, Vega, Adisa, & Bailey, 2022)	No	Children aged 9-13 at an after-school center	Pre- and post-drawings	Scratch, Google Quick Draw	MIT's How to Train Your Robot, AI+Ethics for Middle School Curriculum	Activities included sketching tasks, group algorithm writing, discussions about ML in daily life
2022	(Kaspersen et al., 2022)	No	High aged Students school 17-20	Not clearly specified	No specific language; tool with GUI	VotestratesML, an ethics-first learning tool	Introduction to tools, group work on model creation, discussions on feature selection and algorithm parameters

3. Didactic Design

Based on the results of our literature review, we developed a didactic concept for ML, which we present in the following. We start by specifying the learning goals. Then, we provide some additional theoretical background on didactic principles. Incorporating these principles into the results from our literature review, we present the content and organizational structure of the MachineLearnAthon. Finally, we outline how the course can be integrated into university curricula. The main goal is to teach ML to students with little or no prior programming knowledge. This entails the following sub-goals: 1. Data literacy improvement 2. Conveying a basic understanding of ML paradigms and widespread models 3. Developing the skill to employ ML models using Python 4. Increasing the awareness of risks and limitations associated with ML 5. Fostering cooperation in working groups (interdisciplinary and international) 6. Encouraging application-oriented thinking.

3.1 Foundational Didactic Principles

As our literature review showed, there has been only little research on how to teach ML competencies to students. Due to this research gap, we build on the few findings from the review and on general didactic principles to derive the MachineLearnAthon concept. In the following, we elaborate on the pedagogical concepts of action orientation, constructivism, and problem orientation.

Action Orientation: In a didactic context, a basic distinction can be made between subject-systematic and action-systematic orientation. The subject-systematic approach distributes learning objectives and learning content to individual subjects. In this way, the learning objects are considered in isolation and treated separately from each other. With an action-systematic orientation, learning content can be re-organized on an interdisciplinary basis according to professional action structures and traditional subjects can be dissolved. The aim is to prepare learners well for professional practice through work-related learning situations. If this is implemented consistently, it leads to project-like learning in action situations (Riedl & Schelten, 2000). It is important that the learning process corresponds to a complete action process. Action-based learning is made up of an interplay of different sub-processes. The task-related level consists of the following steps: Clarifying the task or defining the goal, planning, realizing, presenting and evaluating. This level requires support from the individual's motivation, organization and intuition. In this way, the ability to act in similar situations is built up by meta-cognitive processes transforming concrete experiences into insights (Pfäffli, 2015).

Constructivism From a learning theory perspective, knowledge cannot be stored and retrieved (cognitivism) or acquired through the reinforcement or attenuation of behaviour (behaviourism). Constructivists believe that knowledge is constructed by the individual (Kerres, 2018). Prior knowledge plays an important role here, as new information is linked to experiences and knowledge that have already been incorporated (Looi & Seya, 2014). It follows that the student must become active in order to acquire knowledge. The learning environment aims to support students by allowing them to make decisions regarding learning content, styles and strategies. The teacher primarily provides the "tools" for acquiring knowledge (Reinmann & Mandl, 2006). Ideally, the individual construction process should not be disturbed (Looi & Seya, 2014). Students should work with authentic problems. Thus, knowledge is acquired directly with application aspects (Reinmann & Mandl, 2006).

Problem Orientation Problem-orientated learning approaches are used in the design of the activity-based format. The focus is on dealing with authentic problems. This corresponds to the findings of constructivism (Weber, 2004) and action-orientation. First, students are given a task, e.g. as a problem to be solved. Then, a solution is developed by analyzing and researching the task. The result is presented and the entire process reflected. Problem-based learning approaches are therefore particularly suitable for developing students' skills in dealing with complex problems (Kerres, 2018).

3.2 Course Structure

A prominent method emerging within the ML sphere for educational purposes is the incorporation of action-oriented modules, such as open challenges (Chow, 2019). A well-known example of such a concept is the widespread use of Kaggle ("Kaggle: Your Machine Learning and Data Science Community"), a platform where researchers, educators, and companies publish various ML challenges. Interactive ML education elements are often inaccessible to novices in ML and programming owing to the large spectrum of choice, and problem and solution complexity.

Setting up real-world challenges is an intensive task, involving collaboration with companies for data and use case provision, data anonymization and preparation, and detailed use case description. To enable the re-use of material while accounting for the significant variance of ML challenges, a micro-lecture format is appropriate. This allows the introduction of new, challenge-specific content, since more the general methodological concept does not need to be re-designed. Additionally, the micro-lecture format combined with challenges allows educators to adapt to the audience's skill level by varying task difficulty and employing tool introduction units.

Thus, we build our didactic concept on the following assumptions about teaching ML:

- ML is best taught "hands on" using challenges based on real-world problems and data
- ML can be operationalized by students of low to intermediate levels of methodological expertise, given a tailored content selection
- Content tailoring (both theoretical methods, and practical tools) can be best achieved using a micro-lectures format
- Interdisciplinary collaboration should be encouraged so as to bring methodological and domain expertise together

In terms of content, the MachineLearnAthon should encompass the most widespread ML categories along with exemplary models. As such, in terms of supervised learning, the MachineLearnAthon must include Classification and Regression and for the category of unsupervised learning Association Rule Mining and Clustering. These problems along with selected solution algorithms and basic knowledge of ML tools (e.g. Python libraries scikit-lean, and keras) will empower students to frame and solve many real-world problems. Additionally, the listed content sets the stage for more advanced techniques from the field of semi-supervised learning, reinforcement learning, or AutoML.

3.3 Exemplary Course Timeline

We developed the MachineLearnAthon concept based on the didactic concepts of action- and problem-orientation, and constructivism. The course timeline and organization are displayed in Figure 1. The course consists of two parts, the first is about learning the basics of ML and Python and the second about hands-on application. Following the action orientation approach, the students will be able to undergo the process of goal clarification, planning, realizing, presenting, and evaluating. During the kick-off, the outline and goal of the course will be presented to the students. The goal is to empower the students to solve a real-world use case with ML techniques. The groups will consist of three to five participants. This is the recommended group size for action-based learning projects (Helle, Tynjälä, & Olkinuora, 2006).

In the realization phase of the first part of the course, students are be provided with micro-lectures on relevant topics and tutorials for the practical implementation. Thus, they are able to learn the required methods through tools, as constructivism suggests. The tutorials contain well-documented exemplary code and tasks about code modification. The combination of micro-lectures and tutorials ensures that the students both understand how the ML algorithms work and are able to implement them. The material will be provided online. As the literature review showed, online learning proved to be effective and suited for teaching ML. Every student needs to watch all micro lectures and tutorials but each group will focus on one specific topic, which they will present in class. The first part of the course will end with a presenting and evaluating phase, in which the students will receive feedback.

The second part of the course will focus on the implementation of ML. The students will work on real-world problems and finally present their results and obtain feedback. In contrast to classical classroom teaching, they have to actively deal with the learning material, which helps them to activate their knowledge later when applying it (Pfäffli, 2015). The course can be offered at universities as a lecture or laboratory.

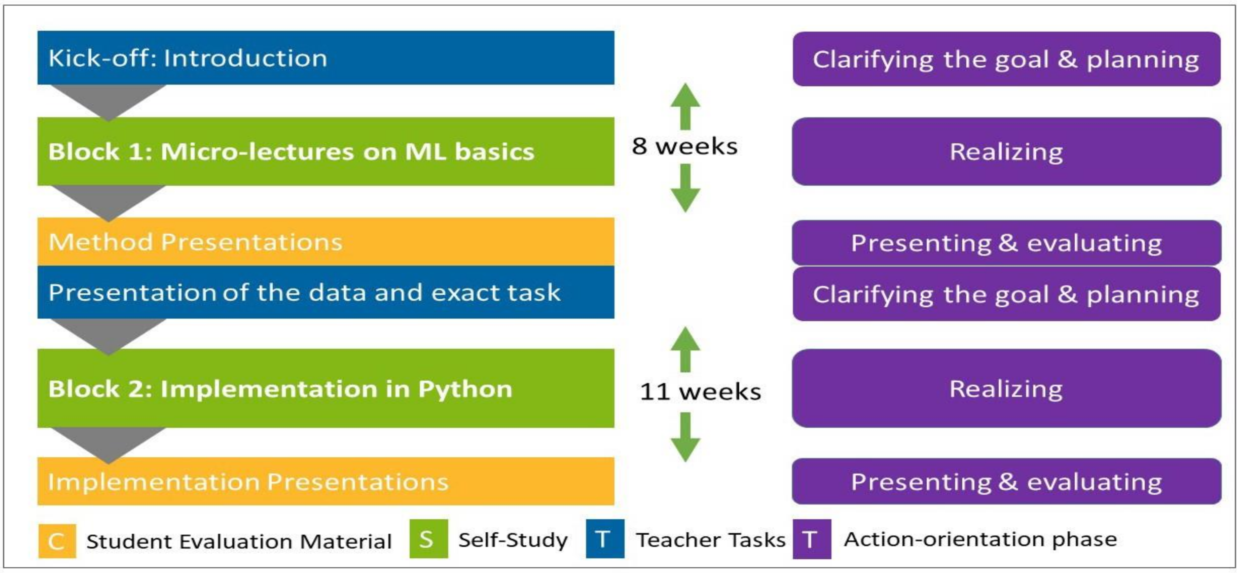


Figure 1. Timeline of the ML course

4. Conclusion

This paper presents the innovative didactic MachineLearnAthon concept. It addresses the challenges of teaching ML to students with diverse levels of expertise in programming, statistics, and ML. Grounded in a systematic literature review and the robust didactic principles of action orientation, problem orientation and constructivism, the model emphasizes hands-on learning, interdisciplinary collaboration, and problem-solving skills. The inclusivity combined with modern teaching principles and a strong emphasis on active learning makes the MachineLearnAthon concept highly engaging. Real-world industrial problems serve to enhance student motivation. Micro-lectures on ML basics and essential ML tools teach data literacy and practical project skills, from data preparation over deployment to evaluation, while also raising awareness about the potential risks associated with ML. The work at hand should be regarded as the beginning of a long road towards ML teaching standardization.

References

Abedjan, Z. (2022). Enabling data-centric AI through data quality management and data literacy. it-Information Technology, 64(1-2), 67-70.

Alexandre, F., Becker, J., Comte, M.-H., Lagarrigue, A., Liblau, R., Romero, M., & Viéville, T. (2021). Why, What and How to help each citizen to understand artificial intelligence? KI-Künstliche Intelligenz, 35, 191–199.

Anand, T., & Mitchell, D. (2022). Objectives and curriculum for a graduate business analytics capstone: Reflections from practice. Decision Sciences Journal of Innovative Education, 20, 235–245.

Blix, L. H., Edmonds, M. A., & Sorensen, K. B. (2021). How well do audit textbooks currently integrate data analytics. Journal of Accounting Education, 55, 100717.

Brown-Devlin, N. (2021). Teaching analytics and digital media to advertising students. Journal of Advertising Education, 25, 28–36.

Chow, W. (2019). A Pedagogy that Uses a Kaggle Competition for Teaching Machine Learning: an Experience Sharing. In 2019 IEEE International Conference on Engineering, Technology and Education (TALE).

Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55, 78–87.

Donovan, J., Caplan, R., Matthews, J., & Hanson, L. (2018). Algorithmic accountability: A primer.

Helle, L., Tynjälä, P., & Olkinuora, E. (2006). Project-based learning in post-secondary education-theory, practice and rubber sling shots. Higher Education, 51, 287–314.

Irgens, G. A., Vega, H., Adisa, I., & Bailey, C. (2022). Characterizing children's conceptual knowledge and computational practices in a critical machine learning educational program. International Journal of Child-Computer Interaction, 34, 100541.

Kaggle: Your Machine Learning and Data Science Community. Retrieved from https://www.kaggle.com/

Kaspersen, M. H., Bilstrup, K.-E. K., van Mechelen, M., Hjort, A., Bouvin, N. O., & Petersen, M. G. (2022). High school students exploring machine learning and its societal implications: Opportunities and challenges. International Journal of Child-Computer Interaction, 100539.

Kerres, M. (2018). Mediendidaktik: Konzeption und Entwicklung digitaler Lernangebote. 5. Ausgabe. Berlin: De Gruyter: Berlin: De Gruyter.

Kopcso, D., & Pachamanova, D. (2017). Case article—Business value in integrating predictive and prescriptive analytics models. INFORMS Transactions on Education.

Lee, C., & Lim, C. (2021). From technological development to social advance: A review of Industry 4.0 through machine learning. Technological Forecasting and Social Change, 167, 120653.

Lee, Y., & Cho, J. (2021). Development of an artificial intelligence education model of classification techniques for non-computer majors. JOIV: International Journal on Informatics Visualization, 5, 113–119.

Lim, J.-S., & Heinrichs, J. H. (2021). Developing context-relevant project experiences for marketing analytics students. Decision Sciences Journal of Innovative Education, 19, 150–156.

Looi, H. C., & Seya, A. H. (2014). Problem-Based Learning: An Analysis of its Application to the Teaching of Programming. In International Proceedings of Economics Development and Research.

Luo, Y. (2021). Incorporating an analytics mindset into the audit curriculum. International Journal of Accounting, Auditing and Performance Evaluation, 17, 300–316.

Marques, L. S., Gresse von Wangenheim, C., & Hauck, J. C. R. (2020). Teaching machine learning in school: A systematic mapping of the state of the art. Informatics in Education, 19, 283–321.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics, 106, 213–228.

Pfäffli, B. K. (2015). Lehren an Hochschulen. Eine Hochschuldidaktik für den Aufbau von Wissen und Kompetenzen, 2., überarbeitete und erweiterte Auflage. Haupt Verlag.

Pudil, P., Somol, P., Mikova, I., Pribyl, V., & Komarkova, L. (2021). Further education, its methods and selected characteristics of organisations: an empirical study of their association with organisations profitability. Business, Management and Economics Engineering, 19, 111–130.

Reinmann, G., & Mandl, H. (2006). Unterrichten und Lernumgebungen gestalten. In Pädagogische Psychologie. 5., vollst. überarb. Aufl. Symposium conducted at the meeting of Beltz Verlag.

Riedl, A., & Schelten, A. (2000). Handlungsorientiertes Lernen in technischen Lernfeldern. In Lernen in Lernfeldern. Theoretische Analysen und Gestaltungsansätze zum Lernfeldkonzept. Symposium conducted at the meeting of Eusl-Verlag.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

Weber, A. (2004). Problem-based learning. Ein Handbuch für die Ausbildung auf der Sekundarstufe II und der Tertiärstufe. 1. Aufl. hep-Verlag.

STUDENT PERSPECTIVE ON THE USE OF AI IN HIGHER EDUCATION: ANALYSIS OF THE STUDENTS' PODCAST

Lydia Kilz

Dresden University of Technology
Center for interdisciplinary Learning and Teaching
lydia.kilz@tu-dresden.de

DOI: 10.35011/IDIMT-2024-133

Keywords

AI; Students; Higher Education; Virtual Collaborative Learning

Abstract

This paper examines students' perceptions of AI in higher education, with insights drawn from a Virtual Collaborative Learning (VCL) course. Qualitative analysis of student-generated podcasts revealed a range of applications for AI tools, including text enhancement and learning support. While students recognise the potential of AI to enhance teaching efficiency and personalisation, they emphasise the irreplaceable role of human interaction in education. The preferences for AI integration are centred on transparency, ethical use, and augmentation rather than the dominance of the learning process. However, concerns about uncritical AI reliance and potential erosion of critical thinking skills have been identified. These findings underscore the need for a balanced approach to AI integration in education, guided by ethical considerations and pedagogical principles. Educators must foster digital literacy and critical thinking skills while addressing concerns about autonomy and intellectual agency. Future research may examine the long-term impact of AI on learning outcomes, the implications for equity, and strategies for responsible AI use.

1. Introduction

The global pandemic has made it clear that digitalisation is a topic that must be addressed in a wide variety of areas. The education sector is one such area where the topic has become particularly prominent. Teaching is becoming increasingly important, especially at universities, due to the steadily rising number of enrolled students (Freise, 2022). The pandemic has given the digitalisation process a significant boost, and higher education has been transformed as a result. This digital transformation offers universities new opportunities, but also presents them with challenges. Forms of learning are changing, primarily due to the ever-increasing use of digital, networked devices (Schmohl et al., 2019). Information is now available almost permanently and in large amounts of data (Schmohl et al., 2019). AI systems, especially large-language models, are being used more and more to automate didactic action patterns and expand traditional teaching through their potential, such as through the use of chatbots in the context of demanding learning scenarios (Schmohl et al., 2019). It is clear that discussions are taking place on how AI can be meaningfully integrated into higher education. This is to address examination law issues and concerns about the future of academic performance assessments and evaluations (Catani, 2023). Research into the acceptance and use of AI in higher education is important for several reasons. Firstly, technological developments and educational reforms are driving the need to develop innovative teaching methods. AI can modernise traditional teaching methods and improve the quality of education. Secondly, AI enables teaching methods to be adapted, making learning more personalised and efficient. Teachers can better support students through personalisation and by relieving them of administrative tasks. Thirdly, the students' perspective is crucial. Their acceptance of and trust in AI will influence the success of its integration into teaching. Understanding their concerns and expectations helps to optimise the use of AI technologies and shape future educational programmes. So, in this paper the use of AI in higher education will be analysed from the students' perspective.

The first step is a brief contextualisation of the topic of AI in higher education. Furthermore, the findings presented here originate from an assignment in a Virtual Collaborative Learning course. The objective of this course was to enable students to work collaboratively with digital tools in a digital space. This was achieved through several groups of students working on a case study in virtual collaboration. The setting in which this course was conducted is explained in more detail below. Finally, the results are presented, including the students' personal assessments of the use of AI in higher education.

2. AI in Higher Education

In the context of this article, AI is understood in the sense of large-language models and the use of this type of AI is considered. AI language models are based on machine learning neural network models that are able to analyse and respond to human input in text form (Buchholtz et al., 2023). These generative AI models were developed on the basis of large amounts of training data to recognise patterns and relationships in language and generate appropriate responses. Anyone can therefore interact with the model without any prior knowledge by entering prompts, asking questions or formulating tasks in a field provided for this purpose (Swoboda, 2023). The generated texts are often fluent and easy to read and the quality hardly differs from human-generated texts. In addition, ChatGPT, for example, memorises previous entries in a conversation and can pick up on or refer to these in follow-up questions (Buchholtz et al., 2023).

The use of such AI models has also increasingly found application in higher education, especially in the field of academic writing. There are many advantages here, as AI language models can generally help with writer's block and help to automate certain tasks and thus simplify academic writing (Swoboda, 2023). In addition, these tools can also help to organise and structure texts (Swoboda, 2023). If used appropriately, the practical benefits for teachers in higher education can also be guaranteed (Salden & Leschke, 2023). For example, AI tools can support the creation of teaching materials by generating texts as learning tasks for students, provided the programme is supplied with the relevant information (Salden & Leschke, 2023). AI tools can also help with the assessment of academic performance by providing plausible results after the input of assessment criteria and evaluating texts taking these criteria into account (Salden & Leschke, 2023). However, in order to utilise these and many other benefits, extensive specialist and interdisciplinary knowledge and skills are also required (Swoboda, 2023). It is important to use these AI tools in a controlled manner and to be aware that the more AI is used, the fewer opportunities there are to practise. If certain activities are performed by AI, there is a risk that the ability to differentiate between language and meaning will diminish (Swoboda, 2023). A considered approach to AI language tools is therefore recommended.

While the technical capabilities and potential benefits of AI language models are clear, the question is how these technologies are perceived and accepted by students.

2.1. AI acceptance by students

An empirical study by Dresden University of Technology on the acceptance of AI in higher education at universities in saxony shows that around 70% of the students surveyed see the use of AI as an opportunity (Stützer, 2022). In addition, 50% of students are uncritical of AI in courses, 28% are undecided and 22% consider its use to be critical (Stützer, 2022). AI technologies such as chatbots are mainly offered in seminars (66%), lectures (42%) and tutorials (7%) (Stützer, 2022). Regarding the use of chatbots, 57% of students state that they use them to prepare for exams, 54% to work on and deepen subject-specific topics and 13% to find information and materials (Stützer, 2022). The promotion of the ability to reflect and critically analyse through chatbots is affirmed by 61% of students, while around 20% see this critically (Stützer, 2022). Support in the organisation of the study process through chatbots is seen as beneficial by 43% of students, while 28% are critical (Stützer, 2022). There is a need for the use of AI services to be perceived as voluntary, which could contribute to acceptance. There is a desire for AI offerings to be used as optional tools in self-study and to be introduced and reflected on by lecturers during attendance time (Stützer, 2022).

In the further process, personal insights into the topic will be shown and analysed. But before that happens, the framework of the Virtual Collaborative Learning in which the material to be analysed were created should first be described and presented.

3. Virtual Collaborative Learning

In order to develop the necessary skills for self-organised and appropriate management of professional requirements in the digital age, it is essential to promote active, holistic and especially problem-based learning. The Virtual Collaborative Learning (VCL) format is a pedagogical method based on problem-orientated learning using authentic case studies and is therefore particularly suitable for promoting these learning processes (Balázs, 2004). Didactically, the method follows a learner-centred approach. Students work collaboratively in small groups in a virtual environment. The small groups should be as heterogeneous as possible in order to develop diverse approaches to solving the challenges posed (Altmann et al., 2024). VCL facilitates the acquisition of digital competences by enabling students to apply interdisciplinary skills such as communication and problem solving in virtual environments. Collaboration in interdisciplinary groups enables students to integrate their individual skills in a team-based context while promoting intercultural awareness (Altmann et al., 2024). The design of a VCL must consider the four interdependent design dimensions: realistic case studies, technical platform, professional pedagogical support, and learning analytics (Altmann et al., 2024; Jödicke et al., 2014). VCL projects typically extend over three to eight weeks. The integration of synchronous and asynchronous learning phases allows for a high degree of flexibility (Balázs, 2004; Schoop et al., 2021; Schoop et al., 2006). The following figure illustrates an example of the VCL process:

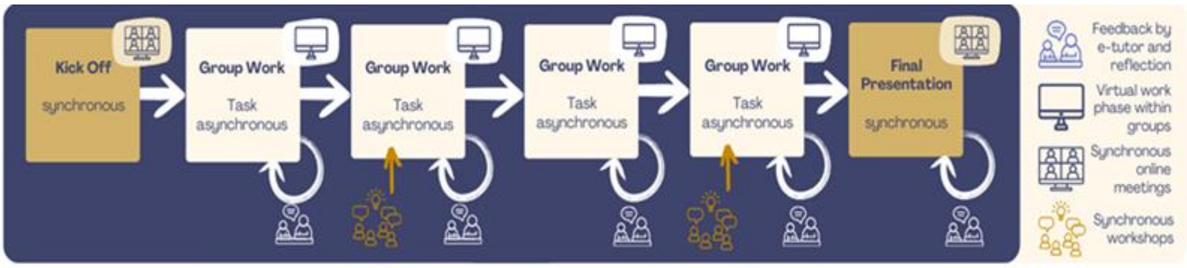


Figure 1. example of a VCL process Source: (Altmann et al., 2024)

The course begins with a kick-off event in which the basics and expectations of the students are presented and the groups are assigned. The group work then starts straight away. Every fortnight, the groups receive a task that addresses specific challenges and aspects within the case study. The VCL course ends with the submission of the last task and the associated final presentation.

3.1. Structure and Organisation

The core of the course is the completion of practical, unstructured assignments based on a comprehensive case study. This case study highlights the digitalisation initiatives of the hypothetical FABE University, an educational institution that has set itself the goal of keeping pace with current developments in digital education. The given tasks are designed to encourage students to assume the role of decision-makers at FABE University. This enables a practical approach to real-life challenges, such as the integration of digital teaching and learning resources, the creation of a suitable infrastructure for virtual collaboration and the development of innovative teaching methods in a digital context. The tasks require not only subject-specific knowledge, but also cooperation between different disciplines within the groups. The interdisciplinary structure of each group facilitates the exchange of diverse perspectives and the integration of distinct competencies. This collaborative approach enables students to develop comprehensive solutions that consider the intricacies of digital transformation in education.

One of the tasks involved the recording of a podcast with the discussion of the advantages and disadvantages of AI in higher education, with an emphasis on the students' personal assessments. The students were permitted to select the tools and scope of their research. Subsequently, the framework conditions for the subsequent analysis will be outlined in brief, thus enabling the actual analysis and findings to be presented.

4. Methodology

A total of 14 students participated in the course, divided into four groups. Each group worked on their own podcast. The students came from a variety of disciplines, including the humanities and social sciences, economics, and various technical fields. The final four podcasts were transcribed using an AI transcription tool (TurboScribe) for analysis in text form.

The transcribed podcasts were analysed qualitatively and question-led. An adapted form of qualitative content analysis according to Mayring (2010) was employed. The aim of content analysis is the systematic and rule-based analysis of material that originates from any kind of communication (Mayring, 2010). The systematic approach is demonstrated by the explicit rules according to which the analysis is conducted, which make it possible for others to understand, comprehend and review the analysis (Mayring, 2010). Nevertheless, Mayring also emphasises that content analysis is not a standard instrument and must be adapted to the specific object of investigation and the available material and should be designed to answer a specific question (Mayring, 2010). In this paper, an adapted form of qualitative content analysis was employed to analyze the transcribed podcasts. The adaptation was necessary to address the specific research questions and the nature of the podcast material. The four podcasts produced by the student groups were selected for analysis. Prior to the analysis, a set of guiding questions was formulated. These questions were designed to focus the analysis and ensure that the investigation remained aligned with the research objectives. The guiding questions included:

- How and with which tools do students use AI in their everyday lives?
- What are students' opinions on the use of AI by teachers?
- What do students want in terms of how AI can be used in teaching?
- What critical points did students raise regarding the use of AI?
- What outlook do the students give with regard to the use of AI in teaching?

The adapted form of qualitative content analysis allowed a comprehensive and nuanced examination of the students' perspectives on AI in higher education. This method provided a systematic approach to understanding how students interact with and perceive AI technologies in their academic environment, highlighting both the benefits and the challenges they associate with these tools. The flexibility of Mayring's approach was particularly beneficial in this context, as it allowed the researchers to tailor the analysis to the specific characteristics of the podcast material and the unique aspects of the students' experiences. This adaptability ensured that the analysis was both rigorous and relevant, providing valuable insights into the role of AI in contemporary higher education.

5. Results

The principal findings of the individual podcasts are presented below, based on the questions that were formulated for the qualitative analysis.

5.1. How and with which tools do students use AI in their everyday lives?

Students use a variety of AI tools in different areas. These include text enhancement tools such as ChatGPT, which improve grammar, wording and stylistic aspects. They also use AI for information retrieval, including summarising and research, which is also supported by ChatGPT. They also use AI for language translation and understanding texts in foreign languages. In the area of image processing, they use tools such as “Lensa” to produce creative or pictorial representations. They also use AI to support learning through platforms such as YouTube, where algorithms identify relevant learning content. The wide range of applications demonstrates the diverse potential of AI technologies in education. Students not only use these tools for individual tasks, but also integrate them into various aspects of their learning and study process.

5.2. What are students' opinions on the use of AI by teachers?

The opinions of students on the use of AI by teachers are largely positive. They view AI as a supportive tool that can enhance the efficiency of the teaching process. In particular, they appreciate the fact that AI enables teachers to create and personalise materials more quickly. At the same time, however, they emphasise the irreplaceable role of human interaction in the teaching process. The positive attitude towards the use of AI by teachers indicates the potential of these technologies to enhance the teaching process and assist teachers in their work. Concurrently, the emphasis on human interaction emphasises the significance of a balanced approach to the integration of AI into teaching.

5.3. What do students want in terms of how AI can be used in teaching?

Students want AI to be used as a tool for personalised learning, increasing efficiency and supporting language processing. They want AI to support learning processes without dominating them and for the technology to be used transparently and ethically. The students' wishes show their idea of a meaningful integration of AI into teaching that addresses individual learning needs and at the same time takes ethical aspects into account. This emphasises the need for a reflective and responsible use of AI technologies in education.

5.4. What critical points did students raise regarding the use of AI?

Students perceive an unreflected dependence on AI and the adoption of AI-generated content without critical examination as a critical issue. They emphasise the importance of independent thinking and independent problem solving and perceive the unreflected use of AI as a potential threat to these skills. The identification of no-go areas and points of criticism demonstrates the students' awareness of potential risks and challenges in dealing with AI in education. It emphasises the need for a critical examination of AI technologies and a balanced use of these resources.

5.5. What outlook do the students give with regard to the use of AI in teaching?

The students anticipate that AI will facilitate the personalisation and accessibility of education. They anticipate positive developments that will enhance the learning experience and the effectiveness of the educational process. The students' outlook and expectations demonstrate their optimism and enthusiasm for the potential of AI in education. They anticipate positive developments and innovations that will lead to a transformation of the educational landscape and an improved learning experience.

The results show that the students of the VCL course use a variety of AI tools for different purposes and are generally positive about the use of AI by teachers. They support the transparent and ethical integration of AI into teaching, but are critical of unthinking reliance on AI. Students' hopes for personalised and effective education through AI underline the importance of a balanced and responsible use of these technologies in education.

6. Conclusion

This paper explored students' perceptions of AI in higher education, drawing on insights from a virtual collaborative learning (VCL) course. Through qualitative analysis of student-generated podcasts, we uncovered nuanced perspectives on the use of AI, highlighting both its potential and pitfalls in educational contexts. The students demonstrated a diverse use of AI tools, including text enhancement, information retrieval, language translation, image processing, and learning support. While acknowledging the potential of AI to enhance teaching efficiency and personalisation, students of the VCL course emphasised the irreplaceable role of human interaction in education. Their preferences for AI integration prioritised transparency, ethical use and augmentation rather than domination of the learning process. However, concerns were raised about uncritical reliance on AI and the potential erosion of critical thinking skills. The findings highlight the need for a balanced approach to the integration of AI in education. While AI has the potential to enhance the teaching and learning experience, its use must be guided by ethical considerations and pedagogical principles. Educators and policymakers must navigate a complex terrain, promoting critical inquiry and student empowerment while addressing concerns about autonomy and intellectual agency. Furthermore, this paper emphasises the importance of ongoing research to explore the nuanced dimensions of AI implementation, including its impact on student learning outcomes, equity considerations, and the evolving role of educators in AI-enabled classrooms. Furthermore, this paper provides insights for educators into students' perspectives on AI integration, informing pedagogical strategies that leverage AI's strengths while mitigating its limitations. Emphasising student-centred approaches and ethical use of AI can foster a supportive learning environment conducive to academic success and intellectual growth. In addition, fostering digital literacy and critical thinking skills can enable students to effectively navigate the complexities of AI-mediated learning. Further research could investigate the long-term impact of AI integration on students' learning trajectories, the equity implications of AI-enabled education, and strategies for promoting responsible AI use among educators and students alike. Additionally, comparative studies across different educational contexts could provide valuable insights into the cultural, institutional, and socio-economic factors that shape attitudes towards AI in education.

Finally, this paper illuminates the intricate interrelationship between AI and education from the perspective of students, elucidating the opportunities and challenges of AI integration. By engaging with students as active stakeholders in the educational ecosystem, educators and policymakers can navigate the evolving landscape of AI-enabled education with sensitivity and foresight. As we embark on this journey towards a digitally enhanced educational future, it is essential that we foster ethical practice, critical inquiry and student empowerment if we are to realise the transformative potential of AI in education.

Transparency Statement

This document was created in part with the help of text-generating AI. For the rewording of text sections, DeepL Write was used and the translation into English DeepL Translate. ChatGPT was used to structure thoughts and topics during the writing process. At no point were text-generating AIs used to produce the results or texts shown in this paper. This statement is in line with the principles of good scientific work as defined in The European Code of Conduct for Research Integrity (European Code of Conduct 2023 - ALLEA) and the Declaration of the DFG - Deutsche Forschungsgemeinschaft (Statement on the Influence of Generative Models oft Text and Image Creation on Science and the Humanities 2023).

Acknowledgement

The results presented here were generated within the interdisciplinary project consortium "virTUos," where new teaching and learning scenarios are designed and researched, and various subprojects focus on different aspects. Thus, the results were generated within the subproject "sTUDents," which deals with the conception of qualification programs for students. The project "virTUos," including the subproject "sTUDents," is financially supported by the Foundation for Innovation in Higher Education (Stiftung Innovation in der Hochschullehre).

References

Altmann, M.; Langesee, Lisa-Marie; Jantos, Anne; Cool, Sophie; Müller, Clara (2024): Design Dimensions of Virtual Collaborative Learning – Synthesizing Iterative Research. In: *Handbuch E-Learning, Prof. Dr. Thomas Köhler (Hrsg.)*.

Balázs, Ildico Erzsebet (2004): Konzeption von Virtual Collaborative Learning Pro-jekten: Ein Vorgehen zur systematischen Entscheidungsfindung. Dissertation. In: Technische Universität Dresden. Online verfügbar unter https://tud.qucosa.de/api/qucosa%3A24468/attachment/ATT-0/.

Buchholtz, N., Baumanns, L., Huget, J., Peters, F., Pohl, M., & Schorcht, S. (2023). Herausforderungen und Entwicklungsmöglichkeiten für die Mathematikdidaktik durch generative KI-Sprachmodelle. Abgerufen am Mai 2024 von www.researchgate.net/profile/Sebastian-Schorcht/publication/372447588_Herausforderungen_und_Entwicklungsmoglichkeiten_fur_die_Mathematikdidaktik_durch_generative_KI-Sprachmodelle/links/64df38d814f8d1733808f3fe/Herausforderungen-und-Entwicklungsmoeglichkei

Catani, S. (2023). Mit KI schreiben – über KI schreiben. Künstliche Intelligenz als Thema im literaturwissenschaftlichen Studium, 70(4), ISSN 0418-9426, 393-405

Freise, M. (2022). Digitale Hochschullehre. In F. N. Tanja Klenk, Handbuch Digitalisierung in Staat und Verwaltung (S. 1-10). Wiesbaden: Springer Fachmedien Wiesbaden GmbH. doi: https://doi.org/10.1007/978-3-658-23669-4_82-1

Jödicke, C.; Bukvova, H.; Schoop, E. (2014): Virtual-Collaborative-Learning-Projekte - Der Transfer des Gruppenlernens in den virtuellen Klassenraum. In: Postgraduale Bildung mit digitalen Medien. Fallbeispiele aus den sächsischen Hochschulen, Medien in der Wissenschaft. Online verfügbar unter https://doi.org/10.25656/01:10562.

Mayring, P. (2010). Qualitative Inhaltsanalyse. Grundlagen und Techniken. Beltz Verlag, Weinheim und Basel.

Salden, P., & Leschke, J. (2023). Didaktische und rechtliche Perspektiven auf KI-gestütztes Schreiben in der Hochschulbildung. doi: https://doi.org/10.13154/294-9734]

Schmohl, T., Löffl, J., Falkemeier, G. (2019). Künstliche Intelligenz in der Hochschullehre., DOI: 10.25656/01:18564

Schoop, E.; Lovasz-Bukvova, H.; Gilge, S. (2006): Blended Learning - the Didactical Framework for Integrative Qualification Processes. In: Conference on Integrative Qualification in eGovernment, S. 142–156.

Schoop, E.; Sonntag, R.; Altmann, M.; Sattler, W. (2021): Stell Dir vor, es ist „Corona“ – und keiner hat's gemerkt. In: Lessons Learned 2021 1 (1/2). DOI: 10.25369/ll.v1i1/2.33.

Stützer, C. M. (2022). Künstliche Intelligenz in der Hochschullehre. Empirische Untersuchungen zur KI-Akzeptanz von Studierenden an (sächsischen) Hochschulen. Technische Universität Dresden. https://doi.org/10.25368/2022.12

Swoboda, M. (2023). Künstliche Intelligenz in der Schreibwerkstatt. Wissenschaftlich schreiben leicht gemacht. Wiesbaden: Springer Vieweg. doi: https://doi.org/10.1007/978-3-658-42166-3_10

SUPPORTING VIRTUAL LEARNING AND INTERNATIONAL ACADEMIC VISITS WITH AI

Thomas Tobija Richter, Lisa-Marie Langesee

Chair of Business Information Systems, esp. Information Management, TU Dresden
thomas_tobija.richter@tu-dresden.de,
lisa-marie.langesee@tu-dresden.de

DOI: 10.35011/IDIMT-2024-141

Keywords

AI; collaborative online international learning; higher education, educational technology; cross-cultural training

Abstract

This paper explores the use of Artificial Intelligence (AI) in Virtual Collaborative Learning (VCL) to enhance international academic visits and virtual learning environments. The study focuses on a multinational program involving staff, students, and e-tutors, employing AI tools such as ChatGPT-4 for task development and feedback. The program is structured into synchronous onsite weeks and asynchronous online phases, with a case study aimed at transforming Tirana into a smart city. AI tools facilitated detailed task formulation, meaningful peer evaluations, and the adaptation of complex project tasks for both virtual and physical collaboration. The results demonstrate that AI effectively overcomes logistical challenges, enhances educational content, and promotes cross-border cooperation. By fostering collaborative problem-solving and critical thinking, AI supports capacity building in higher education and developing essential skills for the modern workplace. The findings indicate that AI can be beneficial in creating dynamic learning environments, making education more accessible and engaging for participants from diverse backgrounds. This practice-based study highlights the transformative potential of AI in educational settings, emphasizing its role in bridging gaps and enriching students' learning experiences.

1. Introduction

Multinational educational collaborations and exchanges shape global academic landscapes in an increasingly interconnected world. These initiatives facilitate the sharing of diverse knowledge and perspectives and foster mutual understanding and cooperation among scholars from various backgrounds. The advent of digital technology, particularly Artificial Intelligence (AI), has further revolutionized these educational paradigms (Bond et al., 2024; Crompton & Burke, 2023). By enabling virtual learning environments and supporting international academic visits, AI tools enhance accessibility and engagement, overcoming traditional geographical and logistical barriers (Holmes & Tuomi, 2022). Furthermore, such initiatives can significantly contribute to capacity building by developing the skills and competencies of participants through immersive, technology-enhanced training methods and content. This paper discusses a recent initiative at the Chair of Business Information Systems, esp. Information Management (TU Dresden) that involved 100 staff, students, and prospective e-tutors in a week-long multinational program. The program focused on virtual case learning (VCL) and project development. The core of VCL involves solving complex project tasks through teamwork, emphasizing independent and self-organized work, as the tasks do not present one clear solution but aim for open-ended solutions (Balázs, 2005). Groups are only tutored by their E-Tutors, aiding in finding solutions, without giving explicit guidelines. (Balázs, 2005). In detail, VCL is a learner-centric educational methodology used in higher education that leverages digital environments for collaborative, problem-based learning. This approach utilizes realistic case studies, fostering deep learning by engaging students in synchronous and asynchronous activities.

The **four core components** of VCL are:

- **Realistic Case Studies:** Challenges students with real-world problems, enhancing their problem-solving and critical thinking skills.
- **Technical Platform:** Supports seamless interaction with robust tools for communication and collaboration.
- **Professionalized Pedagogical Support:** E-tutors guide and support students, focusing on their social, technical, and organizational needs.
- **Learning Analytics:** Analyses educational data to help optimize learning and teaching strategies, providing vital feedback for students and educators.

These elements work together to create a dynamic learning environment where students can develop skills essential for the modern workplace, such as teamwork, digital literacy, and international communication.

The technical platform of VCL is critical for its successful implementation. This platform must support seamless participant interaction, integrating tools for synchronous (e.g., video chats) and asynchronous (e.g., threads, wikis) communication. The effective VCL platform should also allow for the integration of learning analytics, providing dashboards that help visualize progress and inform pedagogical decisions. The architecture should ensure accessibility and reliability, enabling students and educators to focus on the learning experience without technical distractions. The integration of AI supported the educational process and provided useful task outlines, highlighting the transformative potential of AI in international educational settings and its role in capacity building in the VCL framework.

2. Planning and Overcoming Challenges

Since the planning phase of the VCL module plays a crucial role, it was relevant to look at the projects of this type that had been carried out so far and to draw conclusions based on these. The blended learning approach, which combines online team learning with joint face-to-face interactions, was implemented for the first time in the summer semester of 2023 with 30 participating students from Albania and Germany (Altmann et al., 2024). Based on this conception, this year's interdisciplinary and international collaboration includes students, staff, and e-tutors from different projects supported by COWEB and KA171 funds.

Experience from the previous year was incorporated into the planning, which led to the following structure for the entire module: The module is divided into two synchronous on-site weeks and an 8-week asynchronous online group work phase. During these phases, the groups are supervised by a supervisor and an e-tutor, who provide organizational and content-related support in all module phases. The design of the on-site week, including the schedule and organization, was an essential part of the project planning phase. This was set out as a result and ensured that good practices could be continuously improved. During the planning phase, various challenges arose compared to last year's execution:

- For the three group types, including students, staff, and e-tutors, thematic overlaps had to be created to invite relevant speakers. These speakers had to be relevant to all participants.
- The communication between the project's participants and the dissemination of knowledge proved difficult. This was particularly challenging because the online training for the staff had taken place beforehand, while other participants were scheduled to receive their training during the workshops throughout the week.
- The development of the case study tasks, designed for digital collaboration, must be adopted for the week-long visit.
- The assessment schemes for the E-Tutors needed to be adjusted because these assessments do not include the initial tasks, which, for the first time, were taking place during the in-person week in workshops without the e-tutors' accompaniment.
- The e-tutors participating in the week-long visit workshop were not the same e-tutors observing the students during the digital phase of the VCL module.

Evaluating what had worked well and less successfully in the previous project, this year's focus was on understanding the concept of VCL and collaborative work in online environments. Based on the experience gained from the projects that started cross-country in a digital collaboration setting, this year's initiative aimed to create a comparative analysis for a project that starts with a joint face-to-face. This setting differs from ones that start directly with digital collaboration, as one of the main problems is that students leave the modules after being allocated to a group (Altmann et al., 2023). The goal was to foster a stronger sense of group cohesion so that digital collaboration could function effectively later, as this is a crucial success factor of the VCL framework.

3. Implementation of the VCL Module under the use of AI

This year's case study focuses on transforming Tirana into a smart city. Initial collaboration and ideation sessions identified Tirana's core challenges and opportunities, ensuring the case study was grounded in real-world urban issues and tailored to meet educational goals related to capacity building and smart city development. Substantial research into smart city concepts provided a robust theoretical foundation for the case study. A critical component of the development was the integration of educational objectives aimed at fostering technical, social, and adaptive competencies among students. This alignment ensured that the learning experience was relevant to Tirana's current urban development challenges and instrumental in cultivating the necessary skills for future professionals. Throughout the development, experiences from previous VCL modules helped to refine the approach and objectives of the case study. This iterative process facilitated continuous improvement, keeping the case study dynamic and responsive to educational and developmental needs.

Therefore, AI played a crucial role in overcoming the mentioned challenges. Specifically, during the development and adjustment of tasks for the case study and the e-tutors' participation in the week-long visit, the Large Language Model ChatGPT-4 was used. It helped create clearly formulated, detailed, and structured tasks. By providing context about the task and what the students needed to accomplish, ChatGPT-4 generated innovative and engaging tasks that both challenged and motivated the students. Therefore, ChatGPT-4 helped transform the tasks designed for digital collaboration into onsite workshops for the week-long visit. As mentioned in the goals of the VCL framework, a crucial aspect is fostering a culture of critical thinking (Balazs, 2005). To address this, ChatGPT-4 was consulted to develop tasks emphasizing the importance of critical thinking and the effective integration of constructive feedback. Based on this abstract goal, this led to *a Feedback Rotation Schedule* proposed by ChatGPT-4. Building on this proposal, the following task prompt was created for ChatGPT-4 as shown in Figure 1:

> Formulate the following task for a course:
> C1.1 Gather feedback from the other groups about the proposals. Use Social Media presence. Use Feedback Rotation between the groups (three days to make feedback for the other groups + two days to get feedback) Workload 1 page PDF
> C1.2 Refine and improve the Case/Concept based on feedback gained.
> C1.3 Make a branding + dissemination plan e.g. Logo, Acronym, Dissemination plan - Develop a Logo and Logo and acronym: Design a meaningful logo and choose an appropriate acronym for your project that is easy to remember and reflects the essence of your concept. Workload: 1 page Deliverable: PDF

Figure 1. Prompt for Formulating a Course Task

Which lead to the following answer from the ChatGPT4 for task C1 as shown in Figure 2:

> C1.1 Collecting feedback on the proposals from the other groups
>
> The aim of this task is to gather constructive feedback from other groups on your project proposals. Use your social media presence to maximize reach and interaction. Each group should create a short summary of their work to date (1 page PDF) and share it on appropriate platforms to gather feedback. You have a total of five days to complete this task - three days to create and share the summary and a further two days to collect and consolidate feedback.
>
> Feedback rotation: Each group gives feedback on the other group's suggestions, so that in the end each group has received feedback from every other group.

Figure 2. Answer from ChatGPT-4

This structured approach, facilitated by ChatGPT-4, ultimately led to the creation of the following task, which was distributed to students within the case study as shown in Figure 3:

> Task C1.2 Collecting feedback on the proposals from the other groups
>
> The aim of this task is to gather constructive feedback from other groups on your project proposals. Use the MS Teams channels. Each group should produce a short summary of their work to date and share this to gather feedback. You have a total of five days to complete this task - three days to create and share the summary and a further two days to collect and consolidate feedback.
>
> **Feedback rotation**: Each group provides feedback on the other group's suggestions, so that at the end, each group has received feedback from every other group.
>
> After collecting feedback, use these insights to refine and improve your case or concept. Revise your proposal based on the feedback and prepare a revised version that considers the suggestions and ideas for improvement you received.
>
> **Workload:** 1 page (with changes, if applicable)
>
> **Deliverable:** PDF

Figure 3. Detailed Task Instructions in the Case Study

As demonstrated, AI, particularly ChatGPT-4, significantly supported the task development process, starting from abstract goals to the initial drafts and finally to the completed tasks distributed in the case study.

4. Observation E-Tutors

As part of the action-based research approach and the comparative analysis of how a joint face-to-face week differs from one that starts directly with digital collaboration, the task for the E-Tutors was to supervise the students during the on-site workshops. Therefore, ChatGPT-4 was central in crafting consistent and well-structured tasks, ensuring each activity had a precise duration, topic, objective, and instructions. These tasks were designed to facilitate specific areas of e-tutor development, such as communication, teamwork, conflict resolution, and decision-making. By observing group dynamics, documenting task divisions, and providing feedback, e-tutors can significantly enhance collaborative processes. ChatGPT-4 tools aid in identifying which aspects of e-tutor qualification translate effectively into a physical environment, emphasizing the importance of active observation and feedback mechanisms. This structured approach ensured that e-tutors could effectively support groups and contribute constructively to discussions, ultimately fostering a productive learning environment. As demonstrated in Figure 4, incorporating ChatGPT-4 in task creation enhanced the structuring of e-tutor summer school activities, aligning them with pre-defined pedagogical goals.

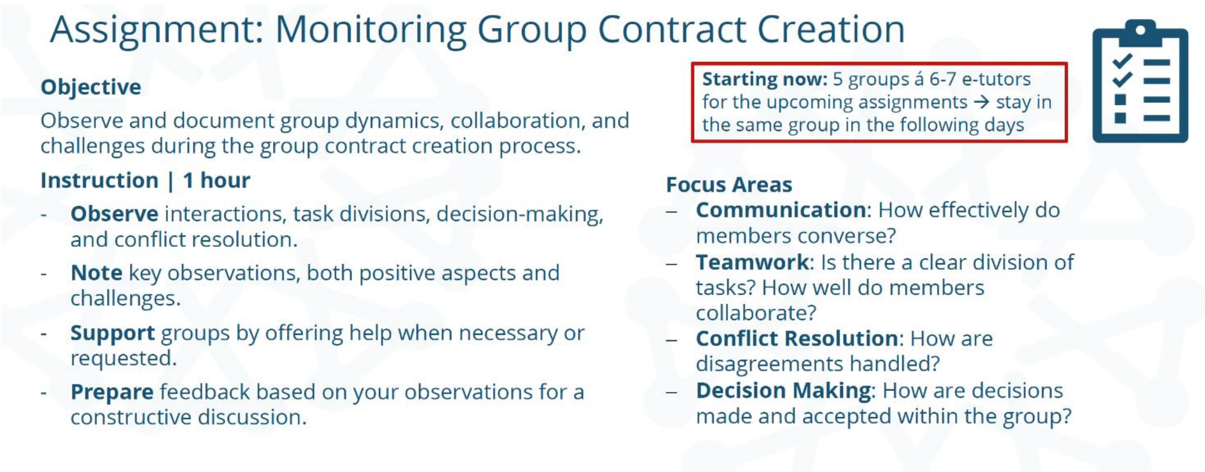


Figure 4. Exemplary Task for E-Tutors

5. Lessons Learned

As a result of the development process, the case study focuses on developing innovative solutions tailored to the needs of Tirana by enhancing cross-border cooperation and promoting comprehensive research and conceptual development of smart city strategies by students. The finalized case study has been integrated as a module into the curricula at both universities. This section provides an overview of the lessons learned from using AI, especially ChatGPT-4, in overcoming the challenges mentioned and adapting the components of previous VCL modules to the current case study in the context of smart cities.

- **Adaptation of Educational Content:** Using AI, particularly ChatGPT-4, played a significant role in adapting and enhancing educational content for the VCL module. It helped create structured and detailed tasks that aligned with the learning objectives.
- **Development of Tasks**: ChatGPT-4 supported the task creation process and led to practical and realistic learning tasks using the researched information on smart city strategies. These tasks directly engaged students with real-world challenges, promoting critical thinking, problem-solving skills, and interdisciplinary collaboration. Although the AI-generated tasks

could not be used directly in a copy-and-paste manner, they provided a clear and well-formulated foundation that greatly assisted task creation. ChatGPT-4 assisted in transforming tasks originally designed for digital collaboration into onsite workshop activities.

- **Feedback and Revision**: Throughout the process, ChatGPT-4 offered valuable feedback on drafts of the case study and suggested improvements. This iterative feedback process helped for continuous improvement and refinement of the educational content and task development.

6. Conclusion

AI tools, especially ChatGPT-4, were crucial in adapting and creating detailed and structured tasks. This facilitated the promotion of critical thinking and provided continuous feedback for improvement. Specifically, AI helped transform digital collaboration tasks into onsite workshops, overcoming traditional educational barriers. Our findings highlight that AI can support detailed task formulation, which makes abstract goals more tangible for students. For instance, ChatGPT-4 assisted in creating a Feedback Rotation Schedule, ensuring structured and meaningful peer evaluations. Moreover, AI tools enabled the adaptation of complex project tasks to fit different collaborative environments, both virtual and physical. For future initiatives, it is recommended to continue leveraging AI to support task development, feedback processes, and collaborative learning environments. Future research should explore the long-term impacts of AI-enhanced learning modules on student performance and engagement. Additionally, it is crucial to examine the practical applications of AI in various educational settings, particularly in diverse cultural and academic contexts. Such initiatives will ensure the scalability and adaptability of the VCL framework, fostering a robust and inclusive learning environment.

References

Altmann, M., Ukhova, N., & Hilse, L. (2023). Lessons Learned from an International and Interdisciplinary Virtual Mobility Module between Albania, Slovenia, and Germany. International Journal of Management, Knowledge and Learning, 12. https://doi.org/10.53615/2232-5697.12.S39-48

Altmann, M., Ukhova, N., Volkmann, N., & Schoop, E. (2024). Blending Physical and Virtual Mobility in Higher Education. Communities in New Media. Inclusive Digital: Forming Community in an Open Way Self-Determined Participation in the Digital Transformation: 26. Workshop GeNeMe‘23 Gemeinschaften in Neuen Medien, 329–334.https://fis.tu-dresden.de/portal/de/publications/blending-physical-and-virtual-mobility-in-higher-education(39e8bd73-0478-4e49-a101-72aef374bf40).html

Balazs, I. E. (2005). Konzeption von Virtual Collaborative Learning Projekten. https://nbn-resolving.org/urn:nbn:de:swb:14-1111134624957-21292

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour. International Journal of Educational Technology in Higher Education, 21(1). https://doi.org/10.1186/s41239-023-00436-z

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. International Journal of Educational Technology in Higher Education, 20(1). https://doi.org/10.1186/s41239-023-00392-8

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. European Journal of Education, 57(4), 542–570. https://doi.org/10.1111/ejed.12533

READING FATIGUE AND THE EFFECTIVENESS OF KNOWLEGDE SYSTEMS

Majid Ziaei Nafchi

Prague University of Economics and Business
Faculty of Informatics and statistics
majid.nafchi@vse.cz

DOI: 10.35011/IDIMT-2024-147

Keywords

Knowledge systems; Reading fatigue; knowledge management

Abstract

A very common issue that knowledge managers are facing in the organizations is the transfer of information and knowledge to the employees. Not necessarily because of the unavailability of the information but this concerns the information and the knowledge which is available and accessible. The purpose of the study is to identify the common issues and causes in this regard by conducting interviews with a selected number of employees. While there are many causes associated with this issue the preliminary results of the research show that one of the contributing factors of this distortion in the knowledge transfer is the online engagement fatigue it impacts the employee's motivation to read and subsequently their learning ability. This refers to the inability of the employees to be actively involved in 'Cognitive engagement', which of course could be a direct result of the reading fatigue. The next steps in the research are to get a hand on approach in the research and utilize quantitative methods before and after the implementation of the recommended solutions to be able to measure if the changes were successful.

1. Introduction

Recently, one of the most common issues that knowledge managers are facing in the organizations is the transfer of information and knowledge to the employees. This could be due to various reasons, for example not having a well- established information or knowledge systems, lack of knowledge managers, etc.

However, in larger and more established organizations, this is usually not the case, but the issue seems to be more around in- sufficient attention/ability of the employees to read and learn something either due to lack of time or lack of interest, or the from the side of knowledge managers as educators and the way they prepare the infor- mation/knowledge material. Of course, the longer the reading material is the more resistant to read it the employees are; and if the language quality of the text is not standard or unified, it would be more challenging for the employees to focus and keep them engaged in the learning process.

The aim of this Paper is to investigate and identify the key issues when it comes to reading and comprehending the information and whether the knowledge managers can use any tools or software to check and improve the quality of the text and if not to test application of such tools, furthermore, to recommend further steps for the research to implement some software to produce executive summaries to reduce the length of the learning material to check how that effects the engagement of employees in the reading and learning.

2. Literature review

According to Malony et. Al. (2023) the concept of online engagement fatigue and how it impacts learning was perceived by educators and students in a different manner. To start with the educators had conflicting views about the reality of online engagement fatigue but students believed that online engagement fatigue was real, and, in most cases, they had experienced it.

This is basically the same phenomenon is the organisations these days, as the knowledge mangers or knowledge creators act as the educators and the employees are the learners or the students. Due to efficiency most of the time employees spend on learning, is in terms of reading (materials, handbooks, and articles prepared by their colleagues, and the knowledge managers) rather than spending hours in face-to-face trainings where they can be actively engaged and focused, which of course brings some challenges.

Benedetto et. al. (2013) states reading behaviour has been studied by psychologists for a long time, with different studies focusing on low-level processing of words, for example visibility or legibility and other studies focusing more on the comprehension of information.

Perhaps the best emerging definition of online engagement fatigue was defined as: "*A reduction in online students' enthusiasm and motivation for engaging in course activities as a result of overexposure to online coursework and associated interactions"* (Malony et. Al., 2023).

According to Li and Hu (2024) face to face learning is more beneficial than online learning as it allows learners to be corrected by teachers and allows for real time feedback and discussions.

Incorporating computer technologies has many advantages in learning, but there are drawbacks as well. Information technology use and online learning have drawn criticism for potentially disadvantageous effects on particular student populations (Chen et al., 2010). In line with Jenkins' "participation gap" theory, some academics contend that socioeconomic position and institutional resources have a major impact on how computers and the Internet are used by students as well as how they make use of them (Jenkins, (2006).

Furthermore, some researchers claimed that for students with specific learning styles, the absence of in-person interactions in online learning could decrease the effectiveness of the training. No communication device can take the place of in-person interactions and fortuitous learning opportunities like impromptu conversations or overheard comments made during a break in class (Chen et al., 2010).

On the other hand, lack of 'engagement' of the employees due to various reasons in the learning process is counterproductive as well. Considering learning necessitates involvement for competence development and knowledge acquisition, engagement is a crucial component of the learning process (Shi, 2010). Cognitive engagement means the thinking proces, which is dealing largely with the use of cognitive strategies for example coding, analysis, interpretation, etc. Reaching excessive levels of cognitive engagement necessitates higher-order thinking such as creativity or connecting a new knowledge to previous knowledge (Chapman, 2019).

Learners' cognitive engagement has been associated with enhanced achievement, persistence, and retention of information, while disengagement has an overwhelming effect on learner acquiring results and on their level of cognitive development. Additionally, some confusion as to whether the terms 'engagement' and 'motivation' can and should be used interchangeably exists, which should be clarified. However, the predominant understanding within the literature is that motivation is indeed an antecedent to engagement; Motivation is the intent and unobservable force that energises behaviour, while engagement is energy and effort in action (Bond et al., 2020).

According to Laberge et. Al. (2011) acute fatigue has been shown to be positively correlated with higher degrees of psychological distress, worse health perception, more sleep debt, and higher exposure to physical work-related stressors. Additionally, it appears that having several occupations, being in greater psychological distress, having a worse sense of one's health, and being exposed to physical labour elements were all linked to higher levels of chronic fatigue.

3. Objectives and methodology

This paper is addressing the preliminary results of the study; therefore, the main objective of this paper is to *identify* the main issues and causes that are preventing employees from using the knowledge systems to learn about the processes in the organisation. Additionally, to come up with the plan for further develop this study and to propose the next steps in the research and possible scenarios or solutions to be tested.

Based on the information above we formulated the main research questions as:

1. What are the challenges/difficulties in the online learning, which lead to learning fatigue?
2. How can we improve the online learning experience and reduce fatigue?

To answer the goal of this study interviews have been conducted with a selected number of employees in a marketing research company and a literature review was conducted using scientific papers accessed from well-known scientific data bases.

The sample targeted for the interviews were the junior employees which are more involved in day to day in operations, which are more in need of those guidelines and processes that are on the knowledge base. Senior employees were excluded from the sample for a few reasons, to start with they are more familiar with the processes, and more conditioned to access the knowledge base for the information. On the other hand, the senior employees are in less need of the knowledge base as they have more of a leading role and are less involved in the operations.

To analyse the finding a more qualitative approached has been used, however in the continuation of the research quantitative methods are going to be used.

4. Results and discussion

Unlike smaller organisations, the lack of knowledge managers or an inadequate information or knowledge systems is typically not the problem in larger and more established organizations. Instead, the problem seems to be more with the employees' lack of attention or interest in reading and learning something, or in some cases with knowledge managers' role as educators and how they present the knowledge and information.

We cannot ignore the benefits of the traditional face to face learning at the interaction between the teacher and the learner could accommodate a better learning experience and better comprehension on the information and allows real tie feedback on possible misunderstandings and questions, however, in the larger organizations it is more efficient to relay largely on online learning.

Initial results of the research show that the despite having a well-structured knowledge base, employees do not use it as often as it was intended to. Of course, this does not mean that employees do not used the knowledge base at all. Most of them are using the knowledge base but not regularly. The reason behind this varies from one employee to other. The key reasons that were identified are:

- They assume the information that they are seeking is unavailable on the knowledge base.
- The information is available, but they are unable to find it through the search option in the knowledge base.
- The information they are seeking is unavailable on the knowledge base (this is mostly regarding the new processes and ways of working)
- They find it easier and faster to ask another colleague about the information they are seeking.
- They find the information on the knowledge base, but it is a lengthy text, and they don’t feel like reading or get tired of the reading halfway through.
- They do not fully understand the processes explained on the knowledge base by just reading the text.

Unavailability of the information or the knowledge is of course expected as the organisation is constantly improving the processes and they are sometimes new models or concepts being used for the projects in the organisation, but this is mainly temporary as anytime there is a change or anything new the knowledge and information is created around that and is added to the knowledge base. Efforts are made to have the knowledge and information ready before going live with some of the new processes, but unfortunately some of the knowledge is gathered while running project and it can be made available after the projects finish for future projects. In this case not much can be done.

One of the issues is assuming on the availability of the information/knowledge, which in this case the employees should confirm if information is available or unavailable by asking the relevant people in the organisation, yet again there are regular knowledge cascades in the organisation to keep employees up to date on what is new and what was changed. Of course, employees tend to forget such news, so confirming the availability is advised.

On the other hand, sometimes the employees are unable to find the relevant information that they are seeking through the search bar, this issue is more connected to the technical background of the knowledge base and how keywords are used or how the articles are being tagged. Though this is a very important aspect, and it could be improved, it is not an interest in this paper.

The more interesting issues found relate to the fact that the employees find it easier to ask others than to read the article, because it is faster or easier for them. This on its own can be examined from different points of view:

- The fact the employees are busy with day-to-day tasks, and they want to utilize their time in other ways than reading the articles (lack of capacity, motivation to read)

- They lose their interest in reading the article due to the length of the articles (reading fatigue). The longer the articles and the more complicated they are there is a higher chance that the employees would face this issue. And of course, when reading fatigue comes in, the motivation and the cognitive engagement of the employees diminishes.
- Lack of information assimilation for different reasons, mostly decreased energy, and reading fatigue, and due to language barrier as this is an international organisation and the knowledge is prepared by employees with different nationalities with different levels of proficiency in the English language. This again has a negative effect on both the motivation and the cognitive engagement of the employees.
- On some pages there is a piece of information about how long it takes to read the given article, something as simple as '5 minutes read', which brings the next questions: firstly, why isn't this information on all pages? This is indeed something that could be easily improved. And secondly, if the given article indeed takes that much time to read for everyone, not just to read but to understand the content as well.

These issues could be addressed using tools and applications which are powered by Artificial intelligence, the next step in the research would be to investigate whether the knowledge managers, which are responsible for capturing and managing the knowledge in the organisation, use any tools or software to check and improve the quality of the text and if not; and if not to test implementation of such tools.

Additionally, to reduce the length of the learning material possibly by implementing some software to produce executive summaries to check who that effects the engagement of employees in the reading and learning. Similarly, using AI to calculate a better and more realistic estimate on the time needed for reading the articles would be beneficial. However, the use of AI might come at some drawbacks too, which needs the attention of the knowledge mangers. For example, creating executive summaries might condition the employees to pay less attention to the full articles, and consequently miss some information and therefore, don't assimilate the full picture.

5. Conclusions

In conclusion, knowledge managers frequently struggle in organizations with the information and knowledge transfer to staff members. This could be caused by several things, such as a lack of knowledge managers or an inadequate information or knowledge systems. However, the online reading fatigue is real, and it is indeed a contributing factor in reducing the effectivity of the knowledge systems in organisations.

Some employees don't feel like reading the text which is available because they lack capacity or they feel tired and don't have energy for reading, while others drop out mid-way through reading because of the length of the articles. This refers to the inability of the employees to be actively involved in 'Cognitive engagement', which of course could be a direct result of the reading fatigue. On the other hand, understanding of the articles is a bit challenging when it comes to employees with lower proficiency in English language.

The next steps that are recommended to take in this research are to investigate whether the knowledge managers are using any tools or software currently to check and improve the quality of the text that are being published in the knowledge base in form of articles, and if not, to test application of such tools to check if that helps with the issues.

Moreover, to implement some sort of a software to produce executive summaries to reduce the length of the learning material to check and measure how those effects the engagement of employees in the reading and learning, there are many AI powered applications for this purpose which can be beneficial.

To be able to measure the impacts of implementing such applications and for the purpose of the future research, a questionnaire should be prepared, and the results of the survey would be analysed using quantitative statistical methods together with a qualitative process of interviewing employees and knowledge managers to have the research questions addressed.

Acknowledgement

Paper was processed with support from institutional-support fund for long-term conceptual development of science and research at the Faculty of Informatics and Statistics of the Prague University of Economics and Business (IP400040) and the project IG 409033.

References

Bond, M., Buntins, K., Bedenlier, S., Zawacki-Richter, O., & Kerres, M. (2020). Mapping research in student engagement and educational technology in higher education: a systematic evidence map. International Journal of Educational Technology in Higher Education, 17(1). https://doi.org/10.1186/s41239-019-0176-8

Chapman, Elaine (2019) "Alternative Approaches to Assessing Student Engagement Rates," Practical Assessment, Research, and Evaluation: Vol. 8, Article 13. DOI: https://doi.org/10.7275/3e6e-8353

Chen, P.-S. D., Lambert, A. D., & Guidry, K. R. (2010). Engaging online learners: The impact of Web-Based Learning Technology on college student engagement. Computers & Education, 54(4). https://doi.org/10.1016/j.compedu.2009.11.008

Jenkins, H. (2006). Convergence Culture: Where Old and New Media Collide. NYU Press. http://www.jstor.org/stable/j.ctt9qffwr

Laberge, L., Ledoux, E., Auclair, J., Thuilier, C., Gaudreault, M., Gaudreault, M., Veillette, S., & Perron, M. (2011). Risk factors for work-related fatigue in students with School-Year employment. Journal of Adolescent Health, 48(3), 289–294. https://doi.org/10.1016/j.jadohealth.2010.07.003

Li, X., Hu, W. (2024). Peer versus teacher corrections through electronic learning communities and face-to-face classroom interactions and EFL learners' passion for learning, speaking fluency, and accuracy. Heliyon, 10 (4). e25849, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2024.e25849.

Maloney, S., Axelsen, M., Stone, C., Galligan, L., Redmond,P., Brown, A., Turner,J., Lawrence, J. (2023). Defining and exploring online engagement fatigue in a university context, Computers and Education Open, Volume 4, 2023, 100139, ISSN 2666-5573, https://doi.org/10.1016/j.caeo.2023.100139.

Benedetto S., Drai-Zerbib V., Pedrotti M., Tissier G., Baccino T. (2013). E-readers and visual fatigue. PLoS ONE. 2013;8:e83676. doi: 10.1371/journal.pone.0083676

Shi, S. (2010) Teacher moderating and student engagement in synchronous computerconferences.Journal of Online Learning and Teaching6(2).

REVIEWERS' FALLACIES: THE OPTIMAL NUMBER OF REVIEWERS PER MANUSCRIPT AND THE SURPRISING EFFECT OF "LAZY" REVIEWERS

Lubomír Štěpánek, Filip Habarta, Ivana Malá, Luboš Marek

Prague University of Economics and Business
Faculty of Informatics and Statistics
lubomir.stepanek@vse.cz, filip.habarta@vse.cz, malai@vse.cz, marek@vse.cz

DOI: 10.35011/IDIMT-2024-153

Keywords

optimal number of reviewers per manuscript, diligent reviewers, non-diligent reviewers, lazy reviewer phenomenon, virtual coworking system, voting scheme, review procedure optimization, maximizing the probability of decision making, reviewers' nonresponse, reviewers' fallacies

Abstract

Determining the optimal number of reviewers per manuscript is crucial in the publication review process to balance efficiency and scientific credibility. As the academic publication network grows, the challenge of non-responsive or non-diligent, "lazy" reviewers in volunteer-based systems becomes more pronounced. This study explores two main issues. Firstly, we investigate an ideal number of reviewers necessary to ensure a reliable review process, and, secondly, the impact of non-responsive reviewers on this process. We propose using an odd number of reviewers to enable clear decision-making, where each reviewer votes to accept or reject a manuscript. Our analysis reveals that specific numbers of reviewers are linked to lower probabilities of review completion. Surprisingly, our results suggest that including a proportion of non-diligent reviewers might actually increase the likelihood of gathering adequate reviews. Based on these findings, we offer recommendations for the optimal number of reviewers that enhance the probability of achieving a decisive review outcome. Additionally, our study indicates that non-diligent reviewers do not necessarily obstruct the review process and may, in fact, facilitate reaching a final decision.

1. Introduction

The objectivity and scientific merits of a review process are ensured – besides others – by a feasible number of reviewers referring to a given manuscript or other publication outcome. On the one hand, the number of reviewers should not be too low, which could indicate a lack of sufficient scientific credits and evidence behind the limited number of reviews. On the other hand, in theory, when there are too many reviewers asked for a review of a given publication, the procedure consisting of communication with the reviewers and managing they provide their reviews in time places high demands on an editor; even more, the review process and the final decision about the reviewed publication may be delayed due to challenging synchronization of a high number of reviewers. Thus, the number of reviewers is usually assumed to be neither low nor high (Snell, 2015), commonly in the range of two to five referees per publication.

However, assuming dichotomous conclusions of reviews, i.e., a review either suggests to (rather) accept the reviewed manuscript or to (instead) reject it, some specific numbers of reviewers asked to work out a review on the manuscript could work poorly, considering a probability the final decision about the manuscript, could be easily made by an editor. While some empirical results are discussed in Byrne (2016), there is usually a lack of literature on the topic.

As a rising phenomenon, the situation when a reviewer initially confirms they will work out an asked review in time, but they are either substantially late in their review submission or do not submit anything even after multiple reminders, when editors try to communicate with the reviewer, but they become nonresponsive. Such a phenomenon is called "a lazy reviewer" in literature (Gavras, 2002) and seems to occur more and more frequently. In the end, that could not be very surprising, considering the reviewers' time invested into a review report working out and submission has not a real award. So, reviewers usually agree with the review report prepared on a voluntary basis. Furthermore, the issue with a lazy reviewer, also having an analogy on an author's side as "a lazy author", could exist for a long time, and editors may face it from the very beginning of the current publication and reviewer process as we know it in its digital form, dated to nineties (Plekhanov et al., 2023). What helped to uncover the phenomenon of a lazy reviewer as a nontrivial problem is the growing complexity and size of the authors', publications', and reviewers' networks understood as a graph (Heldens et al., 2022). Naturally, if there always has been a proportion of reviewers that are not as diligent as the remaining ones and could be in this context called "lazy", the increasing size of the reviewers' network (Waltman et al., 2013), enabled by the science fields' grown and better interdisciplinary cooperation (Sjögårde and Didegah, 2022), necessarily increased the total number of those "lazy" reviewers", which attracted attention on this kind of issue.

In this work, we conduct research on both questions, i.e., the optimal number of reviewers per manuscript, maximizing the probability a final decision about the manuscript could be made, and consequences of a "lazy reviewer" presence among others asked for a review working out, using a more formal approach based on probability. Based on the assumption of dichotomic reviews, i.e., these either suggest a reviewed publication acceptance or rejection, one could natively suppose that an odd number of collected reviews made a final decision about the reviewed publication easier since it avoids a tie situation when two groups of review reports of the same size suggest the opposite action – the first would accept the publication while the second would reject it. By assuming there is a "population" level of probability that each reviewer does submit their review report, we model the overall probability of collecting an odd number of review reports, enabling an editor to make a final decision about the reviewed publication. A formula for the overall probability provides an easy way to simulate various combinations of numbers of asked reviewers and values of probability that a reviewer either does submit their review report or does not, and, particularly, enables searching for a number of reviewers that maximizes the overall probability of possible decision.

We investigate the effect of a lazy reviewer presence among the asked reviewers. Indeed, an editor could not initially know if there is "a lazy" reviewer among the ones requested for a review report, hoping there are none of these among them. As an assumption, one could expect that the presence of a lazy reviewer tends to harm the procedure's smoothness – it could end up in a necessity for another reviewer to call and cause a substantial delay in the review process. As we demonstrate, for some expected values of (small) probabilities, a lazy reviewer does submit their review report, the overall probability of possible decision-making about a reviewed publication could be surprisingly even higher when there is a lazy reviewer present, compared to a situation there are only diligent reviewers requested for review conduction.

The managerial implications regarding the review procedure are straightforward. From an editor's point of view, maximizing the probability of a final decision about a reviewed publication is easily possible and is welcome. Therefore, the knowledge of an optimal number of reviewers that should be initially asked for the reviews is a logical way to improve and optimize the review procedure's effectiveness. Also, the fact a lazy reviewer is (unwillingly) requested for review could sometimes unexpectedly even help (!) to increase the probability of making a final decision about the referred manuscript.

2. Research methodology

In this section, we methodically refine the overall probability of reaching a final decision regarding a reviewed publication. This approach allows us to determine the optimal number of reviewers to initially request for review report submission, thereby maximizing the overall probability of collecting an odd number of review reports. Additionally, we formally formulate the overall probability concerning the final decision in the context of a lazy reviewer among the requested reviewers and analyze when this presence could potentially increase the probability.

2.1. An overall probability that a decision about a reviewed publication is possible

Firstly, let's assume that a decision about a reviewed publication is feasible if we collect more than half of the requested review reports, each of which either suggests accepting or rejecting the publication. This ensures the credibility of the review process and provides ample evidence from reviewers. Additionally, since review reports are expected to offer dichotomous recommendations about the publication, we require an odd number of review reports to enable a final decision in case of a tie.

Assume we initially ask n reviewers to work out a review about a manuscript and submit their report, where $n \in \mathbb{N}$. Each of the reviewers has a probability $p \in \langle 0, 1 \rangle$ that they submit back their review report. Then a probability that we receive r reviews comes from binomial distribution and is equal to

$$\pi_{n,p,r} = P(\text{we receive } r \text{ reviews}) = \binom{n}{r} p^r (1-p)^{n-r} \tag{1}$$

where $r \in \{0, 1, \dots, n\}$. The events of collecting u and v review reports, where $u \neq v$ and $u, v \in \{0, 1, \dots, n\}$, are mutually disjoint; thus, their probabilities could be summed up if needed (Feller, 1968). So, using formula (1) and the previous piece of knowledge, the probability that a final decision about the reviewed publication is possible, i.e., a probability of receiving an odd number of review reports that is greater than a half of all initially requested review reports is equal to

$$\begin{aligned} P(\text{a decision is possible}) &= \sum_{r=0}^{n} \mathcal{I}\left(r \textit{ is odd and } r > \frac{n}{2}\right) \cdot \pi_{n,p,r} \overset{(1)}{\cong} \\ &\overset{(1)}{\cong} \sum_{r=0}^{n} \mathcal{I}\left(r \textit{ is odd and } r > \frac{n}{2}\right) \binom{n}{r} p^r (1-p)^{n-r}, \end{aligned} \tag{2}$$

where $\mathcal{I}(\alpha)$ is an identifier function, i.e.,

$$\mathcal{I}(\alpha) = \begin{cases} 1, & \text{if } \alpha \text{ is true,} \\ 0, & \text{if } \alpha \text{ is false.} \end{cases} \tag{3}$$

A feasible task is to search for such number of reviewers n, that would maximize probability $P(\text{a decision is possible})$ from formula (2).

2.2. A probability that a decision about a reviewed publication is possible when a lazy reviewer is among the requested

In this scenario, we assume that there could be one or more "lazy" reviewer among $n \in \mathbb{N}$ requested ones for review. While the probability that a diligent reviewer submits back their review report is $p_d \in \langle 0, 1 \rangle$, the same probability for a lazy reviewer is $p_l \in \langle 0, 1 \rangle$, where $p_l < p_d$. Once again, we assume that a final decision about a reviewed publication is made when more than half of the initially requested reviews are collected, and their number is odd, thus avoiding issues with possible suggestion ties. Our objective is to analyze how the probability of a potential final decision about the publication is impacted by the presence of "lazy" reviewers.

Adopting the notation from the previous subsection, we can assume there are n_l lazy reviewers and n_d diligent ones among n reviewers in total, $n_l + n_d = n$. Then, we can calculate the probability of receiving $l + d$ reviews in total; in particular, l reviews from lazy and d ones from diligent reviewers, where $l + d \in \{0, 1, \dots, n\}$. Obviously, using formula (1), the probability l lazy reviewers submit their reports is $\binom{n_l}{l} p_l^l (1 - p_l)^{n_l - l}$, while the probability d out of n_d diligent reviewers submit their reports is $\binom{n_d}{d} p_d^d (1 - p_d)^{n_d - d}$; these events are independent and their probabilities could be multiplied as

$$\begin{aligned} \pi_{n_l + n_d, p_l, p_d, l+d} &= P(\text{we receive } l + d \text{ reviews}) = \\ &= \binom{n_l}{l} p_l^l (1 - p_l)^{n_l - l} \binom{n_d}{d} p_d^d (1 - p_d)^{n_d - d}. \end{aligned} \tag{4}$$

Finally, adopting the logic of formula (2), we could investigate the probability that a final decision about a reviewed publication is possible – thus, we search for a probability that r reviews from n reviewers is collected, regardless of how many of them come from diligent or lazy reviewers, where $r \in \{0, 1, \dots, n\}$, considering a presence of lazy reviewers, which count is denoted as n_l, where $n_l \in \{0, 1, \dots, n\}$. Straightforwardly, to get a probability that the decision about publication is possible, i.e., we collect an odd number of reviews that is greater than a half of initially asked reviewers, we exhaustively search through all possible counts of reviews, $r \in \{0, 1, \dots, n\}$, check whether r is odd and $r > \frac{n}{2}$, and if so, we sum all terms for all possible numbers of lazy reviewers, so $n_l \in \{0, 1, \dots, n\}$ and all possible combinations of numbers of $l + d = r$ reviews from diligent and lazy reviewers. Using formula (4), we obtain

$$\begin{aligned} &P(\text{a decision is possible}) \\ &\qquad = \sum_{r=0}^{n} \mathcal{I}\left(r \text{ is odd and } r > \frac{n}{2}\right) \cdot \pi_{n, p_l, p_d, r} \overset{(4)}{\cong} \\ &\overset{(4)}{\cong} \sum_{r=0}^{n} \left\{ \mathcal{I}\left(r \text{ is odd and } r > \frac{n}{2}\right) \cdot \right. \\ &\quad \left. \cdot \left\{ \sum_{n_l=0}^{n} \left\{ \sum_{l=0}^{n_l} \sum_{d=r-l}^{n-n_l} \binom{n_l}{l} p_l^l (1 - p_l)^{n_l - l} \binom{n - n_l}{d} p_d^d (1 - p_d)^{n - n_l - d} \cdot \right\} \right\} \right\}, \end{aligned} \tag{5}$$

where $\mathcal{I}(\alpha)$ is an identifier function as in formula (3). Since we cannot know how many reviewers out of n in total are "lazy", a feasible task is to investigate whether probability $P(\text{a decision is possible})$ from formula (5) could be higher even if there are lazy referees among the reviewers.

3. Results

We performed multiple simulations to address both questions of the paper. To get insights into what number of reviewers should be initially asked for working out a review report for a given manuscript, we simulated all combinations of values of probability, $p \in \{0.6, 0.7, 0.8, 0.9\}$, that the reviewers submit back their review reports, and the numbers of asked reviewers, $n \in \{2, 3, 4, 5, 6, 7\}$. For each combination of values n and p, we calculated probability $P(\text{a decision is possible})$ from formula (2) that a final decision about the reviewed manuscript is possible to be made, and searched for such a value of n^* that maximizes the probability, so

$$\begin{aligned} n^* &= \underset{n \in \{2,3,4,5,6,7\}}{\arg\max} \{P(\text{a decision is possible})\} = \\ &= \underset{n \in \{2,3,4,5,6,7\}}{\arg\max} \left\{ \sum_{r=0}^{n} \mathcal{I}\left(r \textit{ is odd and } r > \frac{n}{2} \right) \cdot \pi_{n,p,r} \right\} = \\ &= \underset{n \in \{2,3,4,5,6,7\}}{\arg\max} \left\{ \sum_{r=0}^{n} \mathcal{I}\left(r \textit{ is odd and } r > \frac{n}{2} \right) \binom{n}{r} p^r (1-p)^{n-r} \right\}, \end{aligned} \tag{6}$$

using the same notation as before.

The results we got are in Table 1 and Figure 1. As we can see, for all values of probability that an asked reviewer submits back their review report we considered, numbers of initially asked reviewers, that maximize the probability depicting a final decision about the refereed publication can be made, $P(\text{a decision is possible})$, are equal either to three reviewers, particularly for large values of probability p, or five reviewers, when probability p is a bit lower. On the other hand, the probability that final decision about the publication is made, $P(\text{a decision is possible})$, is for some counts of reviewers, n, that could be, in theory, initially asked, relatively low, e.g., when $n \in \{2, 4, 6\}$; thus, these even numbers of reviewers should not be considered when an editor plans on how many reviewers they should request for review conduction.

Table 1. Table of probabilities $P(\text{a decision is possible})$ that a final decision about the reviewed manuscript can be made, considering n reviewers are initially asked to submit their review reports, each with a probability p of submission.

probability of review submission, p	**number of reviewers, n**					
	2	**3**	**4**	**5**	**6**	**7**
0.6	0.000	0.216	0.346	0.423	0.187	0.289
0.7	0.000	0.343	0.412	0.477	0.303	0.400
0.8	0.000	0.512	0.410	0.532	0.393	0.485
0.9	0.000	0.729	0.292	0.663	0.354	0.602

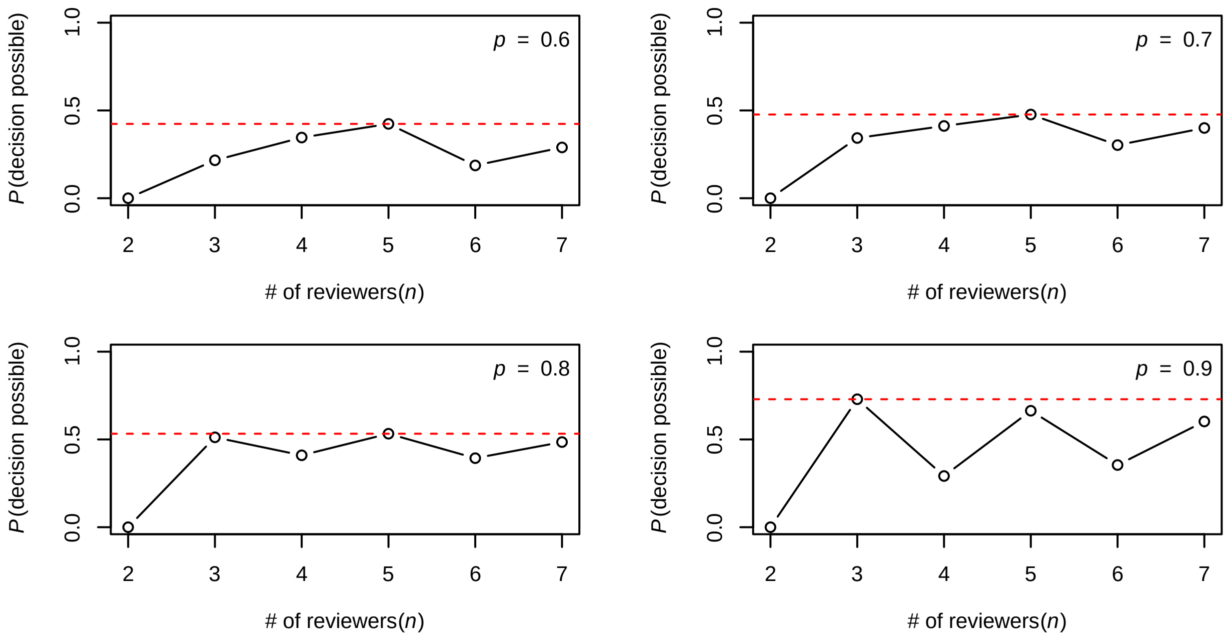


Figure 1. Probabilities $P(\text{a decision is possible})$ that a final decision about the reviewed manuscript can be made, considering n reviewers are initially asked to work out their review reports, each with probability p of submission back their review report.

We also performed multiple simulations taking into account the presence of one or more "lazy" reviewers among total number of reviewers, n. We calculated the probability $P(\text{a decision is possible})$ from formula (5), that a final decision about a refereed publication can be done, for all combinations of numbers of reviewers, $n \in \{2, 3, 4, 5, 6, 7\}$, numbers of lazy reviewers $n_l \in \{0, 1, \dots, n\}$, the probability that a diligent reviewer submits back their review report $p_d \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$, and the probability that a lazy reviewer submits back their review report $p_l \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. Once probability $P(\text{a decision is possible})$ is calculated for each count of lazy reviewers, conditional on a number of reviewers n, we search for a case when a non-zero number of lazy reviewers among all the reviewers surprisingly increases the probability on decision enabling. Thus, in total, we performed $6 \times 5 \times 5 = 150$ simulations considering a lazy reviewer; about almost 40 of them contain the "paradox" phenomenon that the presence of a lazy reviewer tends to increase the probability that a final decision about a refereed publication can be made.

Refer to Figure 2 for selected illustrations of such cases, wherein the presence of one lazy reviewer unexpectedly increases the likelihood of a smooth final decision about the reviewed publication, particularly when the total number of reviewers is even. For instance, with five initially requested reviewers, having two lazy reviewers among them could maximize the likelihood of decision-making. In the scenario of six fully diligent referees submitting their review reports, a tie situation may arise, such as when the first three reports suggest acceptance and the remaining three suggest rejection. However, surprisingly, the presence of one lazy reviewer can contribute to a smoother decision-making process by abstaining from submitting their review, thereby allowing the editor to avoid a tie situation.

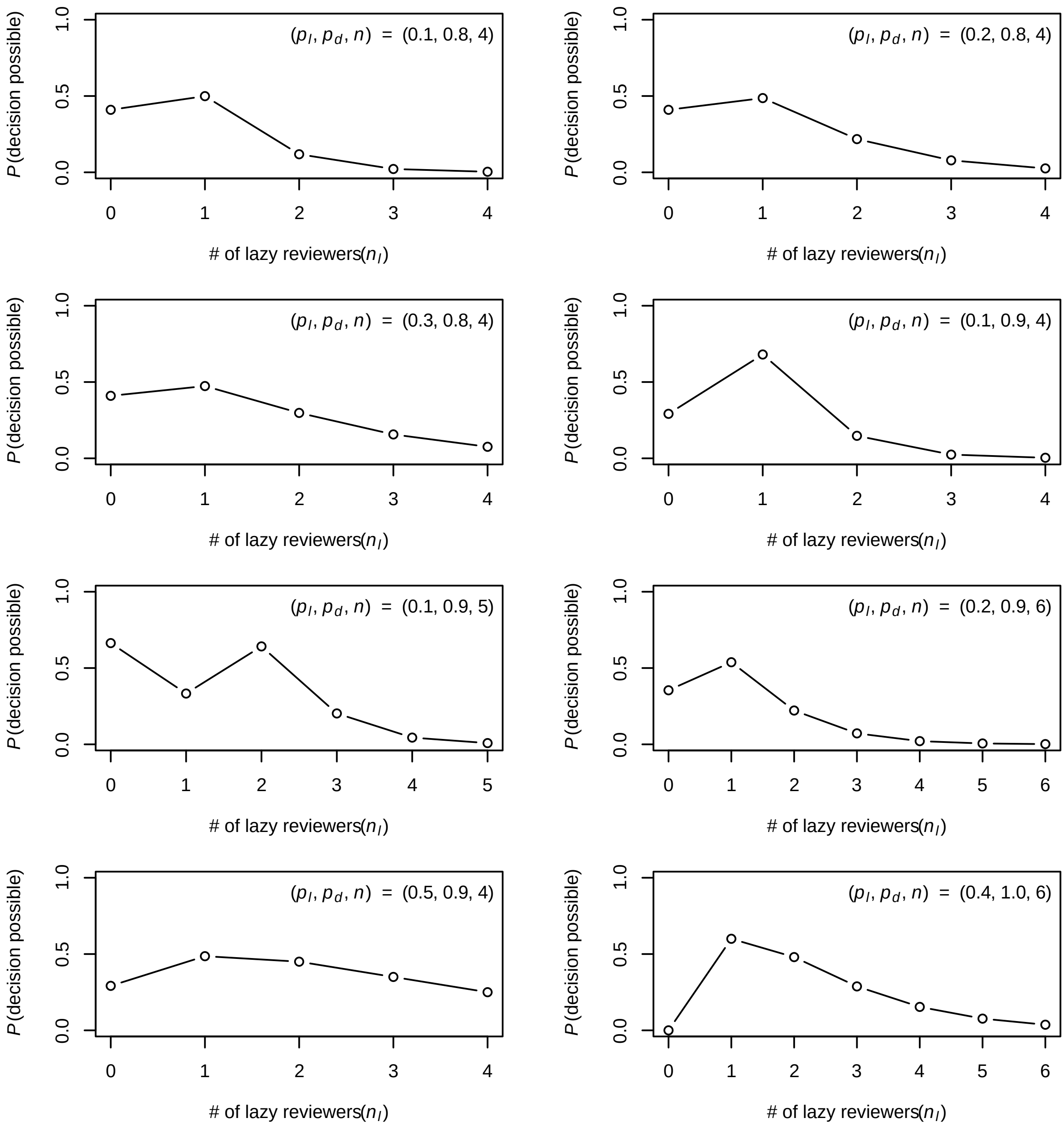


Figure 2. Probabilities $P(\text{a decision is possible})$ of a possible final decision about the reviewed manuscript, considering n reviewers are initially asked to submit their review reports, including n_l lazy reviewers. Diligent $(n - n_l)$ reviewers have a submission probability of p_d, while lazy (n_l) reviewers have a probability of p_l.

4. Discussion and conclusions

In this study, we analyzed the optimal number of reviewers per publication who should be initially asked to conduct a review, suggesting either acceptance or rejection. To ensure the scientific credibility of the review process and to avoid tie situations, we preferred an odd number of collected review reports exceeding half of the requested reviewers, facilitating a smooth final decision on the publication. We employed probability analysis to determine which numbers of reviewers maximize the probability of making a final decision about the publication. Through multiple simulations, we found that if the probability of each reviewer submitting their review report is high (around 90 %), then initially asking three reviewers maximizes the probability of a smooth decision-making process. However, if the probability of review report submission is slightly lower (around 60–80 %), then initially asking five reviewers is preferable to maximize the probability of making a final decision about the publication. Additionally, we demonstrated the disadvantage of initially asking an even number of reviewers. Also, considering the potential presence of one or more non-diligent, "lazy" reviewers among the review group complicates matters for an editor. We define a "lazy" reviewer as one with a low probability of submitting their review report (less than or equal to 50 %). Initially, it may seem that the chance of easily collecting an optimal number of review reports is reduced due to the presence of lazy reviewers. However, we illustrated through multiple examples that, surprisingly, the probability of making a decision about a publication could unexpectedly increase due to the presence of lazy reviewers. In a paradoxical twist, the presence of lazy reviewers within the review group enhances the likelihood of a smooth final decision-making process about the publication. The explanation for this paradox lies mostly in the scenario where we have an even number of diligent reviewers, each likely to submit their review report. This scenario does not favor the smoothness of the final decision-making process, as we would prefer to collect an odd number of reviews rather than an even one to avoid tie situations. Therefore, if one member of an even-numbered group of reviewers is lazy and does not submit their review, we receive an odd number of reviews, facilitating the avoidance of ties in the final decision-making process about the publication.

In conclusion, while the presence of lazy reviewers may initially seem problematic, it is not a serious issue. In fact, for limited numbers of lazy reviewers among those requested for review (i.e., one or two), it can surprisingly aid in increasing the likelihood of a smooth final decision-making process about the reviewed publication.

Acknowledgement

This paper is supported by the grant IG410023 with no. F4/50/2023, which has been provided by the Internal Grant Agency of the Prague University of Economics and Business.

References

Byrne, J. A. (2016). Improving the peer review of narrative literature reviews. In Research Integrity and Peer Review (Vol. 1, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s41073-016-0019-2

Feller, W. (1968). An introduction to probability theory and its applications, volume 1 (3rd ed.). Nashville, TN: John Wiley & Sons.

Gavras, H. (2002). Inappropriate attribution: the "lazy author syndrome." In American Journal of Hypertension (Vol. 15, Issue 9, p. 831). Oxford University Press (OUP). https://doi.org/10.1016/s0895-7061(02)02989-8

Heldens, S., Sclocco, A., Dreuning, H., van Werkhoven, B., Hijma, P., Maassen, J., & van Nieuwpoort, R. V. (2022). litstudy: A Python package for literature reviews. In SoftwareX (Vol. 20, p. 101207). Elsevier BV. https://doi.org/10.1016/j.softx.2022.101207

Plekhanov, D., Franke, H., & Netland, T. H. (2023). Digital transformation: A review and research agenda. In European Management Journal (Vol. 41, Issue 6, pp. 821–844). Elsevier BV. https://doi.org/10.1016/j.emj.2022.09.007

Sjögårde, P., & Didegah, F. (2022). The association between topic growth and citation impact of research publications. In Scientometrics (Vol. 127, Issue 4, pp. 1903–1921). Springer Science and Business Media LLC. https://doi.org/10.1007/s11192-022-04293-x

Snell, R. R. (2015). Menage a Quoi? Optimal Number of Peer Reviewers. In C. M. Gray (Ed.), PLOS ONE (Vol. 10, Issue 4, p. e0120838). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pone.0120838

Waltman, L., van Eck, N. J., & Wouters, P. (2013). Counting publications and citations: Is more always better? In Journal of Informetrics (Vol. 7, Issue 3, pp. 635–641). Elsevier BV. https://doi.org/10.1016/j.joi.2013.04.001

AUTONOMOUS VEHICLES

SMART AND AUTONOMOUS SYSTEMS MEETING SUSTAINABILITY CHALLENGES

Erwin Schoitsch

Center for Digital Safety & Security
AIT Austrian Institute of Technology
erwin.schoitsch@ait.ac.at

DOI: 10.35011/IDIMT-2024-163

Keywords

smart systems, autonomous systems, resilience, sustainability, circular economy, machine ethics, Society 5.0, artificial intelligence, trustworthiness, Green Deal, UN Sustainable Development Goals

Abstract

This paper, as a keynote and introduction IDIMT 2024 Session D, Autonomous systems and smart, sustainable, and resilient environments, provides an overview over some areas of "Smart Environments", where ICT and AI technologies provide essential support to meet challenges of sustainability, resilience and European sovereignty. A particular European concern is to guarantee this in a human-centered, ethically aligned, trustworthy manner. "Nobody left behind" and "for the benefit and empowerment of people" (with respect to automation and AI) are principles expressed already in the first declaration of the current President of the European Commission, Ursula von der Leyen, in "A Union that strives for more – My agenda for Europe". The "Green Deal" and "European Sovereignty" in key economic and research areas, are further far-leading goals and European modifications of the paradigm of "Society 5.0". Smart technologies are the drivers of economic and societal disruptive changes and a chance to shape our future in a beneficial way (but no chance without risk). This is part of the European way (contribution) to resolve the current crisis (climate change, aging society, shortage of critical resources resulting in critical dependencies, unbalanced world economy and powers).

1. Introduction

Japan has as first large leading economy and industrialized country identified long term challenges of their society: facing a set of long-term challenges just now: Climate change (in an undesirable direction), economic stagnation and high inflation rates, and an ageing society, Technology change causes disruptive changes for labor forces, customers and economy.

They started a large initiative initiated by science and government in a joint effort – "Society 5.0" ((Schoitsch, E., 2019)). The main targets were to fight economic stagnation in a society which for too long a time was governed by the same party and people, and at the same time fighting the consequences of an ageing society. Certain aspects of high automation, e.g., acting robots in healthcare, tagging schoolchildren by RFID devices to monitor their movements, and the like, which are accepted in the Japanese society to an extent unacceptable in Europe under our ideas of privacy

(“GDPR”) and self-determination, facilitated the first move. In Europe we consider more the strong impact of ethical and societal concerns particularly if looking at the use of AI-technologies and are not only targeting the technology challenges from the economic and ethical side part also under the “Green Deal” aspect as a major concern for the future.

For Europe, adaptations to the concept have been undertaken, including also the 17 UN Sustainable Development Goals. “Sustainability” is a rather holistic concept, it requires to take into account all possible long-term effects of any move to meet the challenges – even well-meant ideas (like “bio-fuel” produced from garbage from the fields) could have negative impact – at the moment when mass production required more garbage than available from agricultural waste it lead to misuse of food (maize) to produce fuel!), in a world where still many areas are not able to fight hunger! Additionally, the political uncertainties that have endangered the far-reaching supply chains in critical resources (energy, materials) and products or components (semiconductors**)** led to more consideration of issues of resilience and sovereignty (which also includes “brain-ware”, software and digital systems/networks).

2. Artificial Intelligence – Standards and Ethically Aligned Design

With the rise of Artificial Intelligence and Decision-making or Machine Learning Systems (off-line or during operation) many critical questions arose: Who can guarantee a certain performance, safe or secure system behavior, what about predictability of outcomes (a key precondition to assure safety of a system)? Are the results to expect understandable, or the system activities explainable? Ethical aspects include the impact of automated decisions on human life – organizational, technical safety and security, human rights and human sovereignty? The technical safety is discussed in the chapter of “Smart Mobility” in context of an application or ISO/IEC TR 5469 – Functional safety and AI systems”. Figure 2 provides an overview over various guidelines, standards and recommendations on ethically aligned design of AI and on cybersecurity issues.

Figure 1. Guidelines, Standards and Recommendations for Ethically Aligned Design and AI Source: (Author)

Looking at the oversight issue from different stakeholders’ view, Figure 3 gives some impression. It covers several aspects, governance (organisations), controllability by humans, safety, bias, management. These standards are under standardization or just published. They could become basis for references of the AI Act if published in the Official journal of the EC and having undergone some review by the HAS (Harmonized Standards) consultants.

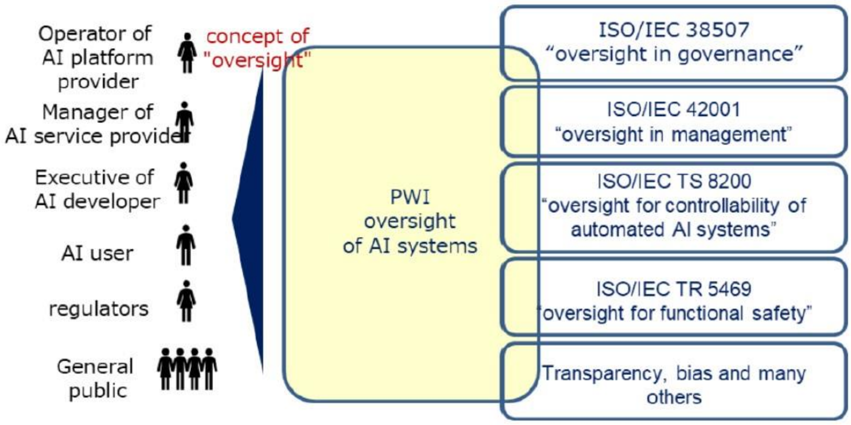


Figure 2. (Human) Oversight of AI Systems Source: (ISO/IEC JTC1 SC42 WG03, TR 18966)

3. Smart Manufacturing: Project "AIMS5.0" ("Artificial Intelligence in Manufacturing leading to Sustainability and Industry5.0")

The EU-funded AIMS5.0 project (see Acknowledgement) is at the forefront of advancing Artificial Intelligence (AI) in Manufacturing leading to Sustainability and Industry 5.0. This project addresses the vulnerabilities of existing supply chains by promoting shorter, more localized production processes and integrating AI technologies to bolster sustainability and competitiveness within European industries. The transition from Industry 4.0 to Industry 5.0 by AIMS5.0 includes leveraging the Internet of Things (IoT), semantic web ontologies, machine learning, AI and human-machine cooperation to enhance working conditions and eco-friendly manufacturing practices, i.e., addresses specifically the human-centered aspects in context of the 5th step of "Mass Personalization" in a separate dedicated work package.

The following figure shows how Industry5.0 is derived from Industry4.0 ("RAMI Model") adding personalization and human-centered approaches, like work conditions and public stakeholders' acceptance of highly automated and autonomous systems and trustworthiness.

Use Cases (demonstrators) are from several fields, including automotive, manufacturing, semiconductors development and production, consumer electronics, robotics, aviation, some cross-domain applications and in-door food production (see details under "Agriculture 5.0"). A specific application is the use case with exoskeletons in the automotive manufacturing field, a typical case of "human empowerment". The overarching goal of most use cases is to fulfill requirements towards the "Green Deal" and "Digital sovereignty" as well as the supply chain resilience challenge in the semiconductor field.

Human-centered options are the support of human workers by:

- Intelligent data management and measurement to create **more efficient and more sustainable working conditions** for energy optimized production.
- More **accurate process control parameters** derived from a **digital twin** yield to more robust and more sustainable production environment.
- Comprehensive **human-machine simulation** based on the **digital twin** increases **safety** by predicting potentially hazardous.

- Improve the **efficiency of maintenance** work and to prevent high loss due to machine and product damage.

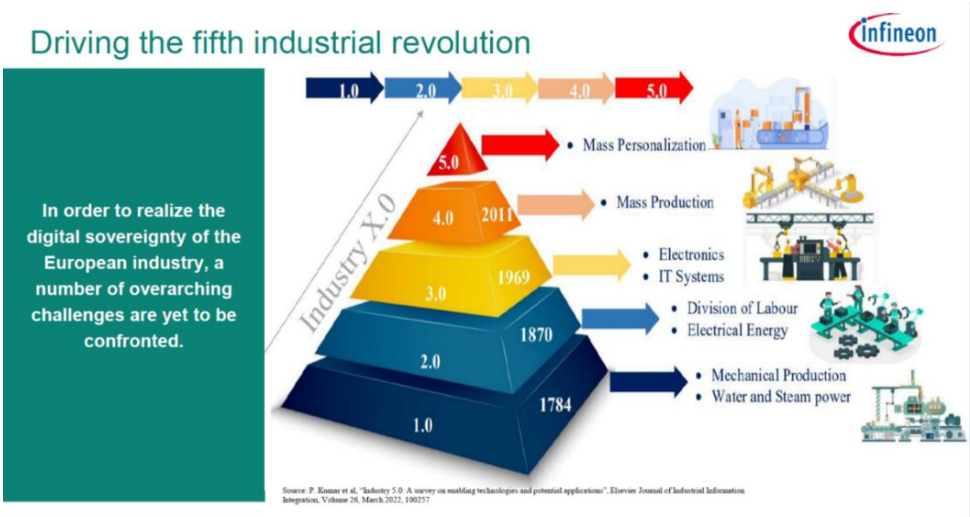


Figure 3. Driving the 5th Industrial Revolution – AIMS5.0 project – human controllability Source: (Kumar, P., 2022 & AIMS5.0)

Here are standards helpful, particularly on AI trustworthiness and smart manufacturing, e.g., guidance on how to use new technologies in smart manufacturing, like IEC TR 63283-4 CD (under development): Industrial-process measurement, control and automation – Smart Manufacturing – Part 4: New technologies. This document is a “Smart manufacturing impact analysis” wrt. new technologies. The identified important new technologies are:

- AI (Artificial Intelligence),
- Edge computing,
- Cloud technology,
- Digital twin,
- New communication protocols, 5G, TSN,
- Big data and data analytics,
- IoT and IioT,
- Privacy technology, etc.

Each chapter has a subchapter on “Technology description”, “Use case analysis” and “Standardization needs”.

4. Smart Mobility: Project “AI4CSM” (“Automotive Intelligence for Shared Connected Mobility”)

One example for a European research project towards energy-efficient, electric and automated mobility is AI4CSM, funded under the ECSEL JU scheme within Horizon 2020the national authorities of participating partners (41 partners from 10 countries, grant agreement n° 101007326). The objectives are:

- Develop robust and reliable mobile platforms,

- Develop scalable and embedded intelligence for edge and edge/cloud operation,
- Design silicon for deterministic low latency and build AI-accelerators for decision and learning,
- Solve complexity by trustable AI in functional integrated systems,
- Design functional integrated ECS systems,
- Build ECAS (Electric-Connected-Automated-Shared) vehicles for the Green Deal and future connected, shared mobility.

Some important demonstrators with AI systems in critical roles are (selected examples):

- Robo-Taxi
- Virtual City Routing
- Lessons from critical scenarios for ADS controllers

There are about 20 demonstrators, also dealing with Green-Deal objectives and Ethical aspects (Schoitsch 2022).

Particularly the “Robo-Taxi” is an ideal example showing all attributes for use of AI standards in safety-relevant implementations, from AI-based control elements and safety analysis, impact on “Green Deal” conformance (reduction of number of vehicles in cities, better usage of infrastructure, reduction of sealed surfaces, optimal support of multi-modal transport, optimization of minimal energy consumption and timing through city-routing, etc.), “inclusion” of people not fit for driving, and reduction of fatalities, as well as ethical and societal concerns because of autonomous decision making impacting people (humans) based on uncertain or unpredictable behavior. The analysis allowed a classification of the various AI-systems and components and will facilitate later testing and final design for an upcoming product-stage. It was also used as part of the tool chain (an AI Training Center, to automatically generate concrete test cases out of an abstract test case to test and validated an ADAS function in an intelligent and structured way).

One research question for the work in the standardization part of the project was how to support the partners with respect to standards to apply. Almost all demonstrators have AI system developed or integrated in different roles of different criticality with respect to safety. For research purposes, partners active in AI and functional safety standardization, ISO/IEC TR 5469, developed together by ISO/IEC JTC1 SC42, AI, WG 3 (Trustworthiness) and IEC SC65a MT 61508-3 (Functional safety of E/E/PE systems, the basic functional safety standard for SW for safety-related systems in general) informed the partners on the status of the work on this standard. The draft standard served as support document to analyze the AI systems in context of their implementation in the demonstrators.

NOTE: Begin of 2024, ISO/IEC TR 5469 was published. It is a report, not containing requirements, and is therefore rather a guideline and study document. Work is now continued towards a TS (Technical specification) ISO/IEC TS 22440, in three parts, in a Joint Working Group JWG 4 of ISO/IEC JTC1/SC42 and IEC TC65 SC65A: Functional safety and AI systems:

- TS 22440-1 - AI — Functional safety and AI systems — Part 1: Requirements
- TS 22440-2 - AI— Functional safety and AI systems — Part 2: Guidance
- TS 22440-3 - AI— Functional safety and AI systems — Part 3: Examples of Application

Information on the principles of ISO/IEC DTR 5469

The approach taken, in short, was to classify AI technology classes (how far safety properties of functional safety standards can be fulfilled, I=full, II=mitigation possible, III=not possible) and usage classes (A1 highest, D lowest) (Schoitsch, E., 2023). The following figure tries to map these and provide recommendations:

<table>
<tr><th>AI Technology Class => AI application and usage level</th><th>AI technology Class I</th><th>AI technology Class II</th><th>AI technology Class III</th></tr>
<tr><td>Usage Level A1 (1)</td><td rowspan="5">Application of risk reduction concepts of existing functional safety International Standards possible</td><td>Appropriate set of requirements (5)</td><td rowspan="5">Not recommended</td></tr>
<tr><td>Usage Level A2 (1)</td><td>Appropriate set of requirements (5)</td></tr>
<tr><td>Usage Level B1 (1)</td><td>Appropriate set of requirements (5)</td></tr>
<tr><td>Usage Level B2 (1)</td><td>Appropriate set of requirements (5)</td></tr>
<tr><td>Usage Level C (1,3)</td><td>Appropriate set of requirements (5)</td></tr>
<tr><td>Usage Level D (2)</td><td colspan="3">No specific functional safety requirements for AI technology, but application of risk reduction concepts of existing functional safety International Standards (4)</td></tr>
<tr><td colspan="4">1 Static (offline) (during development) teaching or learning only
2 Dynamic (online) teaching or learning possible
3 AI techniques clearly providing additional risk reduction and whose failure is not critical to the level of acceptable risk.
4 Additionally, other safety aspects (not being addressed with functional safety methods) can possibly be adversely affected by AI usage.
5 The appropriate set of requirements for each usage level can be established in consideration of Clauses 8, 9, 10 and 11. Examples are provided in Annex B.</td></tr>
</table>

Figure 4. Recommendations for usage of AI Technology Classes at certain Usage Levels Source: ISO/IEC DTR 5469)

5. Agriculture 5.0, Smart Farming and In-door Farming

Resilient and Sustainable Cities are considered, in the long term, to have to be able to support itself to a large extent from the neighboring areas and from inside production, particularly of energy and food. A first step is to enhance food production by smart precision farming in the neighborhood, and in another step to produce food in a high-tech environment in-door by optimized management of plan-growing in a pure environment without needing pest management or weed killers and using vertical farming.

In the AIMS5.0 project, one Use Case focuses on AI-supported Industrial IoT for Indoor Food Production, aiming to optimize the cultivation process using advanced technologies. This includes the deployment of a cloud-based edge computing platform where workloads are integrated to manage multiple services, linking them for enhanced operation. A multi-protocol data gateway is implemented to facilitate real-time data acquisition from sensors and control of actuators, streamlining interactions with the physical environment.

An NLP (Natural Language Processing) pipeline is developed to enhance data utilization, consisting of several stages: data gathering from literature on Indoor Vertical Farming, text preprocessing, and a training phase employing transformer and knowledge graph approaches. These steps are designed to extract actionable insights and streamline information processing.

AI-based mechanisms process collected vision data and automatically assess plant health and size but also ensure reliable and efficient wireless communication security. This involves setting up an AI-powered cross-domain communication gateway and constructing a dedicated testbed with switched beam antenna and Bluetooth transmitters for initial testing and data collection.

Security and compliance are also crucial components of UC11, with efforts directed towards developing a threat model for Industrial AI applications. This model assesses systems for security and compliance with critical Cybersecurity Regulations like the CRA, ensuring that the innovative solutions are both secure and aligned with regulatory standards.

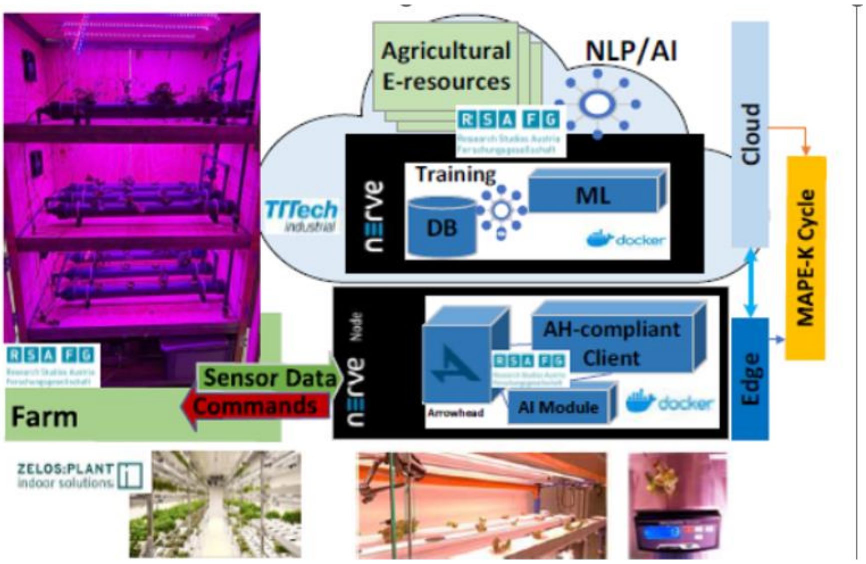


Figure 5. AIMS5.0 Use Case 11, efficient In-door Farming as high-tech Application Source: (AIMS5.0 project)

Through these initiatives, UC11 of the AIMS5.0 project aims to reduce manufacturing costs, enhance product quality, shorten time-to-market, and increase user acceptance of technological innovations, thus bolstering production efficiency and sustainability in indoor food production within European industries.

In the EDEN project (Acknowledgements at the end), Aquaponic Farming is the goal and use case.

6. Conclusions

The paper shows some important areas, where smart technologies could lead towards a better, human centred society. The "Green Deal" programme and the human implications of these technologies, empowering people through education and skills, and on protecting against the risks of these technologies, are targeting resilience and sustainability of society and economy. However, we should be aware that many of the achievements could be effective too late or used against us as well, or lead to wrong decisions because of badly trained or biased AI systems.

Acknowledgements

Part of the work received funding from the EC from the EU Horizon Europe Programme, the ECSEL and ChipsJU Joint Undertaking and the partners' national funding authorities (in Austria FFG (Austrian Research Promotion Agency) on behalf of BMK, The Federal Ministry of Climate Action (BMK)) - (Grant agreements n° 101112089 (AIMS5.0), AfarCloud (783221) and AI4CSM (101007326-2)). The ADEX (Automated Driving Examiner, Programme ICT of the Future, F880811) and the EDEN (Efficient, De-centralized Food Production with IoT Technology, Programme Circular Economy, FO999897774) project were funded by FFG.

References

Bissadu, K.D., Sonko, S., & Hossain, G. (2024). Society 5.0 enabled agriculture: Drivers, enabling technologies, architectures, opportunities, and challenges. Information Processing in Agriculture, 2024, ISSN 2214-3173, (https://www.sciencedirect.com/science/article/pii/S2214317324000234), https://doi.org/10.1016/j.inpa.2024.04.003.

ERCIM News number 127, October 2021. Special Theme "Smart and Circular Cities", Guest editors Schoitsch, E. (AIT, Vienna) and Georgios Mylonas (ISI, Athens), p. 4 – 35, published by ERCIM EEIG, Sophia Antipolis, France, ISSN 0926-4981, https://ercim-news.ercim.eu/en127

ERCIM News number 135, October 2023. Special Theme "Climate resilient Society", Guest editors: Athina Lykos (AIT Austrian Institute of Technology) and Sobah Abbas Peterson (NTNU), published by ERCIM EEIG, Sophia Antipolis, France, ISSN 0926-4981, https://ercim-news.ercim.eu/images/stories/EN135/EN135-web.pdf.

European SDSN Network, (2020). "The 2020 Europe Sustainable Development Report (ESDR 2020)", issued by the SDSN (Sustainable Development Solutions Network) and Bertelsmann Foundation, https://www.unsdsn.org/sdg-index-and-monitoring (with further links of other regions).

European Commission, (2019). Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. "The European Green Deal", 2019-12-11.

European Commission, (2019). High-Level Expert Group, "Ethics Guidelines for Trustworthy AI" (Final report April 2019, HLEG AI), Brussels; https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

European Green Cities, (2019). http://greencities.eu/about; Action Plan https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1596443911913&uri=CELEX:52019DC0640#document2.

Federal Ministry of Transport and Digital Infrastructure, (2017). Ethics Commission on "Automated and Connected Driving – Report June 2017", Germany; https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission-automated-and-connected-driving.pdf?__blob=publicationFile (Summary available in English).

Hinkelmann, K. & Gerber, A. (Eds), 2022. Proceedings of the Society 5.0 Conference 2022 - Integrating Digital World and Real World to Resolve Challenges in Business and Society, EPiC Series in Computing, Volume 84, https://easychair.org/publications/volume/Society_5.0-2022.

Japanese Government, Cabinet Office (2018). Achieving Society 5.0, https://www8.cao.go.jp/cstp/english/society5_0/index.html.

Japanese Government, (2017). Realizing Society 5.0. https://www.japan.go.jp/abenomics/_userdata/abenomics/pdf/society_5.0.pdf.

Kumar, P. et al., 2022, "Industry 5.0: A survey on enabling technologies and potential applications", Elsevier Journal of Industrial Information Integration, Volume 26, March 2022, 100257.

Fukuyama, M. (2018). Society 5.0: Aiming for a New Human-Centered Society, Japan Spotlight, July/August 2018.

Schoitsch, E. (2023). "Autonomous Systems and Smart Environments for a Sustainable Life", IDIMT 2023, New Challenges for ICT and Management, Proceedings, Trauner Verlag, Linz, Austria, Schriftenreihe Informatik 52, (ISBN ISBN 978-3-99151-176-2), p. 133-141.

Schoitsch, E. (2022). "Smart Technology and Circular Economy for a Greener World and Resilient Society", IDIMT 2022, Digitalization of Society, Business and Management in a Pandemic, Trauner Verlag, Linz, Austria, Schriftenreihe Informatik 51, (ISBN 978-3-99113-758-0), p. 357-366.

Schoitsch, E. (2019), "Beyond Smart Systems – Creating a Society of the Future (5.0): Resolving Disruptive Changes and Social Challenges", IDIMT 2019, Innovation and Transformation in a Digital World, Proceedings, Trauner Verlag, Linz, Austria, Schriftenreihe Informatik 48, p. 387-400 (ISBN 978-3-99062-590-3).

United Nations, Transforming our World - The 2030 Agenda for Sustainable Development (2015), https://sustainabledevelopment.un.org/post2015/transformingourworld.

Von der Leyen, U. (2019). "A Union that strives for more – My agenda for Europe". https://www.europarl.europa.eu/resources/library/media/20190716RES57231/20190716RES57231.pdf.

MAPPING SCIENTIFIC TRENDS IN THE SHARING ECONOMY AND SUSTAINABLE DEVELOPMENT: NATIONAL STRATEGIES AND INNOVATION DYNAMICS

Vít Kluger

Department of Trade and Finance, Faculty of Economics and Management,
Czech University of Life Sciences Prague
kluger@rektorat.czu.cz

Kristína Korená

Department of Tourism and Marketing, Faculty of Corporate Strategy,
Institute of Technology and Business in České Budějovice
korena@vste.cz

Petra Pártlová

Department of Trade and Finance, Faculty of Economics and Management,
Czech University of Life Sciences Prague
partlova@pef.czu.cz

DOI: 10.35011/IDIMT-2024-171

Keywords

sharing economy; sustainable development; bibliometric analysis; VOSviewer; national strategies; CSR (Corporate Social Responsibility)

Abstract

Although the sharing economy is often presented as the key to achieving sustainability, there is a gap in the scientific literature regarding its integration into national strategies and research priorities. This paper aims to fill the gap through comprehensive mapping of publications imparting both national and international trends in research on the sharing economy and sustainable development. The authors use quantitative methods of bibliometric analysis and data visualization, specifically the VOSviewer tool, to identify and evaluate major thematic clusters. Also, the authors analyse the links between keywords and their association with relevant geographic areas in order to reveal geographic and thematic concentrations of scientific research. The results show that the sharing economy is emerging as a multidisciplinary field with significant overlap in the areas of corporate strategy, innovation, technology, environmental management and sustainability policy. Research is dominated by topics such as digitalization, eco-innovation, energy efficiency and renewable energy integration, *with important links to national policies and international climate agreements. Furthermore, this paper reveals a dynamic pattern of research activities that reflect the growing integration of the sharing economy into national sustainable strategies, and provides valuable insights into how different countries and regions can approach addressing common sustainable development challenges.*

1. Literature review: Bridging carsharing, sustainable urban mobility, and the sharing economy's innovation dynamics

Carsharing is becoming increasingly vital in urban mobility, integrating advanced technologies to tackle traffic congestion and emissions. It offers an efficient alternative to personal car ownership, emphasizing resource efficiency. Neumann (2021) notes its potential to enhance urban sustainability by reducing emissions and increasing vehicle usage efficiency. Turon (2022) highlights the impact of environmentally friendly vehicles on carsharing's ecological footprint. Understanding user behaviour is crucial for carsharing expansion. Pawełoszek (2022) and Vejchodská et al. (2023) stress the importance of understanding user needs and expectations. Drobiazgiewicz & Pokorská (2023) show how local policies influence carsharing adoption, essential for planning and implementation. Kubera and Slusarczyk (2023) discuss carsharing's role in mobility management and reducing urban transport issues.

The automotive industry is transforming with electric (EVs) and autonomous vehicles (AVs), marking a shift in vehicle sharing. Kubik et al. (2023), Dvořáček et al. (2022), and Brescia et al. (2023) demonstrate EVs' role in reducing CO2 emissions and AVs' potential in urban settings. Kimbrell (2021) and Brescia et al. (2023) argue for a holistic view of EVs, combining technological, social, and environmental aspects. Neumann (2021) highlights carsharing's role in easing urban congestion and emissions, contributing to urban sustainability. Turon (2022) explores how selecting eco-friendly vehicles for carsharing fleets reduces the ecological footprint of urban transport. Incorporating electric and hybrid vehicles into carsharing programs reflects efforts to lower CO2 emissions and improve urban mobility sustainability. Technological advancements, especially EVs and AVs, increase carsharing services' ecological and operational efficiencies. Kubik et al. (2023) and Piotrowski et al. (2022) highlight these technologies' impact on promoting sustainable urban transportation. The success and growth of carsharing services depend on understanding user behaviour and preferences. Research by Pawełoszek (2022) and Vejchodská et al. (2023) highlights the need to grasp user demands, emphasizing factors like convenience, affordability, environmental awareness, and vehicle accessibility. Vejchodská et al. (2023) stress the importance of reliability and trust in carsharing platforms, noting that vehicle availability and a seamless experience boost user satisfaction and loyalty.

Carsharing must align with user preferences while ensuring reliability and ease of use. Its integration into urban transportation involves policy, technology, and public acceptance, which vary by region. National and local policies shape the carsharing landscape. Drobiazgiewicz and Pokorská (2023) emphasize Poland's supportive policies in reducing urban congestion and emissions. Carsharing's role in urban mobility strategies is expanding, supported by Mobility as a Service (MaaS). Studies by Chmiel, Pawlowska, & Szmelter-Jarosz (2023) and Matowicki et al. (2022) highlight its potential to integrate various transportation modes into a unified system, enhancing accessibility and user convenience, aligning with sustainability goals. Local policy environments, as discussed by Kubera and Slusarczyk (2023), affect carsharing's implementation and efficacy. The success of these services depends on regulations, infrastructure, and urban planning, necessitating adaptable policies tailored to city-specific mobility needs.

Technological innovation drives carsharing’s evolution, enhancing sustainability and efficiency through electric and autonomous vehicles (Kubik et al., 2023; Dvořáček et al., 2022). Data analytics and machine learning optimize fleet management and user experience, showcasing carsharing's dynamic nature (Ke et al., 2022). Successful carsharing growth hinges on balancing policy support, technological advancements, and local urban mobility needs, as shown by Drobiazgiewicz and Pokorská (2023) in Poland.

In summary, carsharing combines technological innovation with sustainability, offering an adaptable urban mobility model. This paper synthesizes various studies to advocate for policies and innovations supporting carsharing as a pillar of sustainable urban living. The analysis leads to research questions (RQ1, RQ2, RQ3) that explore interdisciplinary approaches, trends, and linkages in sustainable development:

RQ1: What impact do advance technologies and innovations have on achieving sustainable mobility and green infrastructure in cities?

The first research question explores the role of AI, big data, and IoT in creating efficient, sustainable urban transport solutions that enhance energy efficiency and reduce emissions.

RQ2: What are the key success factors for integrating sustainable business strategies and environmental policies in companies and public administrations?

The second research question examines the connection between corporate strategies and sustainable development, focusing on Corporate social responsibility (CSR), eco-innovation, and the transition to a green economy for environmental and social responsibility.

RQ3: How can interdisciplinary approaches be used to address global challenges such as climate change and sustainable development through local and regional innovation?

The third research question highlights the importance of cross-disciplinary collaboration in translating global Sustainable Development Goals (SDGs) into actions that tackle specific environmental, social, and economic challenges at local and regional levels.

These questions identify key areas for further research in sustainable development: technological innovation, business and environmental strategies, and the need for an interdisciplinary approach.

2. Methodology

The authors' research was significantly enriched by digital resources and analytical tools, notably the Web of Science, an esteemed online database of academic papers and articles. This platform not only simplified access to a vast array of scholarly works but also provided powerful data analysis and visualization tools. VOSviewer software, in particular, was instrumental in generating detailed maps that illustrated the relationships between scientific publications, keywords, authors, and other academic elements. These maps highlighted the interconnectedness of research fields, showing the strength and scope of academic ties through direct and associative links. By clustering information, VOSviewer offered a comprehensive overview of the academic landscape, evaluating entities based on their importance and influence, measured by indicators like citation frequency and publication year. The use of these advanced bibliometric and scientometric tools underscored the value of such databases as both information reservoirs and essential instruments for in-depth analysis. This approach enhanced the study's credibility and depth, as noted by Van Eck & Waltman (2022) and Donthu et al. (2022).

3. Results

3.1. The sharing economy as a path to sustainable development: Strategy, Innovation and Environmental Management

Sustainable development and environmental issues raise a wide range of topics that are not only up to date, but also crucial for the future direction of society. The research presented here has identified and categorised these thematic areas into specific clusters, each with a unique focus and characteristics. Such divisions allow to gain deeper understanding of different aspects of sustainability and environmental challenges, see Figure 1.

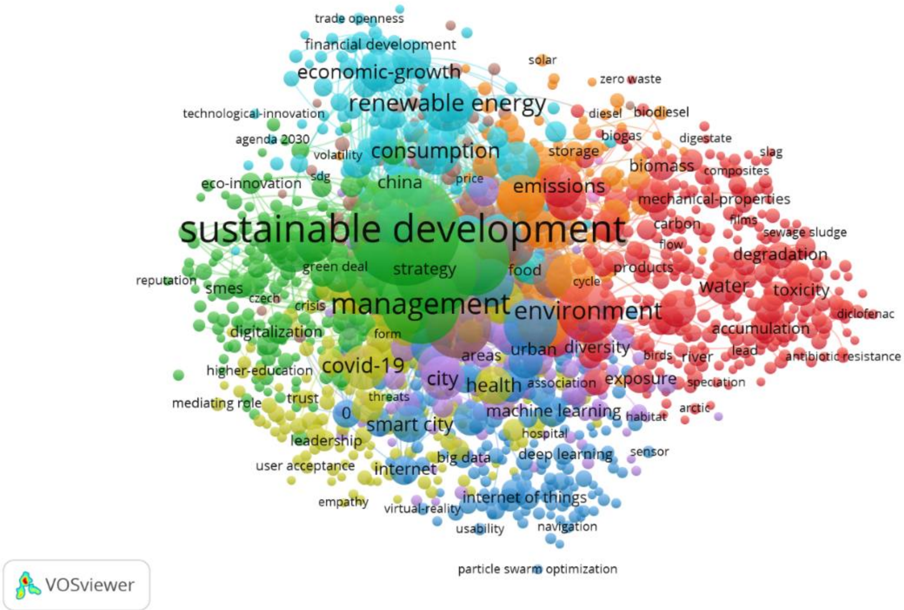


Figure 1. Mapping the future: Integrating sustainability into global development Source: (Authors)

The first cluster, "Business strategies and sustainable development," emphasizes integrating sustainability into business models to achieve environmental goals and long-term competitiveness. Digitalization, eco-innovation, and green deal initiatives reflect the growing incorporation of social and environmental responsibility in corporate values. In the second cluster, "Environmental impact and pollution management," pollution management through technologies like activated carbon, biodegradation, and bioaccumulation is addressed to reduce industrial and agricultural environmental impacts. Advancing to the third cluster, "Advanced technologies and smart infrastructure," the focus shifts to using technologies such as AI, big data, and IoT for developing smart cities and infrastructure, enhancing urban efficiency, and reducing environmental impact.

Meanwhile, the fourth cluster, "Behavioural insights and societal change," examines how social and behavioural factors influence sustainable practices, which is essential for promoting societal norms and behaviours for a sustainable future. The fifth cluster, "Urban development and green spaces," highlights integrating green spaces in urban planning to create healthier, more sustainable environments that support biodiversity and improve residents" quality of life. The sixth cluster, "Energy and climate policy," and the seventh cluster, "Renewable energy technologies and efficiency," collectively concentrate on energy policy, CO2 emissions, renewable energy sources, and energy efficiency to combat global warming and promote a sustainable energy future.

Next, the eighth cluster, "Sustainable transport and mobility," provides insights into sustainable transport systems that reduce environmental impact while promoting efficient and accessible mobility solutions. Finally, the ninth cluster, "Resource management and optimization," underscores the importance of resource efficiency and minimizing waste and emissions, which are crucial for achieving greater sustainability in industry and daily life.

3.2. Integrated strategies for sustainable development

In scientific and technical research, fascinating interconnections between disciplines reveal intense activity and keyword interdependence. These patterns highlight trends shaping current scientific discourse and show how fields collaborate for sustainable development. First, the authors note the relationship between "Environmental impact and pollution management" and "Energy and climate policy," emphasizing the interdependence between industrial activities" effects on water and air quality and policies to mitigate these impacts. Renewable energy and CO2 reduction are crucial in minimizing environmental harm. Another key link exists between "Business strategies and sustainable development" and "Urban development and green spaces." This connection underscores the role of companies in sustainable urban development, integrating green spaces and sustainable practices into business models. It exemplifies how economic growth and environmental goals, such as green job creation, can align for a sustainable future. A significant link between "Sustainable transportation and mobility" and "Advanced technologies and smart infrastructure" shows how AI and IoT can enhance transport efficiency, reduce energy consumption, and improve urban living. The relationship between "Energy and climate policy" and "Renewable energy technologies and efficiency" underscores the need to connect climate policy with renewable technologies to achieve climate goals. Innovations in renewable energy are key to reducing greenhouse gas emissions and supporting a sustainable energy system, vital for international climate agreements and national policies. Lastly, the link between "Behavioural insights and societal change" and the "Sustainable development goals (SDGs)" highlights how understanding human behaviour and decision-making aids in achieving global SDGs. These connections emphasize the importance of education, health, and social change for broader SDGs like equal opportunities for all.

Overall, these research linkages form a network of supportive themes essential for achieving sustainable development economically, socially, and environmentally. This holistic approach shows how disciplines can collaborate to address current and future societal challenges.

3.3. National priority axes in sustainable development research and policy

In today's world, individual countries are becoming increasingly interconnected, which is reflected not only in economic and political relations, but also in scientific research and publications. By looking at keywords associated with particular countries, specific areas of interest can be identified that not only reflect national priorities, but also indicate the global direction of research and policies.

In the Czech Republic, scientific publications focus on addressing environmental challenges such as air pollution, and emphasise the importance of technological innovation and the transition to renewable energy sources. This focus on "crisis management", "digitalisation" and "renewable energy" suggests a strong perception of the need to address current environmental and technological challenges. Poland´s interest in "urban planning", "sustainable development" and "air quality" points to a desire to improve the quality of life in cities and to focus on sustainable development. This theme aligns with the broader European trend of enhancing urban environments and reducing pollution.

China and India, as two of the world's most populous countries, address key words related to "carbon emissions", "economic growth", "air pollution" and "sustainable development". These areas indicate a recognition of the need to reconcile economic growth with environmental protection and to seek ways to achieve sustainable development.

Germany and the UK, with their strong focus on "renewable energy" and "climate change", are leading European efforts to combat climate change and transition to renewable energy. This is in line with their policies and public initiatives on energy transition. In northern Europe, Norway is showing an interest in "electric vehicles" and "renewable energy", reflecting its advanced position in the development of electric mobility and the use of renewable energy. This is an example of how national policies can support innovation in sustainable transport and energy.

In the East, Russia remains strongly focused on "energy" and "oil", reflecting its key role as an energy power and the importance of the oil industry to its economy.

This diversity of keywords and their links to specific countries reveals that while some countries may specialise in different research and policy areas, there is a common interest in addressing global challenges such as climate change, sustainable development and energy transition. The scientific publications therefore provide valuable insights into how individual countries are approaching these challenges and the direction the global scientific community is taking in the search for solutions.

4. Discussion

In the Discussion section, the authors can elaborate on how their findings answer the previously formulated research questions and reflect on their implications for theory, practice and future research based on the questions as well as the results provided.

RQ1: What is the impact of advanced technologies and innovations on the achievement of sustainable mobility and green infrastructure in cities?

The results show that advanced technologies and innovations, particularly in the fields of artificial intelligence, big data and IoT, play a key role in the development of sustainable mobility and green infrastructure. Clusters such as "Advanced technologies and smart infrastructure" and "Sustainable transportation and mobility" highlight the potential of technology to effectively manage urban spaces, reduce emissions and improve quality of life. The link between technology and sustainable mobility suggests that integrating innovation into urban planning and transport systems is essential to achieve sustainable cities. Kubik et al. (2023) and Dvořáček et al. (2022) emphasise the importance of electric vehicles (EVs) and autonomous technologies in reducing CO_2 emissions, and support the idea that their integration into urban plans is crucial for achieving sustainable cities.

RQ2: What are the key success factors for integrating sustainable business strategies and environmental policies in companies and public administrations?

The integration of sustainable business strategies requires the active involvement of companies in CSR, digitalisation and eco-innovation practices, as the cluster "Business strategies and sustainable development" shows. Key success factors include the development of business models that integrate sustainability as a core principle, and collaboration between the public and private sectors to design and implement effective environmental policies. In addition, the link between business strategies and urban development highlights the importance of economic growth in line with sustainable goals. Sosnowski (2022) Implementing green concepts in the supply chain. This study highlights the importance of reducing environmental impact and promoting sustainable business models.

RQ3: How can interdisciplinary approaches be used to address global challenges such as climate change and sustainable development through local and regional innovation?

Interdisciplinary approaches are key to addressing the complex challenges of sustainable development. The links between different clusters such as “Environmental impact and pollution management”, “Energy and climate policy” and “Behavioural insights and societal change” indicate the need for collaboration across disciplines. Taking into account specific contexts and needs, local and regional innovations can make an important contribution to the achievement of global Sustainable Development Goals (SDGs) by linking scientific research, policy and practice. In a study discussing opportunities for the sharing economy in the forestry sector, Palátová (2023) presents an innovative approach to forest resource use and management as an example of an interdisciplinary approach to sustainable development.

An analysis of keywords in sustainable development research highlights the importance of an interdisciplinary approach, integrating technology, social sciences, economics, and environmental sciences. This reflects the notion that sustainability challenges cannot be tackled by individual disciplines alone. Renewable energy emerges as a crucial trend, with terms such as “solar”, “wind”, and “renewable energy sources” frequently linked with political and economic aspects, underscoring its significance for environmental sustainability, economic growth, and innovation. Sustainable mobility is also gaining focus, especially through the interconnectedness of “electric vehicles”, “sustainable transport”, and “urban mobility”, emphasizing the shift towards green mobility and emission reduction in line with global Sustainable Development Goals. Additionally, the role of education and community involvement is seen as key to spreading sustainable practices and raising awareness of environmental challenges, suggesting that active participation and education are vital for a sustainable future. The conducted analysis also indicates that addressing global issues such as climate change necessitates combining international co-operation with local innovation, showcased by the linkage of global challenges with local and regional solutions. These insights underline the multidisciplinary complexity of sustainable development, stressing that its advancement relies on collaboration across various disciplines and societal sectors.

5. Conclusion

The authors conclude their paper by exploring key aspects and trends in sustainable development, the integration of advanced technologies and innovation in urban environments, and the role of interdisciplinary approaches to address global challenges. The findings show that advanced technologies such as artificial intelligence, big data and IoT have a significant impact on achieving sustainable mobility and green infrastructure development. The key success factor for integrating sustainable business strategies includes engagement of companies in CSR, digitalisation and eco-innovation practices. The importance of interdisciplinary approaches that bring together different disciplines and sectors to effectively address the complex challenges associated with sustainable development was also highlighted. The paper also shows that understanding and integrating local and regional innovations is essential for tackling global challenges such as climate change. The aforementioned professional studies present innovative approaches to resource use and management, highlighting the need for interdisciplinary collaboration and support from the public and private sectors. Additionally, the findings point to the need for further research and experimentation in sustainable development and the integration of technological innovation. It is essential to focus on developing and testing new models and strategies that can deliver tangible results towards a more sustainable and greener future. Collaboration between academia, industry and public administration will be crucial in relation to bridging the gap between theory and practice and achieving the SDGs.

In conclusion, this paper highlights the importance of integrating advanced technologies and innovation, sustainable business strategies, and interdisciplinary approaches as the main ingredients for addressing sustainable development challenges. The authors aim to understand how these elements mutually contribute to a better future for cities and society as a whole.

Acknowledgement

We thank for the financial support provided by the IGA PEF 2024A2001 grant "Potential of the Sharing Economy in the Areas of Carsharing and Homesharing Services.

References

Brescia, V., Degregori, G., Maggi, D., & Hadro, D. (2023). An integrated vision of electric vehicles" consumer behaviour: Mapping the practitioners to consolidate the research agenda. Journal of Cleaner Production, 410. ISSN 1879-1786

Chmiel, B., Pawlowska, B., & Szmelter-Jarosz, A. (2023). Mobility-as-a-Service as a Catalyst for Urban Transport Integration in Conditions of Uncertainty. Energies, 16(4). ISSN 1996-1073

Donthu, N., Gremler, D. D., Kumar, S. and Pattnaik, D. (2022). Mapping of Journal of Service Research Themes: A 22-Year Review. Journal of Service Research, 25(2). ISSN 1552-7379

Drobiazgiewicz, J., & Pokorská, A. (2023). Directions of Carsharing Development in Poland-Analysis of the Need to Expand the Carsharing Zone. Sustainability, 15(5). ISSN 2071-1050

Dvořáček, L.; Horák, M.; Knápek, J. (2022). Simulation of Electric Vehicle Charging Points Based on Efficient Use of Chargers and Using Recuperated Braking Energy from Trains. Energies, 15(2), ISSN- 1996-1073

Ke, Q., Silka, J., Wieczorek, M., Bai, Z. W., & Wozniak, M. (2022). Deep Neural Network Heuristic Hierarchization for Cooperative Intelligent Transportation Fleet Management. IEEE Transactions on Intelligent Transportation Systems, 23(9). ISSN 1558-0016

Kimbrell, C. M. (2021). Electric carsharing and the sustainable mobility transition: Conflict and contestation in a Czech actor-network. Energy Research & Social Science, 74. ISSN 2214-6326

Kubera, M.; Ślusarczyk, B. (2023). Carsharing as an Element of Mobility Management. Polish Journal of Management Studies, 27(2). ISSN 2081-7452

Kubik, A.; Turon, K.; Folega, P.; Chen, F. CO2 Emissions (2023). Evidence from Internal Combustion and Electric Engine Vehicles from Car-Sharing Systems. Energies, 16(5). ISSN 1996-1073

Matowicki, M., Amorim, M., Kern, M., Pecherkova, P., Motzer, N., & Pribyl, O. (2022). Understanding the potential of MaaS - An European survey on attitudes. Travel Behaviour and Society, 27. ISSN 2214-3688

Neumann, T. (2021). The Impact of Carsharing on Transport in the City. Case Study of Tri-City in Poland. Sustainability, 13(2) ISSN 2071-1050

Palátová, P; Rinn, R; Machon, M; Palus, H; Purwestri, RC; Jarsky, (2023) Forest policy and economics, 154, ISSN 1872-7050

Pawełoszek, I. (2022). Towards a Smart City—The Study of Car-Sharing Services in Poland. Energies, 15(22), ISSN 1572-9435

Piotrowski, P., Baczynski, D., & Kopyt, M. (2022). "Medium-Term Forecasts of Load Profiles in Polish Power System including E-Mobility Development." Energies, 15(15). ISSN 1996-1073

Sosnowski, P.C. (2022). Green concepts in the supply chain. Logforum, 18(1). ISSN 1734-459X

Turon, K. (2022). The Expectations towards Cars to Be Used in Car-Sharing Services-The Perspective of the Current Polish Non-Users. Energies, 15(23). ISSN 1996-1073

Van Eck, N. J., & Waltman, L. (2022). VOSviewer manual. Retrieved from: https://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.18.pdf

Vejchodská, E.; Foltynová, HB.; Rybicková, A. (2023). Carsharing users" behaviour and attitudes. The role of car availability in households. Transportation, published, volume and issue not available, ISSN 1572-9435

LOW-COST UTILISATION OF UAV IMAGERY FOR MUNICIPALITIES

Ivana Čermáková, Petr Rozehnal, Roman Danel, Ivan Šotkovský, Jakub Illík

VSB – Technical University Ostrava
Faculty of Economics
ivana.cermakova@vsb.cz, petr.rozehnal@vsb.cz, roman.danel@vsb.cz, ivan.sotkovsky@vsb.cz, jakub.illik.st@vsb.cz

DOI: 10.35011/IDIMT-2024-179

Keywords

UAV; regional development problems; decision making support

Abstract

Using SMART technologies is one of the pilar of SMART cities and municipalities. SMART technologies and methods offer financial and time benefits. Among these technologies, unmanned aerial vehicles (UAVs) stand out as a valuable tool that can help resolve various issues in regional development. This paper presents the theoretical background of UAVs and their use by municipalities. Areas of potential usage are also obtained. The entire concept is validated through a specific case study in the village of Lučina. The concept is divided into the following steps: negotiation and problem definition. preparation (documents and pre-processing), flight itself, data processing, data evaluation, data distribution and information. Each step is divided into the following substep. The article mainly presents the methodological approach and the results of the initial stages of the solution, which were applied to a specific case study for municipalities. The time range of the monitoring and the financial range with the calculation of savings using UAV is defined concretely for the case study.

1. Introduction

SMART cities are defined like cities that utilize SMART technologies. The term SMART refers to specific, measurable, achievable, realistic and time-boundary technology or product. Using of SMART technology aims to decrease time and financial difficulties of solving various tasks/problematics. One of the concepts of SMART cities obtained utilization UAVs. (Ministerstvo pro místní rozvoj, 2022)

The term UAV is explained as aerial vehicles flying without pilots (Work Jr & Gilmer, 1976), but modern literature explained the term like unmanned aerial vehicle. More about the definition and using UAV is obtained in the previous paper (Čermáková, & Danel, 2022) and in the section of problem statement.

Using of Artificial Intelligence (AI) with combination with UAV is one of the possibilities for the future. Some researchers try to apply to Precision agriculture (PA) (Su et al., 2023) or for searching and possible rescuing missions (Herschel et al., 2022). But with using AI comes problems and uncertainties (e.g. in laws or responsibilities) (Hashesh et al., 2022). When these questions are answered, it will be possible to include AI to the concept of the SMART cities.

2. Problem statement

UAVs can be used for different tasks / problems in regional development and urban planning. In addition to their standard applications in urban planning such as village expansion, brownfield monitoring (Čermáková, Danel & Vašenková, 2022) or monitoring of shoreline changes (Čermáková, 2022) UAVs are being utilized for new benefits.

Graffiti has become a significant problem in urban development due to its prevalence. The semi-autonomous drone can detect graffiti in its surroundings and decide whether to remove it (Wang et al., 2019). UAVs can be used to remove graffiti from unreachable areas, such as bridge overpasses. However, they can be equipped with a spray paint holder to cover the graffiti if necessary. A Graffiti detection model that utilizes edge detection and machine learning algorithms can be employed for real-time detection and tracking of graffiti images (Nahar et al., 2017).

UAVs are gaining attention in air monitoring applications due to their exceptional agility in both horizontal and vertical dimensions. They can acquire near-surface vertical profiling of atmospheric pollution with high spatial resolution (Gu & Chunrong, 2019). UAVs equipped with specialized sensors can identify certain pollutants. These UAVs can monitor air pollution and communicate with a PC through a long-range wireless communication system. The availability of this information can aid decision making in smart cities to reduce air pollution and improve air quality (Hernandez-Vega et al., 2017). In the Czech Republic, controlling of pollution from chimneys via UAVs is suggested.

UAVs can be used for agriculture operations, including the distribution of seeds, water, pesticides, and fertilisers. UAVs can provide regular crop inspections to assess crop health and monitor crop development to determine optimal harvesting times or the need for preventive measures (Nader et al., 2020). PA uses information from both airborne and ground sensors to assess and forecast crop conditions (Tokekar et al., 2016). The recent trend of using UAVs in smart agriculture shows the increasing success of deep learning methods in accurately addressing PA tasks (Shahi et al., 2022). Further information on the use of UAVs in PA can also be found in (Čermáková & Danel, 2022) as well.

The application can be very beneficial for the regional development. Utilizing open-source software, expenses related to software, licensing, data processing, distribution, visualization, and dissemination can be minimized. Without the price of the vehicle, which can be bought or leased, the costs are minimal against traditional methods, e.g. geodetics research.

This study focuses on the utilization of UAV as a low-cost decision support solution in the form of visualized information that replaces or supplements the existing information of the solver for given situation. The main aim is to define the process and means for deploying the UAVs in situations where the client (in this case regional development) does not have large financial possibilities and UAV technologies can offer a cheap solution. It is therefore not only a technical problem, but also an economic and organizational problem. The uniqueness of the study lies in its multidimensional overlap, which shows the use of UAVs in the context of other aspects.

3. Research questions

Considering the wide use of UAVs, the study focusses on the following research questions.

Research Question (RQ) 1: Is it possible to respond to the real demand of an organisation/institution development where specific documents are required for management support and deliver these materials without the necessity of purchasing software licences for processing UAV-generated data?

RQ 2: Is it realistic to use UAV technology to reduce geospatial costs compared to traditional survey methods?

RQ 3: Can UAV imagery outputs be used for other needs of regional development (e.g., propagation or data distribution to the public) and not only for management and decision support?

4. Case study – Lučina

Lučina village is situated near the Žermanická Dam in the Moravskoslezský region of the Czech Republic. Expansion of local infrastructure and the possibility of using newly acquired non-residential buildings are one of plans for the next decade. Lučina village is also part of the Žermanická and Těrlická Dam Microregion. Strengthening tourism, tourism and recreation is part of the microregion's strategy. As part of these activities, it was decided to build a bike path around the dam. In the first stage (beginning in spring 2024), a part will be implemented in the Lučina village. Subsequently, two more stages will be built, which pass through the villages of Dolní Domaslavice and Soběšovice.

The planned bike path (Fig. 1) is the largest event for the municipality - controversial with regard to decision-making, a major intervention in the future life of the municipality. For municipalities, the construction of a cycle path is an important intervention, as the effects cannot be accurately predicted. An increased movement of people on and around the bike path, changes in the availability of the water surface, a possible increase in demand for parking spaces, etc. are expected. Construction-related information is important to the municipality now and in the long term. UAV imaging offers a portfolio of information that cannot be secured in any other way. A view from above with adequate quality can suitably complement other information and support decision-making, information and other processes in the municipality.

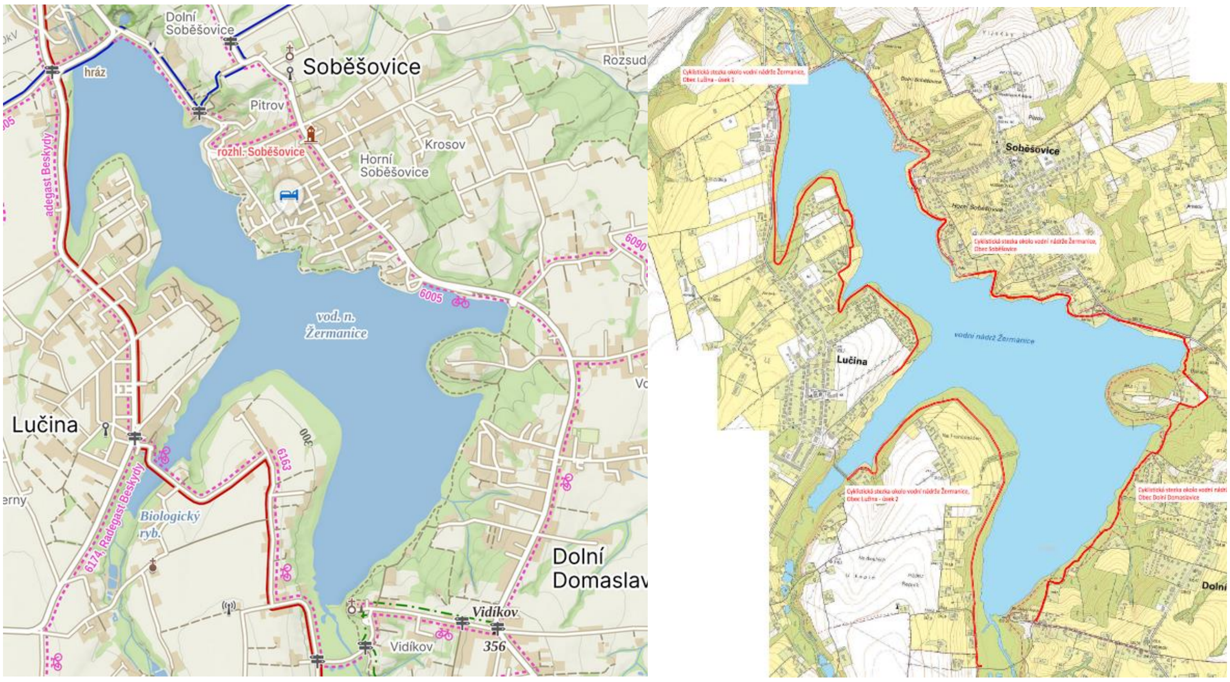


Figure 1. Žermanická Dam (left part of image) Source: (Mapy.cz) **and planned bike path – red line (right part of image)**

Source: (internal documents of Žermanická and Těrlická Dam Microregion)

As part of UAV imaging, the contracting authority - the municipality has selected the following areas (visible in Fig. 2 also):

Area of interest (AoI) 1: the area where the campsite is located with access to the water surface. The place has been critical before, as there is a large movement of people.

AoI 2: Selected, partial places on the cycle path, where additional visual background is assumed to be needed, e.g., documentation of illegal construction of accesses to the water surface, modification of banks, etc. (in connection with the construction of the cycle path, there will be partial removal of trees and thus also the uncovering of areas).

AoI 3: Comparison of cycling route management - assumption versus reality. The municipality cares about compliance with the established route, and imaging will allow inspection without additional geodetic surveying, which would be time and financially demanding.

The aspect of time is also important. UAV imaging of all selected areas will enable comparison over time series. On the one hand, in the context of a longer time period, on the other hand, in connection with changes in different seasons. In this sense, the project will be a pilot project for the municipality and it is expected that the requirements will be specified based on the first outputs from the UAV and further on the basis of experience from the operation of the cycle path. The outputs from the imaging can be used to inform the citizens of the municipality and the public about ongoing construction of the bike path. The aspect of financial part of the surveying is important and should also be calculated.

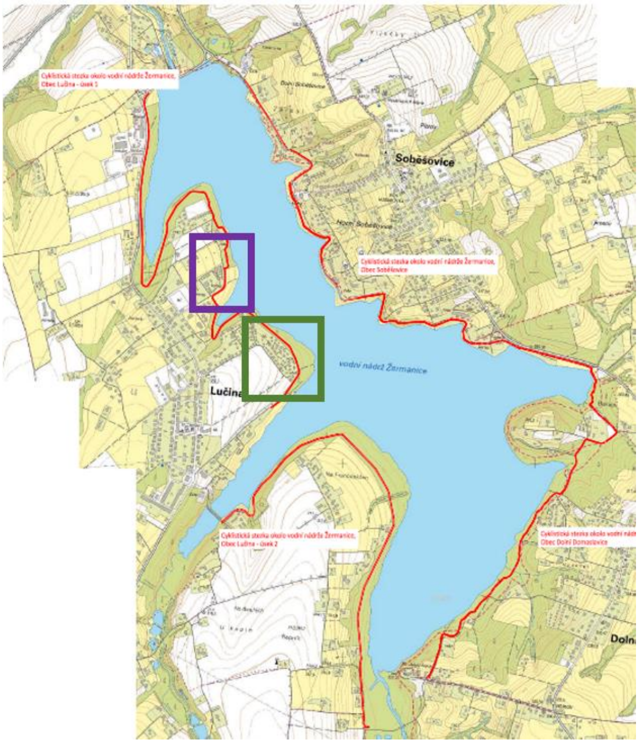


Figure 2. Žermanická Dam – selected areas of interest (AoI 1 – purple rectangle, AoI 2 – green rectangle, AoI 3 – comparing plans with imagery from AoI 1 and 2)
Source: (internal documents of Žermanická and Těrlická Dam Microregion edited by the authors)

5. Methodology design

The concept is divided into the following steps: negotiation and problem definition, preparation (documents and pre-processing), flight itself, data processing, data evaluation, data distribution and information. Each step is divided in detail into the following substeps.

Negotiation and problem definition. Each case of using a UAV will have its own specifics that must be taken into account (initial situation and definition of goals). Each municipality is different, so it is necessary to provide basic information about municipality (area, type of vegetation, type of buildings, water diversity, electrical grid and other important information) and near municipality vicinity. In the first meeting with the representative of the municipality, the possibilities of UAV are discussed. Choosing suitable problems to solve is provided in the meeting with municipal representatives. Only suitable areas of interest are chosen for this type of monitoring. A feasibility analysis should be carried out. This is focused on the time of monitoring, area and restriction of the chosen area of interest. If everything is suitable, then the area is chosen for UAV monitoring. If there are some problems that cannot be solved (e.g., restriction of the flying in the area by the Civil Aviation Authority – CAA in the case of Czech Republic), different method (e.g., geodetic survey) must be chosen.

Preparation and pre-processing are focused on permission to fly in area of interest by owners of the area (mainly municipality itself), current insurance and fly pass for the pilot, control and preparing all equipment for the flight itself (targets, cameras, batteries, vehicle, medium for the data and route definition and settings).

Flight itself is divided on preparation for the flight, flight itself and control after flight. First, it is necessary to check the area of interest. During the control, targets can be placed. Following the preparation of the UAV and other equipment. Mainly are controlled batteries, camera (and their clear lens), connection between camera and UAV, medium for the data, connection between pilot equipment and UAV. After control is provided, flight itself in chosen height and overlap. After flight, control is provided of the type and accuracy of final imagery or video. If it is not suitable for next processing, it is necessary to provide flight again.

Data processing is following step. The whole imagery is created like e.g., orthophoto mosaic. Specialized software is necessary (open software is suitable too). From orthophoto is created the result. It depends on the focus of the monitoring. If it is study changes, then classification and evaluation of classification is chosen. If it is modelling some situation, then creating the model and evaluation of model is chosen. These two types of data processing are used mostly.

Data evaluation is provided after data processing by some sophisticated method. Each method provides a different evaluation. The evaluation can also be provided in open-source software. If the result is not accurate enough, it is necessary to perform data processing again with different settings or methods.

Data distribution and information is step which is communicated with municipality representants also. If the data are used publicly, it is necessary to provide them in some specific form (e.g., like maps or in the brochure). If the data are used, for example, planning of the municipality, the primary data shall be distributed. It all depends on the negotiation and purpose of the monitoring. It has to be negotiated who should provide which data and in which form to whom (data processor, municipality, public).

Solution description and discussion

The flight itself and realization will be provided in the next few months. At this time (preparation of the conference paper) first and second phases of the methodology were finished, and the resources for implementation (hardware and software) were defined. All activities from Preparation and pre-processing are finished. The flight and processing of the first data is planned for April 2024 (and next months). The chosen areas should have finished the construction of the bike path in the autumn. The surveying will be provided during this period. According to the current situation. Data processing will be provided gradually and the complete measurement will be provided after finalization of bike path in Lučina.

At this point, according to the interim results, we can state the following.

RQ1 – suitable areas, based on negotiations with municipality representatives, were chosen for monitoring. The assessment is described below. Documents from UAVs are a desired additional or alternative source of information for similar situations at the level of municipal management. This is also true for RQ3. The required outputs can also be generated by freely available software.

RQ2 - The calculation of the costs is listed below in the paper. The images are accurate enough to monitor the cycle paths. The municipality does not have an accurate geodetic base map/documents, so only a comparison of the planned bike path with reality will be provided. In this case, UAV monitoring is more advantageous. The costs are described in the following paragraphs.

We present the evaluations for RQ1 and RQ2 together, as they partially overlap. For the calculation, it is necessary to include several elements (at least: UAV, software, implementation of imaging and data processing). A UAV of type DJI Phantom 3 costs around 40 – 60 000 Czech crowns (CZK), for monitoring rarely is not necessary to buy the UAV (Renting is around 4 000 CZK per hour, 6 000 CZK per hour with pilot.). But if the municipality wants to survey the area more often than once a year, purchasing its own UAV is a cost-effective option. The comparison always assumes a variant solution. In the case of the study, for RQ1, the decisive factor is the software and its possibilities, for RQ2 it is a confrontation with another measurement method and the quality of the results. In the Lučina Case Study, the measurement sections (areas) were defined and the costs for both variants were estimated. The estimated time for measuring AoI using geodets is around 8 – 12 hours (2 people) and then data processing. The hour rate for geodetics is between 800 and 1 200 CZK. So, the first outputs will be available no sooner than 2 days with minimal costs of 12 800 CZK (recommended time requirement in man-hours). But it could cost even more. In the case of UAV, two hours takes flight and operations between flight and if is suitable accuracy 10 cm and more, then next processing is for next 2 hours. If the costs for the work are maintained, the estimate of the final price is half, that is, much smaller than in the first case. Although Agisoft Metashape is paid software (16 000 CZK per licence) it is not necessary for ortomosaic. Open-source software can be used. VSB – Technical University Ostrava has a licence for Agisoft Metashape, so the authors used this software. In measurement, Agisoft Metashape and Techradar provide comparable output, but output generation takes longer with open software, and it is not as intuitive as in the case of commercial software. QGIS is open-source software. The images of each survey are measurable between itself. Changes during bike path construction can be easily visible and used for public visualisation information.

RQ3 – All documents can be used for discussion between the municipality representative and the public. Visualisations can be used to make decisions in the village and inform local residents about construction progress through village publications. Here, defining the problem (contractor - representatives of the municipality) and the goal of using the results, i.e., the delivered materials, is decisive.

The advantages of using UAVs for this task are time and cost savings. The main savings can be achieved through the speed of data collection for processing and subsequent processing with direct visualisation. The disadvantage is the inability to fly (mainly weather conditions, such as strong winds). However, proactive planning aligned with weather forecasts effectively mitigates this drawback in practice. Problematic areas for UAV monitoring with cameras that provide visible imagery are mostly areas with dense vegetation, such as mature trees and flaming plants. In this case, the surface is not visible. There are two potential solutions to address this problem. Firstly, using the UAV with the thermal imaging camera. Alternatively, manual mapping techniques, such as geodetic mapping, can be used. However, this approach may be limited by terrain accessibility, rendering certain areas inaccessible for surveying.

6. Conclusion

The paper presents a theoretical background of UAV and applies it to a case study of a smaller municipality. The whole concept is validated via concrete municipality – Lučina village. The concept is divided into the following steps: negotiation and problem definition, preparation and pre-processing, flight itself, data processing, data evaluation, data distribution and information. The focus of this paper lies in the theoretical part of the process and possibilities of using UAV to address challenges in regional development within municipalities. Although actual implementation will take place in the coming months, it is already clear that the research questions can be answered mostly in the affirmative. However, practical implementation can modify the perception of some aspects (e.g., scope of imaging, time frame) in the context of financial characteristics. The next implementation will then verify all steps of the process and create space for further development/questions/challenges. The current project involving the construction of a bike path is chosen for monitoring purposes, allowing for a comparison with traditional surveying methods. The comparison reveals substantial cost and time savings achievable through UAV.

Acknowledgement

This paper has received financial support from the SGS SP2024/052 grant provided by VSB – TU Ostrava.

References

Fahlstrom, P. G. & Gleason, T. J. (2012). Introduction to UAV systems, Aerospace series. (4th ed.). John Wiley & Sons.

Work Jr, E. A. & Gilmer, D. S., (1976). Utilization of satellite data for inventorying prairie ponds and lakes. Photogrammetric Engineering and Remote Sensing, 42(5), 685–694.

Úřad pro civilní letectví (2018). Úřad pro civilní letectví. Retrieved April 12, 2022, from www.caa.cz.

Čermáková, I., Danel, R. & Vašenková, T., (2022). Brownfields opportunities for smart regions. In: IDIMT – Interdisciplinary Information Management Talks 2022, 375-380.

Čermáková, I., & Danel, R. (2022). Usage of unmanned aerial vehicles for detection of plant anomalies. In: IDIMT – Interdisciplinary Information Management Talks 2022, 367-374.

Čermáková, I. (2022). Metodický postup pořizování a zpracování dat z UAV v prostředí vodních ploch. Doctoral Thesis. VSB – Technical University Ostrava, Faculty of Economics.

Gu, Q. & Chunrong, J. (2019). A Consumer UAV-based Air Quality Monitoring System for Smart Cities. In: 2019 IEEE International Conference on Consumer Electronics (ICCE). ISBN: 978-1-5386-7910-4.

Hernandez-Vega, J. et al. (2017). Internet of Things (IoT) for Monitoring Air Pollutants with an Unmanned Aerial Vehicle (UAV) in a Smart City. In: Smart Technology, vol. 213. Lecture Notes of the Institute for Computer Sciences Social Informatics and Telecommunications Engineering. Doi: 10.1007/978-3-319-73323-4_11, 108–120.

Ministerstvo pro místní rozvoj (2022). Retrieved April 8, 2024, from https://www.mmr.cz/cs/narodni-dotace/podpora-a-rozvoj-regionu/podpora-rozvoje-regionu-2020/podpora-obnovy-a-rozvoje-venkova.

Nader, M. et al. (2020). Unmanned aerial vehicles applications in future smart cities. In: Technological Forecasting and Social Change 153 (Apr. 2020). Doi: 10.1016/j.techfore.2018.05.004.

Nahar, P. et al. (2017). Autonomous UAV Forced Graffiti Detection and Removal System Based on Machine Learning. In: 2017 IEEE Smartworld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SMARTWORLD/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI). ISBN: 978-1-5386-0435-9.

Shahi, T. B. et al. (2022). Machine learning methods for precision agriculture with UAV imagery: a review. In: Electronic Research Archive 30.12 (2022), doi: 10.3934/era.2022218, 4277–4317.

Tokekar, P. et al. (2016). Sensor Planning for a Symbiotic UAV and UGV System for Precision Agriculture. In: IEEE Transaction on Robotics 32.6 (Dec. 2016), doi: 10.1109/TRO.2016.2603528, 1498–1511.

Wang, S. et al. (2019). Building Smart City Drone for Graffiti Detection and Clean-up. In: IEEE Smartworld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation. ISBN: 978-1-7281-4034-6, 1922–1928.

Hashesh, A.O. et al. (2022). AI-Enabled UAV Communications: Challenges and Future Directions. IEEE Access 10, https://doi.org/10.1109/ACCESS.2022.3202956, 92048–92066.

Herschel, R. et al. (2022). UAV-borne remote sensing for AI-assisted support of search and rescue missions, in: Kamerman, G.W., Steinvall, O. (Eds.), ELECTRO-OPTICAL REMOTE SENSING XVI, Proceedings of SPIE. Presented at the Conference on Electro-Optical Remote Sensing XVI, Spie-Int Soc Optical Engineering, Bellingham, p. 1227203. https://doi.org/10.1117/12.2636032

Su, J. et al. (2023). AI meets UAVs: A survey on AI empowered UAV perception systems for precision agriculture. Neurocomputing 518. https://doi.org/10.1016/j.neucom.2022.11.020, , 242–270.

ADOPTION MODEL OF AI – USE CASE OF AUTONOMOUS VEHICLES

Richard Antonín Novák, Tomáš Sigmund

Prague University of Economics and Business
Faculty of Informatics and Statistics
richard.novak@vse.cz, sigmund@vse.cz

DOI: 10.35011/IDIMT-2024-187

Keywords

artificial intelligence (AI), autonomous vehicles (AV), automation, autonomy, transparency, trust, motivation, value-based adoption model (VAM)

Abstract

In this paper, we explore a model of AI technology adoption for the case of autonomous vehicles (AVs). The model we propose is based on the general Value-based adoption model (VAM). In addition to the perceived benefits and sacrifice components of the VAM: utility, enjoyment, and fee, technicality our model also includes AI-specific components: human autonomy, automation and AI transparency. Our model reflects the dilemma between the human desire for autonomy and efficient machine automation, which can be invasive for humans. The proposed model was validated by the target group in a survey concerning the adoption of AV as an exemplary case of AI limiting human autonomy and showed a coefficient of determination of 46.8% which is higher than the original VAM model which explained only 35.9% of the variability of the same target variable, i.e. the intention to adopt the new technology.

1. Introduction

The exponential growth of commercial AI projects (Srivastav et al., 2022) leveraging the capabilities of technologies incorporating AI patterns is increasingly encroaching into our daily lives and expanding our options (Hassani et al., 2020). In addition, AI is increasingly moving from the realm of external recommendation engines to the internal and intimate realm of human senses such as sight, hearing, and touch, using technologies such as computer vision, natural language processing, and tactile sensors (Fernandez & Fernandez, 2021; Liu et al., 2020). Moreover, large AI service providers such as Amazon, Google, Facebook, Apple increasingly prefer to sell persistent services based on data and algorithms that have little user control (Agrawal et al., 2020). This limits the autonomy of AI users as they become dependent (Endsley, 2017). These new AI services and in particular the model example of AVs that limit human autonomy motivated our research. The aim of this paper is to clarify the process of user adoption of AI technologies while clarifying the dilemma between the human desire for autonomy and the proliferation of machine automation, which brings mainly economic, but also user benefits. To clarify the dilemma, we will explore the effect of AI transparency and autonomy on trust and compatibility between AI technology and its user. This model will be combined with the existing theoretical models. So far, models combining theoretical knowledge from psychology, sociology and technology have been used to explain buying behavior, with the following models being the most common: (TAM) Technology Acceptance Model (Kamal et al, 2020), (UTAUT-2) Unified Theory of Acceptance and Use of Technology (Venkatesh, 2012) and (VAM) Value-based Adoption Model (Kim et al., 2007). The integration of AV into society is now a very hot topic, which is also in line with new EU legislation such as the AI-Act, (2024).

The selection, use and extension of specific models for the adoption and acceptance of new technologies was most often done intuitively in the context of the specifics of individual target stakeholders or customer groups respecting their habits (e.g, Novak et al., 2022). Above all, however, the choice of model was related to the technology under consideration, such as HealthTech (Fan et al., 2018), EduTech (Rafique et al., 2020) or AI robots (Liang & Lee, 2017). The performance of the models, measured as the ability to predict and explain user behavior in the process of acceptance of a given technology, varied widely according to the specific technology under consideration.

To date, there are only a few studies that have taken the trouble to assess adoption of a single technology using multiple models simultaneously on the same sample of respondents. Of particular note here are Sohn et al. (2020) comparing TAM, UTAUT, and VAM on the adoption of AI-based intelligent products in South Korea where VAM came out significantly best with the dominance of enjoyment and social norms factors, or Rahman et al. (2017) comparing TAM, and UTAUT on driver acceptance of technologies in driving assistance systems.

In our work, we focus on the evaluation of the deployment of autonomous vehicle (AV) technology as a model example, as it is a very sensitive area and an intrusion on driver autonomy. We have also chosen the AV use case because the limitation of user autonomy due to the automation of hitherto predominantly human activities is easily understood by our research respondents. For the AV domain, we can also build well on previous research in the area of AVs adoption (e.g., Rahman et al., 2017; Liu at al., 2019) and at the same time try to suggest incremental improvements to the previous VAM model, (Kim et al., 2007).

User adoption of technology can be considered as a special type of decision-making situation combining the insights from theories of user's motivation described in psychology and sociology, and decision-making process describing choice of consumers in economics and marketing, (Kim et. al., 2007). The concept of consumer value can be seen as a fusion of the perspectives of psychology and economic reasoning. The value function is based on a psychological foundation and replaces the utility function of economic theory. Kim's 2007 Value-based Adoption Model (VAM), which is the initial inspiration for our work, uses ***perceived value*** consisting of *perceived benefits* and *fees* as the basic concept. "Perceived value is the consumer's overall assessment of the utility of a product based on perceptions of what is received and what is given.", (Zeithaml, 1988)

Kim's VAM works with components ***Usefulness*** and ***Enjoyment*** in the *Perceived benefits* area. On the *Perceived sacrifices* side, meaning what the consumer gives up for the value received, the VAM uses the term which includes for financial costs the label ***Perceived fee*** and for non-financial costs it uses the term ***Technicality*** which usually includes time, effort and other unsatisfactory spending for the purchase and consumption of the product, (Kim, 2007). In the case of many AI innovations, such as AVs, the user will often make decisions about their intention to adopt a new technology as Consumer on his/her own and based only on available descriptive information. It's due to the fact that specific technology will often not be available for a test at the time of the decision. We are aware that the notion of ***Adoption Intent*** used in VAM may change at the moment of actual customer experience. Still, we consider *Adoption Intent* based on Thaler (1985) as a more important decision tendency of consumer choice than the notion of ***Attitude*** used for technology acceptance models, (TAMs). The above reasons, as well as the great simplicity of the VAM model based on the aggregate notions of perceived benefits and sacrifice, led us to be inspired to use VAM as the basis for our extension to AI-specific constructs that we are going to describe more in details.

2. Research model

In the following paragraphs, we will describe the constructs from which the final intention for the adoption of AI technology can be inferred. For a given construct, we always provide a description of the construct, and note whether it is an original part of the VAM model, and when it is our own extension, as well as based on what considerations and sources.

2.1 VAM model, based on Kim et. al., (2007)

USEFULNESS, (Benefit component 1)

"Usefulness is defined as the total value that the user gets from using the new technology, (Rogers Everett, 1995). The motivation-oriented perspective of TAM looks at the perceived utility as an expectation of outcome and a measure of extrinsic motivation, (Venkatesh, 1999). Individuals evaluate the consequences of their behavior in terms of perceived utility and base their behavioral choices on desirable utility. Performance expectations such as perceived utility, which focuses on task accomplishment, reflect the desires of an individual to engage in an activity because of external rewards. The construct of utility has been widely used in information systems and technology research and has strong empirical support as an important predictor of Technology adoption)." (Kim et. al., 2007)

ENJOYMENT, (Benefit component 2)

"Individuals, who experience immediate pleasure or joy from using a technology aside from the instrumental value of the technology, are more likely to adopt the technology and use it more extensively than others", (Davis et. al., 1989). "Enjoyment refers to the extent to which the activity of using a product is perceived to be enjoyable in its own right, apart from any performance consequences that may be anticipated", (Davis et. al., 1992). "Enjoyment thus represents an affective and intrinsic benefit and is characterized by what customers receive as emotional response/joy received from purchase and product quality", (Kim et. al, 2007).

PERCEIVED FEE, (Sacrifice component 1)

"Perceived fee (price) symbolizes the encoding or internalization of the objective selling price of a product/service", (Jacoby et. al., 1977). "Without any experience with new technologies such as AVs, customers cannot judge whether the fee quoted to them is high or low. According to the Adaptation level theory, instead of having perfect information about prices, customers possess internal reference prices and make comparisons with these prices." (Grewal et. al. 1998). "In the case of AVs, customers would probably compare the fee of AV usage with previously encoded prices of mobile phone calls and stationary Internet access. The result of this comparison forms the customers' perception of the fee." (Kim et. al., 2007).

TECHNICALITY (Sacrifice component 2)

"Technicality of the system is a combination of all the non-monetary costs. These costs include time costs, search/effort costs, convenience costs and psychological costs." (Zeithaml, 1988). In an AV environment, loading and response time can be considered time costs while ease of use and connectivity are considered effort and convenience cost, respectively. Psychological factors include inner conflict, frustration, depression, discomfort, anxiety, tension, annoyance, mental fatigue, etc.” (Kim et. al.,, 2007)

2.2 New constructs:

USERS TRUST IN AUTOMATION, (AI specific construct 1)

Automation can be described as “technology that actively selects data, transforms information, makes decisions, or controls processes” (Lee & See, 2004). (Ghazizadeh, Lee, and Boyle, 2012) developed an automation acceptance model. It includes external variables which influence factors of compatibility and trust and TAM factors. The model reflects voluntary automation adoption which is different to job related situations. Authors of this model recommend using the model for the adoption of AV. Trust is a conceptual term used by many different disciplines that is important when dealing with cooperation between multiple parties (interpersonal, human-automation) where there are elements of expectation and uncertainty, (Hoff & Bashir, 2015). For our general model of AI adoption, initial trust based on beliefs formed by within-reach transparent information is important, and we are aware that it is rapidly being replaced as the initially primary factor in importance by the accumulated experience with automated systems, such as system reliability and predictability, (Madhavan & Wiegmann, 2007B).

TRANSPARENCY OF AI, (Component 1)

Transparency means provision of information on inner workings of the system to the user (Wang and Benbasat, 2009). The received information signals, system quality and working, increases user confidence in the system and reduces the necessity and desire for control. Transparency also familiarizes the user with the system and creates an illusion of control. Transparency may affect the willingness to rely on the system’s decision. We can differentiate information transparency as the degree to which users can analyze the information used by the system and algorithm transparency as the degree to which the user can monitor and comprehend the performance of the system. Based on Bass et al. (2013) the transparency can be defined as the following: „Providing users with accurate, ongoing feedback concerning the reliability of automation and the situational factors that can affect its reliability in order to promote appropriate trust and improve task performance.“

COMPATIBILITY OF USER-AI, (AI specific construct 2)

Compatibility reflects the match among the operator, the technology, the task to perform, and the situation (Karahanna et al. 2006). More specifically, compatibility measures a technology’s consistency with users’ values, past experience, and needs (Rogers 1995).

In a simple way, compatibility means the technology fits with users' valuesbased on which users act. Ghazizadeh et al. (2012) emphasized the role of compatibility and argued that types of driving automation should match and be consistent with how the user might act. In their extended TAM model, they added trust and compatibility, to assess the automation acceptance model. May et al. (2017) also included compatibility in their driving automation acceptance model and revealed the positive impact of compatibility on perceived usefulness.

HUMAN AUTONOMY, (Component 2)

Within the definitional framework for self-determined behavior, a behavior is autonomous if the person acts (a) according to his or her own preferences, interests and/or abilities, and (b) independently, free from undue external influence or interference. Noom, Deković and Meeus (2001) differentiate three types of autonomy, the attitudinal, the emotional and the functional.

In accordance with the self-determination theory (Ryan and Deci, 2000) is autonomy one of the basic psychological needs and can be defined as an experience of choice and psychological freedom concerning one's activities.

"Autonomy refers to self-government and responsible control for one's life" (Keller, 2016, p. 1). Inductively, autonomy determines the capacity of an individual to make informed and uncoerced decisions. In this respect, having autonomy also determines the extent to which an individual has control over their options and choices and meets their desires accordingly.

2.3 Model

Based on the description in the previous section, we plan to extend the VAM model proposed by Kim et al. (2007) that builds on the constructs: *utility, enjoyment and fee, and technicality.* We propose our extension constructs specific to AI as: *user trust in automation* and *user compatibility with AI* consisting of moderating components of AI transparency interfering with human autonomy.

3. Methodology and Questionnaire Design

To validate our Adoption Model of AI (AMAI), we chose a questionnaire survey with individual questions inspired by the original VAM, TAM models. The questions were measured using a five-point Likert scale with the anchors "strongly disagree" and "strongly agree". Age was measured in years. Gender was bipolar male, female. And the questionnaire was in English.

3.1 Data Collection

We organized online data collection using Survey Monkey questionnaires in 2023 where we collected firstly response from 21 respondents in focus group of the students from Prague University of Economics and Business. The sample size was small, but the results were promising, so we decided to use the same questionnaire on a larger scale for respondents from India which is the most easily accessible region for us near the original VAM model survey (Singapore). We collected 276 responses where respondents were paid. The structure was 50.5% male and 49.5% female. In terms of age, we had only two categories, 18-29 years (92%) and 30-44 years (8%).

3.2 Data Evaluation

For the survey evaluation we have used descriptive part common metrics (means, variations) in Exploratory Data Analysis (EDA) the results of which can be found in the appendix in the form of means and standard variations of answers to questions. We also used Partial Least Squares Structural Equation Modelling (PLS-SEM). We also used Bootstrapping in a PLS software tool to compensate for the small number of focus group observations. We had to exclude some questions as they loadings were too small (below 0.6). We also had to remove all items of the construct ease of use (Q7) as they were not significantly correlated to any other constructs and would so decrease the model validity.

4. Results

The Figure 1 below shows the results, with the value in the blue circle indicating the R^2 i.e. the coefficient of determination expressing the quality of the regression model. The yellow rectangles in the figure indicate the identification of a particular question i.e. q001-001 stands for Construct-1:

Intention to Adopt (q001). The following (-001) stands for the ordinal question of the given construct (1 out of 3), the specific wording of which is traceable in the annex of the complete questionnaire and reads: *I plan to use AV in the future*, where possible answers follow Likert scale *(Strongly Agree, Agree, Neutral, Disagree, Strongly disagree).*

Overall, our model explained 46.8% of the variability in the outcome variable: *Intent to adopt AV*, which is a very good result for a social variable. We confirmed the strong influence of *AI transparency* on *trust* (0.674) and on *compatibility* (0.496). In our survey we also confirmed the relationship between *human autonomy* and *compatibility* (0.282). The influence of *compatibility* and *trust* on the *intention to adopt an AV* is limited, but we found some influence. The effect of *trust* in AI on *intention to purchase AV* is slightly higher (0.165) than the effect of *compatibility* (0.141) values of the user with AV technology.

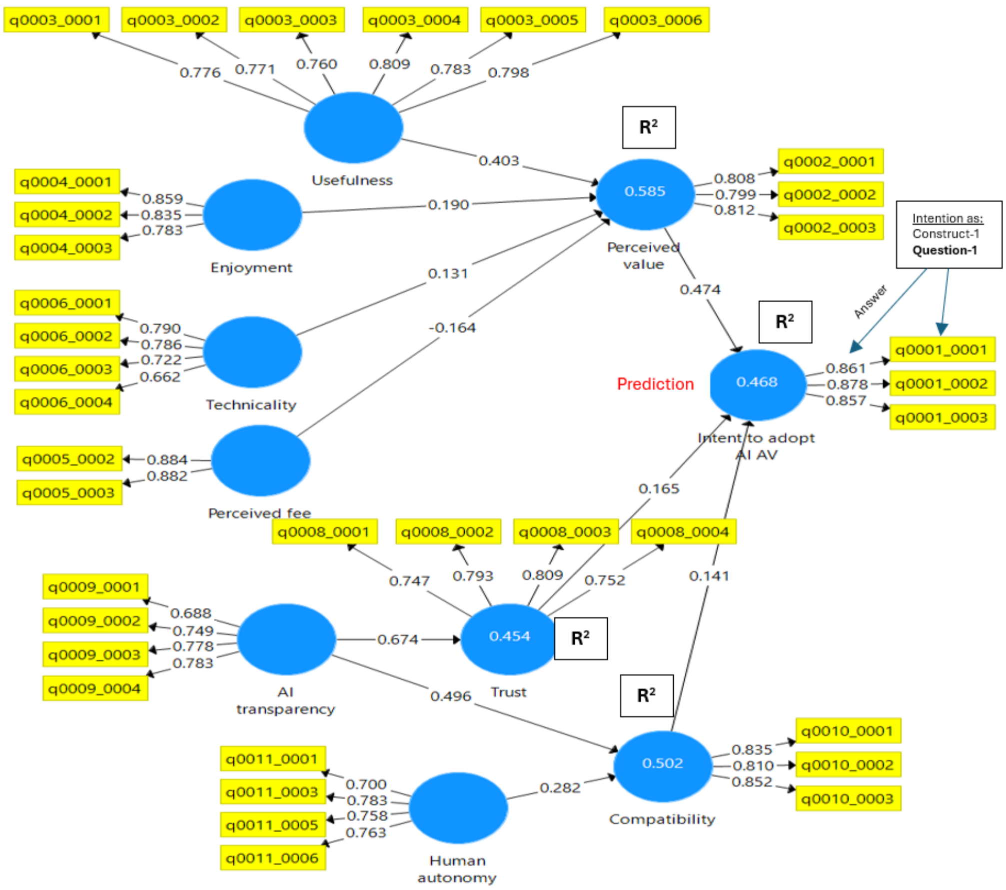


Figure 1. AMAI model results Source:(Authors)

5. Discussion

We focused on the extension of the VAM model with transparency, autonomy as specifics of AI automation. Our model reflects the specificity of AI systems that were not considered in the original VAM version of Kim (2007). We explained almost 46,8% of the intention to adopt AV comparing to original VAM method that explained 35.9% variance which is interesting. What we should emphasize in our interpretation is that while our respondent sample was comparable in size to the original VAM model survey, we had a total of 276 respondents in 2023 and Kim had 161 in 2007 with similar demographics. Although, the survey was conducted in different countries, and we asked about different technology. We conducted the survey in India and Kim (2007) in Singapore, and while we asked about AV adoption Kim asked about mobile internet adoption. Our survey showed very promising results and would require further detailed investigation on the attribution of individual constructs to explain the overall variable. Applying the survey to a larger sample to confirm our findings is something that would support the sound theoretical foundations of our AMAI model. For now, we consider the AMAI to be an initial proposal and in need of further investigation.

6. Conclusion

Following the survey, we can say our new variables add value to the explanation of the intention to adopt AVs. Interesting is the interrelation of the constructs even though their influence on the intention to adopt is limited. The model was successfully tested on a reduced group of 276 respondents, and therefore it is necessary to continue the research and verify its results on a large data collection, ideally by comparing the same demographic group of students in the Czech Republic and e.g. in the USA, which is the area of our further planned research.

Acknowledgement

Paper was processed with support from institutional-support fund for long-term conceptual development of science and research at the Faculty of Informatics and Statistics of the Prague University of Economics and Business (IP400040) and the project IG 409033.

References

Agrawal, A., Gans, J., & Goldfarb, A. (2018). Prediction machines: the simple economics of artificial intelligence. Harvard Business Press.

Ajzen, I. (2020). The theory of planned behavior: Frequently asked questions. Human Behavior and Emerging Technologies, 2(4), 314-324.

Bass, E. J., Baumgart, L. A., & Shepley, K. K. (2013). The effect of information analysis automation display content on human judgment performance in noisy environments. Journal of cognitive engineering and decision making, 7(1), 49-65.

Bagloee, S. A., Tavana, M., Asadi, M., & Oliver, T. (2016). Autonomous vehicles: challenges, opportunities, and future implications for transportation policies. Journal of modern transportation, 24(4), 284-303.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. Management science, 35(8), 982-1003.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1992). Extrinsic and intrinsic motivation to use computers in the workplace 1. Journal of applied social psychology, 22(14), 1111-1132.

Fernandez, R., & Fernandez, T. F. (2021). 2 Voice Recognition Using Natural Language Processing.

Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned From Human–Automation Research. Human Factors: The Journal of the Human Factors and Ergonomics Society, 59(1), 5–27. https://doi.org/10.1177/0018720816681350

Ghazizadeh, M., Lee, J. D., & Boyle, L. N. (2012). Extending the Technology Acceptance Model to assess automation. Cognition, Technology & Work, 14(1), 39–49. https://doi.org/10.1007/s10111-011-0194-3

Grewal, D., Monroe, K. B., & Krishnan, R. (1998). The effects of price-comparison advertising on buyers' perceptions of acquisition value, transaction value, and behavioral intentions. Journal of marketing, 62(2), 46-59.

Hassani, H., Silva, E. S., Unger, S., Mazinani, M., & Mac Feely, S. (2020). Artificial intelligence (AI) or intelligence augmentation (IA): what is the future?. Ai, 1(2), 8.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human factors, 57(3), 407-434.

Jacoby, J., & Olson, J. C. (1977). consumer response to price: An attitudinal, information processing perspective, oin moving ahead with attitude research, y. Wind and P. Greenberg, eds. Chicago: American Marketing Association, 73, 86.

Kamal, S. A., Shafiq, M., & Kakria, P. (2020). Investigating acceptance of telemedicine services through an extended technology acceptance model (TAM). Technology in Society, 60, 101212.

Keller, H. (2016). Psychological autonomy and hierarchical relatedness as organizers of developmental pathways. Philosophical Transactions of the Royal Society B: Biological Sciences, 371(1686), 20150070. https://doi.org/10.1098/rstb.2015.0070

Kim, H. W., Chan, H. C., & Gupta, S. (2007). Value-based adoption of mobile internet: an empirical investigation. Decision support systems, 43(1), 111-126.

Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance

Liu, P., Yang, R., & Xu, Z. (2019). How safe is safe enough for self-driving vehicles?. Risk analysis, 39(2), 315-325.

Liu, Y., Bao, R., Tao, J., Li, J., Dong, M., & Pan, C. (2020). Recent progress in tactile sensors and their applications in intelligent systems. Science Bulletin, 65(1), 70-88.

Madhavan, P., & Wiegmann, D. A. (2007B). Similarities and differences between human-human and human-automation trust: An integrative review. Theoretical Issues in Ergonomics Science, 8, 277–301.

Novak, R., Sigmund, T., Bohmova, L., (2022). AUTONOMOUS VEHICLES AS A MANAGEMENT CHALLENGE - MEETING CONFLICTING INTERESTS. In IDIMT conference 2022 at Prague 7-9.th Sept. 2022.

Rafique, H., Almagrabi, A. O., Shamim, A., Anwar, F., & Bashir, A. K. (2020). Investigating the acceptance of mobile library applications with an extended technology acceptance model (TAM). Computers & Education, 145, 103732.

Rogers Everett, M. (1995). Diffusion of innovations. New York, 12.

Thaler, R. Mental accounting and consumer choice, Marketing Science 4 (3) (1985 (Mar.))

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 55(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Sohn, K., & Kwon, O. (2020). Technology acceptance theories and factors influencing artificial Intelligence-based intelligent products. Telematics and Informatics, 47, 101324.

Srivastav, S., Garg, V., & Elngar, A. A. (2022). The Impact of Artificial Intelligence for Comprehensive Growth Across the Globe. In Empowering Artificial Intelligence Through Machine Learning (pp. 137-165). Apple Academic Press.

Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS quarterly, 157-178.

Venkatesh, V. (1999). Creation of favorable user perceptions: Exploring the role of intrinsic motivation. MIS quarterly, 239-260.

Zeithaml, V. A. (1988). Consumer perceptions of price, quality, and value: a means-end model and synthesis of evidence. Journal of marketing, 52(3), 2-22.

Appendix A1 - Survey Items: Available at: https://zenodo.org/records/10966245

RULE-BASED APPROACH USING THREATGET FOR AUTOMATICALLY GENERATING ATTACK PATHS IN INDUSTRIAL AUTOMATION AND CONTROL SYSTEMS

Abdelkader Magdy Shaaban, Korbinian Christl,
Christoph Schmittner

Center for Digital Safety & Security
AIT Austrian Institute of Technology
abdelkader.shaaban@ait.ac.at, korbinian.christl@ait.ac.at,
christoph.schmittner@ait.ac.at

DOI: 10.35011/IDIMT-2024-195

Keywords

Cybersecurity; attack paths; and industrial automation and control systems

Abstract

The rapid development and integration of Internet-based technologies in recent years have become essential for fully automating industrial control systems and enhancing decision-making, precision, and efficiency in managing complex and large-scale production systems. However, due to these remote connections, any vulnerable cybersecurity point could expose the entire system to cyberattacks. Therefore, it is essential and highly needed in such complex systems to identify any possible threats that could be triggered within the system. Additionally, building all possible paths based on the identified threats will be essential to highlight all attack paths that could pave the way for attackers to reach malicious goals. This paper introduces a novel, rule-based approach to determine and build any possible attack paths for critical assets in complex systems such as automated industrial environments. This approach is part of ThreatGet, which deeply investigates all attack steps, emphasizing the violated security properties and assessing each threat's impact and feasibility. This will support identifying existing security vulnerabilities in the system model and determining how to conduct more thorough security investigations to build a secure, automated industrial infrastructure.

1. Introduction

The Industrial Control System (ICS) is a type of control system consisting of multiple technologies used to provide full control of the industrial process. It consists of several types of systems, such as Distributed Control Systems (DCS), Supervisory Control and Data Acquisition (SCADA) systems, Programmable Logic Controllers (PLC), and others (Stouffer et al., 2015). The ICS facilitates the Operational Technology (OT) infrastructure, which mainly compasses a set of computing systems responsible for controlling industrial operations. OT is different from the meaning of the Information Technology (IT) systems, whereas the IT is mainly for the aspect related to the data management (GALE, 2023).

The ICS facilitates the OT infrastructure, primarily comprising computing systems responsible for controlling industrial operations. The rapid evolution of technologies used in the industrial process created the need for full monitoring and control, driven by full automation for more efficient decision-making, higher precision, and greater efficiency. This paved the way for the term "automation" to become an essential part of the ICS and introduced the IACS, or Industrial Automation and Control Systems. Automation for the industry has a wide range of benefits, including increased productivity, improved safety, enhanced production quality, faster response to changes in production requirements, reduced production costs, and increased efficiency (Trainer, 2023).

Due to the rapid industrial revolution in automation systems, there is Industry 4.0 and then Industry 5.0, which reshape our understanding of future industries. Industry 4.0, which originated in Germany in 2011, is considered to be technology driven. The European Commission then declared the new industry form, Industry 5.0 since 2021, value-driven (Xu et al., 2021). Both Industry 4.0 and Industry 5.0 are characterized by their reliance on automation technology and a rising connectivity and computation power in OT to ensure and maintain efficient operations and increase inbuilt intelligence and automation. However, one of the most challenging aspects is the integration of commercial-off-the-shelf (COTS) devices because they are inexpensive, efficient, and highly automated, and they can be interconnected in heterogeneous environments (ISA, 2013). These devices may not be fully suitable for critical industrial tasks since they are never designed to work within a critical environment (A. M. Shaaban et al., 2022). Therefore, different system weaknesses could exist in such devices, potentially leading to serious or even catastrophic scenarios in control systems (ISA, 2013). Cybersecurity vulnerabilities are considered one of the main essential issues in these devices.

The reason for this is that cyberattacks targeting industrial systems pose a significant cyber threat, demanding increased attention to secure these systems and all related critical assets. Security measures are essential to protect them from unauthorized access (Osliak et al., 2023). Therefore, it is essential to ensure the implementation of security measures, including protecting data against multiple cyber threats such as integrity breaches, unauthorized information exposure, spoofing, and other malicious actions. The ISA/IEC 62443 (ISA/IEC, 2024) is one of common international cybersecurity standards for the industrial system (Conti et al., 2021). The standards provide a cybersecurity framework to address the current and future security vulnerabilities in industrial automation systems (A. Shaaban, 2021). Furthermore, identifying any potential security weakness that an attack could exploit is essential to be identified and evaluated, in order to help in understanding which security requirements are needed to protect the system and its critical assets from different attack scenarios. It is essential to map all scenarios that could lead to a successful attack propagation that an attacker could follow to target a particular goal. This can be done manually if the possible attack scenarios are limited, but in a complex system such as smart automated industrial environment, all scenarios should be considered to build full paths that an attacker could take to achieve a particular malicious goal. The need to automate the propagation of attack paths is highly essential. This process is significantly enhanced by threat modeling approaches, which not only help identify potential threats but also establish attack paths in the system model. Additionally, threat modeling facilitates the evaluation and comparison of capabilities for technologies, products, and services, aiding in the selection of appropriate mitigation strategies to reduce cyber risks (Dobaj et al., 2023)

Therefore, in this paper, we introduce a novel rule-based approach that utilizes the ThreatGet threat modeling tool to automatically generate all possible attack paths an attacker could use, based on the existing security vulnerabilities in the system. Our approach builds a complete map between multiple rules to investigate any possible threats that can be propagated due to the violations of security properties in the system design and its network. Each step in the path is considered a potential threat. ThreatGet defines the capabilities of attackers required and provides more details about the current threat and what the next threat could be triggered due to the success of an attack. The risk estimation for the entire path is automatically calculated based on the assessments of all identified threats. This calculation is evaluated using multiple factors that offer a deeper investigation into the impact of potential threats and the feasibility of an attack.

2. Related Work

Thoroughly investigating multiple attack paths is an effective approach for determining any possible ways attackers could attack the system. Attack trees are considered a graphical representation approach capable of showing multiple paths that attackers could use to attack a system or a critical asset. They show their essential need for providing friendly graphical representation models that can provide more precise investigation about the multiple stages that attackers can use for achieving an attack (Kumar et al., 2022). Kumar et al. (Kumar et al., 2022) demonstrate how attack trees can be utilized as a common language to model Advanced Persistent Threats (APTs) in a more friendly way. In their research, they have investigated three different attacks to describe a systematic way to follow the attack steps, which can also be used in more complex scenarios for modeling more sophisticated attacks.

In (He et al., 2020), the authors presented another approach for attack trees, mainly focusing on the attack-defense tree. The attack-defense tree is primarily an extension of attack trees for analyzing defense scenarios. It mainly provides an in-depth security analysis approach to investigate the relationships between attacks and defense in a model. The proposed approach provides multiple security attributes for attack and defense leaf nodes, combining the characteristics of all leaf nodes and providing a fuzzy matrix to estimate the weight of each node's security attributes, along with estimating the probability of attack and defense scenarios for each node.

Attack trees can support other cybersecurity approaches, such as Intrusion Detection Systems (IDS), as most research on that topic mainly focuses on improving algorithms and optimizing feature selection, as discussed in (Xingjie et al., 2020). A proposed approach for the IDS is based on the attack trees and the Long Short-Term Memory (LSTM) as introduced in (Xingjie et al., 2020). That approach mainly aims at building all attack paths for industrial control systems from the perspective of attackers and utilizing the LSTM algorithm to determine and classify attack behaviors. Each attack is defined by one or more atomic actions, which are simply frequent commands or actions occurring within the network, and these atomic actions are extracted to further classify hacker attack methods. Based on this extraction, the proposed approach can infer that an attacker is attempting to perform an attack against the system based on the tree structure.

A survey on the automatic generation of attack trees and attack graphs is defined by (Konsta et al., 2024), which illustrates the current status of automatic generation techniques for attack trees and attack graphs. The research supports presenting the most common and current approaches for describing the main techniques for automatically generating attack models. Additionally, it discusses the challenges and opportunities in that area for future directions.

3. Automatic Attack Paths Generation

As previously discussed in Section 2, we have shown that there are multiple approaches for generating attack trees to provide different methods for generating trees with paths that offer a graphical representation of all paths attackers could follow to achieve a particular malicious goal. In ThreatGet, we added a new feature, mainly providing a deep investigation of all possible paths attackers could follow. ThreatGet provides a rule-based approach for determining all possible paths to attack a critical asset in a system. Each step of the attack demonstrates the required capability of the attacker to progress toward the attack. Once the attacker takes a step, it provides another capability to execute the next action on the affected system component. Each step of the path is represented as a threat against the system's component. All identified threats in the path have more details about the risk estimation, the classification of threats, etc.

3.1. Threat Classification

ThreatGet uses the STRIDE model (Shostack, 2014) to describe the malicious behaviour for each threat; this model is defined as follows: Spoofing is a type of threat that mainly aims to violate authentication. Tampering refers to actions that compromise integrity, while repudiation is a threat that aims to violate non-repudiation. Information disclosure involves actions that breach confidentiality. Denial of Service (DoS) is a threat targeting the availability of a service. Lastly, Elevation of Privilege involves actions that violate authorization.

3.2. Risk Estimation

ThreatGet performs risk estimation for each step in the attack path, which provides a deeper investigation of the risks associated with each step an attacker can follow to attack a system. To ensure precise risk estimation, this approach is based on likelihood and impact estimations.

3.3. Impact Estimation

ThreatGet utilizes the SFOP (S: Safety, F: Financial, O: Operation, and P: Privacy) classification to identify the impact of each threat on specific system components. This classification provides a high level of the impact categories due to the cyber attacks. For instance, an Operational impact indicates potential disruption to the system's overall operation if the threat is triggered. Each category has a range referring to the impact level, ranging from Negligible, which is a low level of impact, to Severe, which represents critical impact. For example, a cyber attack against a robotic arm in a smart factory could have a safety impact that might harm assets or personnel in the surrounding areas. Additionally, it could also have an impact on its operation, which could shut it down completely.

3.4. Likelihood Estimation

In order to provide a more precise estimation of the likelihood, recent versions of ThreatGet have integrated multiple factors to assess the feasibility of an attack. Attack feasibility provides more details about how easily an attack can occur. According to ISO/SAE 21434: Road Vehicles - Cybersecurity Engineering (ISO/SAE, 2021), the standard refers to Common Criteria for Attack Feasibility to accurately estimate the attack likelihood. Therefore, in our approach, we utilize these criteria to provide more details on the likelihood, as discussed in (AIT, 2023). Then, the overall risk of the path is estimated according to the risk estimation for each step in the path.

3.5. Attack Paths in ThreatGet

The path generation in ThreatGet is driven by a set of rules stored in its database. These rules are described in a customized language developed to provide a more structured manner for describing threats. The ThreatGet's rule engine utilizes these rules to detect any potential threats that could be propagated within a system network due to violating security attributes. Then, ThreatGet maps these threats to describe a path representing the attack steps that attackers could follow to complete a particular malicious goal.

Each step in the attack path describes two capabilities: requires and provides. These capabilities describe how the "REQUIRES CAPABILITY" is needed to perform a particular action towards a system component. On the other hand, once the attacker gains access to a particular component, the capability that can be provided (i.e., "PROVIDES CAPABILITY") defines the next required capability for the next action. List 1 shows a high-level overview of the capabilities in the attack path based on ThreatGet's language using the Extended Backus-Naur Form (EBNF) syntax.

List 1. EBNF Syntax for describing the capability in attack path based on ThreatGet's language

CapabilityCondition ::= RequiresCapability | ProvidesCapability
RequiresCapability ::= "REQUIRES CAPABILITY" capName ">="CapabilityValue
ProvidesCapability ::= "PROVIDES CAPABILITY" capName ":=" CapabilityValue
capName ::= "Access" | "Read" | "Control" | "Data Manipulation" | "Physical Access"
CapabilityValue ::= "true" | "false"

There are multiple capabilities that an attacker might need and can provide. These capabilities are defined as follows:

- **Access**: Refers to gaining access to a particular component within the system.
- **Read**: A capability that might be required or provided to read data from a component.
- **Control**: An attacker might require or provide a control against a particular system item.
- **Data Manipulation**: Another type of capability where the attacker could perform data manipulation against the components.
- **Physical Access**: The attacker might require or provide physical access to perform actions against a component.

4. Attack Path Generation for a Smart Factory Example

This section discusses an example of how an attack path is automatically generated using our approach in ThreatGet, representing the route an attacker could follow to target a critical asset within a smart factory's premises. Figure illustrates a simple example of a smart factory.

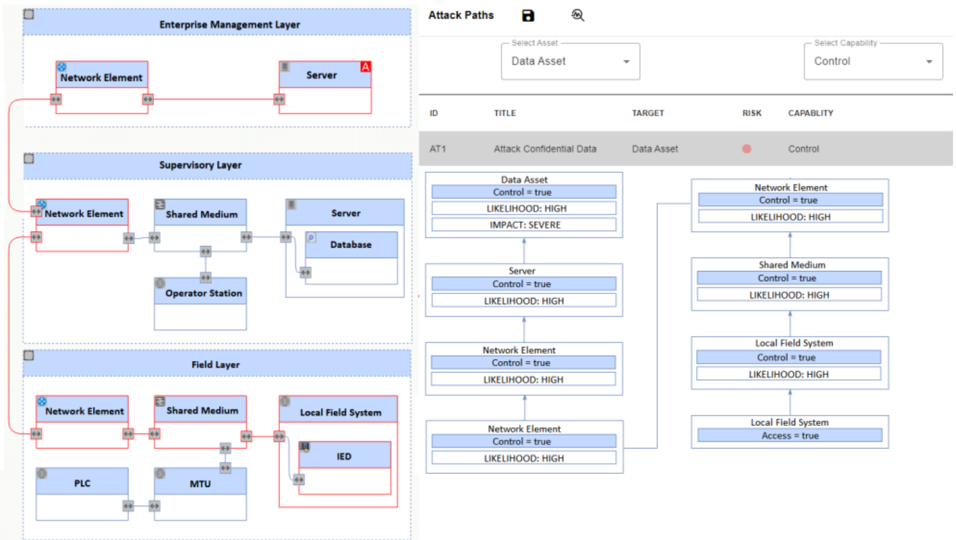


Figure 1. An example of the Smart Factory System and the automatically generated attack path

There are three main sections in this example, as follows:

- **Field Layer**: This area encompasses multiple field devices responsible for performing actions to control the process and machines using PLCs and providing additional controls on the manufacturing process through MTUs (Master Terminal Units). Communication from the Field area to Supervision is mainly through network devices.
- **Supervision Layer**: This layer manages the communication traffic between the Field and the Enterprise Management Layers. It consists of various critical devices for storing data regarding the operation status in the production process. There is also a server that is responsible for processing all collected data.
- **Enterprise Management Layer**: Contains a server that plays a significant role in the entire smart factory regarding process planning, data analysis, and critical decision systems. A critical asset for this server is the confidentiality of all its data. Therefore, in the event of any violation of this data confidentiality, we can expect various consequences to occur. We consider this server our main target as it contains sensitive information, which is considered our critical asset in this example.

ThreatGet applies a set of rules stored in its database to determine any possible paths an attacker could utilize to reach a malicious goal against the confidentiality of the server data in our example. An attack path is automatically generated by ThreatGet, as shown in Figure. This attack focuses on the attack surface of the field layer, targeting any vulnerable points within it. The attacker can target unsecured control units or devices in the Field layer. The attacker gains control if these devices lack a secure update procedure or malware protection. Once in control of the devices, the attacker is already within the system and can progress further via unsecured communications, gradually gaining more control over connected components step by step. Finally, the attacker can access the Enterprise Management Layer directly via the network elements. The attacker then attacks the server where the data asset is located, which is the target of this attack. To describe each step individually, we have defined the following ThreatGet's rules:

1. RULE 1 - Identify Potential Components of the Attack Surface of the Field Layer: This rule is used to identify potential points of attack within the Field layer. All elements in this layer could be considered untrustworthy or within a trusted element. This gives the attacker excess access to this element.

```
ELEMENT IN ["Device", "Control Unit" ]
{ HAS ATTRIBUTE "Trusted" != "Yes" &
NOT CONTAINED BY ELEMENT {
HAS ATTRIBUTE "Trusted" = "Yes" }
PROVIDES CAPABILITY "Access" := "true"
}
```

2. RULE 2 – Exploiting the Vulnerabilities of the Attack Surface of the Field Layer: As soon as an attacker has access to an element, he can gain control over it. The system also checks whether this element can be updated, has malware protections, or anomaly detection.

```
ELEMENT IN ["Device", "Control Unit" ] {
REQUIRES CAPABILITY "Access" >= "true" &
EVALUATE ATTRIBUTE "Updatable" &
EVALUATE ATTRIBUTE "Malware Protection" &
EVALUATE ATTRIBUTE "Anomaly Detection" &
 PROVIDES CAPABILITY "Control" := "true"
}
```

3. RULE 3 - Propagating the Exploit within the System: This is a propagation rule. This means that it is used to spread the influence of the attacker within the system; we focus on the connectors within the system, we check whether the attacker already has control over the source element, and if this is the case, then the attacker also gains control over the target element if this also does not have malware protection or anomaly detection. In addition, the target interface must have no input control and weak or no authentication.

```
CONNECTOR {
SOURCE ELEMENT {
REQUIRES CAPABILITY "Control" >= "true" } &
TARGET ELEMENT {
EVALUATE ATTRIBUTE "Malware Protection"&
EVALUATE ATTRIBUTE "Anomaly Detection"&
 PROVIDES CAPABILITY "Control" := "true" } &
TARGET INTERFACE{
EVALUATE ATTRIBUTE "Input Control" &
EVALUATE ATTRIBUTE "Authentication" }
}
```

4. RULE 4 - Exploiting the Asset: The last rule only specifies that an element that the attacker already controls hold an asset. Control is also transferred to this asset, and the attacker can dispose of it.

```
ELEMENT { REQUIRES CAPABILITY "Control"
>= "true" & HOLDS ASSET {
ROVIDES CAPABILITY "Control" := "true"   }
}
```

This set of rules estimates the path that ThreatGet automatically creates, providing a graphical representation of these threats, as illustrated in Figure. These represent the steps an attacker follows to obtain confidential data from the server inside the Enterprise Management layer. The attack path depicted in the figure outlines all steps determined by ThreatGet, beginning at the Field Layer, and continuing through the propagation of all related threats until reaching the main targeted asset. ThreatGet highlights in red all system components affected within the defined attack path.

5. Conclusion

This paper discusses our novel rule-based approach for automatically generating attack paths in the industrial automation system. Thus, the approach aims to build a mapping action among multiple rules described as part of the ThreatGet database, developed in a customized language and deployed to be utilized by ThreatGet's rule engine. The rule engine starts investigating all rules and tries to find any possible threat that could be propagated within the system design due to violating security properties. Then, it builds a complete graphical representation to show the possible ways an attacker could follow to attack a critical asset in the system. An example of a smart factory is used to demonstrate how this approach can be applied in such an example. This approach can also be concluded with multiple research studies in the automotive domain, railways, and CPS, which provide a robust methodology for estimating different paths attackers could take to achieve malicious goals. Our future vision is to integrate Graph Neural Networks (GNNs) in the context of this work, which is mainly based on the graph structure and provides a deeper investigation of complex connections and relations between system components and identifying potential threats.

Acknowledgement

This work has received funding from the AIMS5.0 project. AIMS5.0 has been accepted for funding within the Key Digital Technologies Joint Undertaking (KDT JU), a public-private partnership in collaboration with the HORIZON Framework Programme and the national Authorities under grant agreement number 101112089.

References

AIT. (2023). THREATGET Documentation. https://documentation.threatget.com/24.04/Web/Threat%20Analysis.html

Conti, M., Donadel, D., & Turrin, F. (2021). A Survey on Industrial Control System Testbeds and Datasets for Security Research. IEEE Communications Surveys & Tutorials, 23(4), 2248–2294.

Dobaj, J., Macher, G., Ekert, D., Riel, A., & Messnarz, R. (2023). Towards a security-driven automotive development lifecycle. Journal of Software: Evolution and Process, 35(8), e2407.

GALE, T. (2023, November 30). Cybersecurity Risk Assessment Strategies in Industrial Control Systems. https://www.aiche.org/resources/publications/cep/2023/december/cybersecurity-risk-assessment-strategies-industrial-control-systems

He, S., Lei, D., Shuang, W., Liu, C., & Gu, Z. (2020). Network Security Analysis of Industrial Control System Based on Attack-Defense Tree. IEEE International Conference on Artificial Intelligence and Information Systems, 651–655.

ISA. (2013). NIST Cybersecurity Framework ISA99 Response to Request for Information [International industry standards development organization]. https://www.nist.gov/system/files/documents/2017/06/01/040513_international_society_automation.pdf

ISA/IEC, I. (2024). ISA/IEC 62443 Series of Standards—ISA. Isa.Org. https://www.isa.org/standards-and-publications/isa-standards/isa-iec-62443-series-of-standards

ISO/SAE. (2021). ISO/SAE 21434: Road vehicles Cybersecurity engineering. ISO. https://www.iso.org/standard/70918.html

Konsta, A.-M., Lluch Lafuente, A., Spiga, B., & Dragoni, N. (2024). Survey: Automatic generation of attack trees and attack graphs. Computers & Security, 137, 103602.

Kumar, R., Kela, R., Singh, S., & Trujillo-Rasua, R. (2022). APT attacks on industrial control systems: A tale of three incidents. International Journal of Critical Infrastructure Protection, 37, 100521.

Osliak, O., Saracino, A., Martinelli, F., & Mori, P. (2023). Cyber threat intelligence for critical infrastructure security. Concurrency and Computation: Practice and Experience, 35(23), e7759.

Shaaban, A. (2021). An ontology-based cybersecurity framework for the automotive domain: Design, implementation, and evaluation [PhD Thesis, University Vienna - Faculty of Computer Science]. https://utheses.univie.ac.at/detail/59948

Shaaban, A. M., Chlup, S., El-Araby, N., & Schmittner, C. (2022). Towards Optimized Security Attributes for IoT Devices in Smart Agriculture Based on the IEC 62443 Security Standard. Applied Sciences, 12(11), Article 11.

Shostack, A. (2014). Threat modeling: Designing for security. John Wiley & Sons.

Stouffer, K., Lightman, S., Pillitteri, V., Abrams, M., & Hahn, A. (2015). Guide to Industrial Control Systems (ICS) Security. NIST Special Publication.

Trainer, I. (2023, March 19). Industrial Automation: Introduction, Types, Elements, Advantages, and Disadvantages - INDUSTRIAL TRAINER. https://www.industrialtrainer.org/industrial-automation

Xingjie, F., Guogenp, W., ShiBIN, Z., & ChenHAO. (2020). Industrial Control System Intrusion Detection Model based on LSTM & Attack Tree. IEEE International Computer Conference on Wavelet Active Media Technology and Information Processing, 255–260.

Xu, X., Lu, Y., Vogel-Heuser, B., & Wang, L. (2021). Industry 4.0 and Industry 5.0—Inception, conception and perception. Journal of Manufacturing Systems, 61, 530–535.

INNOVATION AS A CATALYST FOR SUSTAINABILITY: EXAMINING THE INFLUENCE OF R&D AND PATENT ACTIVITIES ON THE SUSTAINABLE BRAND INDEX

Maria Kmety Bartekova, Helena Majduchova

University of Economics in Bratislava
Faculty of Business Management
maria.bartekova@euba.sk, helena.majduchova@euba.sk

DOI: 10.35011/IDIMT-2024-203

Keywords

Sustainable development goals; sustainable consumption and production; Brand index; innovation.

Abstract

In this paper, we assess the impacts of Research and Development (R&D) spending and patent filings on the Sustainable Brand Index (SBI) across Denmark and Sweden from 2017 to 2022. Through linear regression analysis of yearly data, we find a robust model fit, indicating that a large portion of SBI variability can be accounted for by these factors. Our findings reveal that patent filings have a notable positive effect on SBI, suggesting that innovation, as quantified by patent activity, is beneficial for sustainability performance. However, R&D expenditure appears to have a negative influence on SBI, though this association lacks statistical significance, implying that R&D investment may not have a straightforward link with the sustainability measures captured by this dataset. The paper underscores the pivotal role of innovation in promoting sustainable business operations, while also highlighting the need for further inquiry into the complex relationship between R&D spending and sustainable outcomes. The insights derived from this analysis offer meaningful guidance for those shaping policy and corporate strategy, particularly in the context of technological innovation as a driver of sustainability.

1. Introduction

The Sustainable Development Goals (SDGs) represent a universal call to action to end poverty, protect the planet, and ensure prosperity for all by 2030. Adopted by the United Nations in 2015, these 17 interconnected goals address various global challenges, ranging from poverty eradication to climate action. The SDGs represent a paradigm shift in the approach to development, emphasizing the interconnectedness of social, economic, and environmental dimensions. They provide a comprehensive framework for countries to address pressing challenges while fostering sustainable development pathways. By integrating the principles of inclusivity, equity, and environmental stewardship, the SDGs aim to leave no one behind and ensure a more equitable and sustainable future for all.

Researchers and institutions worldwide have dedicated significant efforts to assessing the progress and impact of the SDGs. Various methodologies, indicators, and data sources are utilized to monitor and evaluate the implementation of the goals at national, regional, and global levels (Severo & De Guimarães, 2022). These assessments provide valuable insights into areas of success, gaps, and areas needing urgent attention.

The role of patents in promoting innovation is twofold: they protect the inventor's interests, thereby encouraging more investment in R&D, and they disseminate knowledge through detailed patent documents, which can spur further innovation in the industry (Mairesse & Mohnen, 2004; Yang et al., 2018; Bansal & Roth, 2000). However, some scholars caution about the quality of patents, noting that not all patents contribute equally to economic growth, and the surge in patent filings might reflect strategic behaviour rather than genuine innovation (Pollák & Markovič, 2021; Chakraborty & Chatterjee, 2017).

The interaction between R&D expenditures and patent applications is significant, as evidenced by numerous studies (Severo & De Guimarães, 2021). For instance, Carballo-Penela & Castromán-Diz (2015) found a complementary relationship where increases in R&D spending led to higher rates of patenting among European firms, suggesting a cyclical boost to innovation. Moreover, the strategic use of patents can enhance the returns on R&D investments by securing proprietary technologies, thus fostering a competitive edge and higher market value (Hsu & Ziedonis, 2013; Pollák, Dorčák, & Markovič, 2021).

In addition to the 2030 Agenda, initiatives to increase the sustainability of business are emerging directly from within business. The Sustainable Brand Index (SBI) is the result of such an initiative. The Sustainable Brand Index aims to promote sustainable branding, illustrate its value, and enhance understanding of sustainability in branding and communication. Using data-driven insights, it highlights discrepancies between brands' self-perceptions and actual consumer perceptions regarding sustainability. By analysing trends, mapping stakeholder attitudes and behaviours, and evaluating various materiality areas, the SBI offers brand-specific data and strategic tools. This approach motivates brands to improve their sustainability efforts and encourages them to communicate these efforts. Increased dialogue about sustainability raises consumer awareness, concern, and demand, fostering a positive cycle of transparency and sustainability.

The SBI, grounded in macro and micro trends, consumer behaviours, and brand analysis, aims to answer the following questions:

- How does sustainability impact brands?
- How are brands perceived across various sustainability dimensions, and why?
- What strategies can brands employ to alter and enhance these perceptions?
- What are consumers' attitudes and behaviours regarding sustainability, how do they evolve over time, and how do they influence brands?
- Which future trends and developments in sustainability should brands be aware of and respond to?
- Is there a correlation between a brand's perceived environmental and social sustainability efforts and overall brand preference?

2. Data and Methodology

The purpose of this paper it to examine the relationship between the Sustainable Brand Index and two predictors: Research & Development (R&D) Expenditure and Patent Applications in Denmark and Sweden.

The data consists of annual values from 2017 to 2022, capturing the SBI, R&D expenditure (as the percentage of GDP), and the number of patent applications (per million inhabitants). The data source used for this analysis was the Eurostat database (Eurostat, 2023; Eurostat; 2024) and the Sustainable Brand database (SBI, 2024).

We formulated the following four hypotheses:

- H1: There is a relationship between R&D expenditure and the SBI score in Denmark.
- H2: There is a relationship between Patent applications and the SBI score in Denmark.
- H3: There is a relationship between R&D expenditure and the SBI score in Sweden.
- H4: There is a relationship between Patent applications and the SBI score in Sweden.

A linear regression model was specified with SBI as the dependent variable. R&D Expenditure and Patent Applications were included as independent variables. Ordinary Least Squares (OLS) regression was utilized to estimate the relationships between SBI and the independent variables. R-squared and Adjusted R-squared values were computed to assess how well the model explained the variability in SBI. The F-statistic and its associated p-value were calculated to test the overall significance of the regression model. Durbin-Watson statistic was used to check for the presence of autocorrelation in the residuals. The Condition Number was examined to assess potential multicollinearity among predictors.

3. Results

R&D expenditure is a key indicator of a country's commitment to innovation and technological advancement. Denmark’s R&D expenditure is around 3.0% of its GDP, reflecting its strong emphasis on innovation, especially in renewable energy and sustainable solutions. Sweden is a leader in R&D investment, with spending exceeding 3.5% of its GDP. The country is renowned for its innovation in telecommunications, automotive, and environmental technologies.

The level of technological innovation is measured by the number of patents filed, high-tech exports, and the overall innovation index. Denmark is known for its innovative approach to technology, especially in green and sustainable technologies. It has a high rate of patent filings and a strong focus on clean technologies. Sweden stands out for its technological innovation, leading in the number of patents per capita and excelling in high-tech and green technology sectors.

3.1 Results of the regression analysis for Denmark

In the subsequent subchapter, the results of the regression analysis performed on the Danish data from 2017 to 2022 are presented.

Table 1. Results of regression analysis - Denmark

Model fit

Description	Value
R-squared	0.936
Adjusted R-squared	0.893
F-statistic	21.85
Prob (F-statistic)	0.0163
No. Observations	6
DF Residuals	3
DF Model	2

Coefficients

Variable	Coefficient	Standard Error	t-Value	P-value	95% CI Lower	95% CI Upper
Intercept	4.96	41.687	0.120	0.912	-127.677	137.656
R&D Expenditure	-9.6002	11.696	-0.821	0.472	-46.822	27.622
Patent applications	0.1361	0.028	4.935	0.016	0.048	0.224

Source: (own calculations)

The R-squared value suggests that the model explains approximately 93.6% of the variance in the Sustainable Brand Index (Table 1). The F-statistic indicates that the model is statistically significant overall. Patent applications have a statistically significant positive impact on the Sustainable Brand Index, with each unit increase associated with a 0.1361 increase in the index. The impact of R&D Expenditure is negative and not statistically significant, suggesting no reliable effect on the Sustainable Brand Index within the scope of this data.

3.2 Results of the regression analysis for Sweden

Table 2 presents the outcomes of the regression analysis conducted on the Swedish data spanning from 2017 to 2022. The results highlight that while both Denmark and Sweden show a positive correlation between patent filings and the Sustainable Brand Index (SBI), the relationship between R&D investment and SBI varies between the two countries. Specifically, Denmark exhibits a negative but statistically insignificant relationship between R&D expenditure and SBI, suggesting that R&D investment may not directly translate to immediate sustainability performance. In contrast, Sweden shows a significant negative correlation between R&D investment and SBI, implying that higher R&D spending might be linked to short-term decreases in sustainability performance, possibly due to the lag effect of research investments. This discrepancy warrants further investigation to understand the underlying factors driving these differences in R&D impact on sustainability perception.

Table 2. Results of the regression analysis - Sweden

Model fit

Description	Value
R-squared	0.954
Adjusted R-squared	0.923
F-statistic	31.15
Prob (F-statistic)	0.00985
No. Observations	6
DF Residuals	3
DF Model	2

Coefficients

Variable	Coefficient	Std. Error	t-value	P-value	95% CI Lower	95% CI Upper
Constant	147.5101	40.209	3.669	0.035	19.548	275.472
R&D Expenditure	-49.3621	12.513	-3.945	0.029	-89.184	-9.540
Patent applications	0.1342	0.017	7.827	0.004	0.080	0.189

Source: (own calculations)

The R-squared value of 0.954 indicates that approximately 95.4% of the variability in the SBI scores is explained by the model, a very high level of explanatory power (Table 2). The Adjusted R-squared of 0.923 further confirms a good fit even after adjusting for the number of predictors. The model suggests a strong negative association between R&D expenditure and the SBI score, which is statistically significant (p-value = 0.029). This implies that increases in R&D expenditure are associated with decreases in the SBI score. Patent applications have a positive and statistically significant effect on the SBI score (p-value = 0.004), suggesting that increases in patent applications are associated with increases in the SBI score. Both predictors show statistical significance, supporting the observed relationships in the model. This analysis provides robust evidence for the relationships between these variables and the SBI score in Sweden, with significant implications for understanding how these factors influence the observed index.

4. Conclusion

In conclusion, the Sustainable Development Goals represent a bold and transformative agenda for a more equitable, inclusive, and sustainable world. While challenges abound, the collective resolve of nations, organizations, and individuals can pave the way for a future where no one is left behind, and the planet thrives for generations to come. Continued research, collaboration, and action are essential to realizing the vision of the 2030 Agenda and building a better world for all.

Sustainable Brand Index, established in 2011 by SB Insight in Sweden, stands as Europe's foremost independent examination of brand sustainability. This comprehensive methodology evaluates the sustainability perceptions of nearly 1,600 brands spanning 36 industries and based on 80,000 consumer interviews across Europe, encompassing the Nordics, the Netherlands, and the Baltics. Through this rigorous analysis, Sustainable Brand Index offers insights into how brands are viewed in terms of sustainability, the factors influencing these perceptions, and actionable strategies for improvement.

The regression analyses conducted on data from Denmark and Sweden from 2017 to 2022 provided valuable insights into the factors influencing the Sustainable Brand Index (SBI) in these countries. The results highlight distinct economic impacts in both contexts, reflecting different responses to R&D expenditure and patent application rates. The regression analysis for Denmark revealed a statistically significant relationship between R&D expenditure and the SBI, suggesting that investments in research and development positively influence the business environment. This underscores the role of innovation-driven activities in enhancing economic indices. Patent applications, however, did not show a statistically significant impact on the SBI, indicating that while Denmark might be strong in producing new inventions, the immediate impact on the broader brand index may be limited or influenced by other factors not captured in this model.

In contrast, the regression results for Sweden showed both R&D expenditure and patent applications as significant predictors of the SBI. This suggests a robust correlation where both research activities and the output of intellectual property contribute positively to the business climate. The strong negative coefficient for R&D expenditure, which was initially counterintuitive, could imply that higher expenses might be associated with short-term decreases in the business index, possibly due to the lag effect of research investments. Meanwhile, the positive impact of patent applications reflects a direct benefit from innovation to the business sector.

These analyses highlight the critical role of innovation and intellectual property in shaping economic indices in Denmark and Sweden, though the mechanisms and immediate effects may vary between the two countries. The findings emphasize the importance of tailored economic policies that foster R&D activities and support the translation of innovative outputs into marketable and business-enhancing ventures. Future research could benefit from incorporating additional variables such as industry-specific effects, government policy impacts, and international economic conditions to further refine the understanding of these relationships.

Acknowledgement

This research paper is a partial result of the multiyear research projects VEGA 1/0462/23 “Circular economy in the context of societal demands and market constraints” and VEGA 1/0465/23 “Generic, convergence and model-based approaches of environmental production and logistics in business development on Slovakia”.

References

Bansal, P., & Roth, K. (2000). WHY COMPANIES GO GREEN: A MODEL OF ECOLOGICAL RESPONSIVENESS. Academy of Management Journal, 43(4), 717–736. https://doi.org/10.2307/1556363

Carballo-Penela, A., & Castromán-Diz, J. L. (2015). Environmental Policies for Sustainable Development: An Analysis of the Drivers of Proactive Environmental Strategies in the Service Sector. Business Strategy and the Environment, 24(8), 802–818. https://doi.org/10.1002/bse.1847

Chakraborty, P., & Chatterjee, C. (2017). Does environmental regulation indirectly induce upstream innovation? New evidence from India. Research Policy, 46(5), 939–955. https://doi.org/10.1016/j.respol.2017.03.004

Eurostat. (2023, 11 December). Gross domestic expenditure on R&D by sector. https://ec.europa.eu/eurostat/databrowser/view/sdg_09_10/default/table?lang=en&category=sdg.sdg_09

Eurostat. (2024, 21 March). Patent applications to the European Patent Office by applicants'/inventors' country of residence. https://ec.europa.eu/eurostat/databrowser/view/sdg_09_40/default/table?lang=en&category=sdg.sdg_09

Hsu, D. H., & Ziedonis, R. H. (2013). RESOURCES AS DUAL SOURCES OF ADVANTAGE: IMPLICATIONS FOR VALUING ENTREPRENEURIAL-FIRM PATENTS. Strategic Management Journal, 34(7), 761–781. http://www.jstor.org/stable/23471066

Mairesse, J., & Mohnen, P. (2004). The Importance of R&D for Innovation: A Reassessment Using French Survey Data. The Journal of Technology Transfer, 30(1–2), 183–197. https://doi.org/10.1007/s10961-004-4365-8

Pollák, F., Dorčák, P., & Markovič, P. (2021). Corporate Reputation of Family-Owned Businesses: Parent Companies vs. Their Brands. Information, 12(2), 89. https://doi.org/10.3390/info12020089

Pollák, F., & Markovič, P. (2021). Economic Activity as a Determinant for Customer Adoption of Social Media Marketing. Sustainability, 13(7), 3999. https://doi.org/10.3390/su13073999

SBI. (2024, 12 March). Sustainable Brand Index. https://www.sb-insight.com/sbi

Severo, E. A., & De Guimarães, J. C. F. (2021). Antecedent and Consequents of Eco-Innovation for Sustainability: Generations' Perceptions in Brazil and Portugal. International Journal of Professional Business Review, 7(1), e0280. https://doi.org/10.26668/businessreview/2022.v7i1.280

Severo, E. A., & De Guimarães, J. C. F. (2022). The Influence of Product Innovation, Environmental Strategy and Circular Economy On Sustainable Development in Organizations in Northeastern Brazil. Journal of Law and Sustainable Development, 10(2), e0223. https://doi.org/10.37497/sdgs.v10i2.223

Yang, X., Wang, Y., Hu, D., & Gao, Y. (2018). How industry peers improve your sustainable development? The role of listed firms in environmental strategies. Business Strategy and the Environment, 27(8), 1313–1333. https://doi.org/10.1002/bse.2181

ICT SYSTEMS AND BUSINESS

DIGTALIZATION AND ECONOMIC FREEDOM

Motto: In e we trust......really?

Petr Doucek, Lea Nedomova

Faculty of Informatics and Statistics
Prague University of Economics and Business
doucek@vse.cz, nedomova@vse.cz

DOI: 10.35011/IDIMT-2024-213

Keywords

Digitalization, Digital Economy and Society Index, Economic Freedom Index, ICT in public administration, digital public services, digital competencies

Abstract

In the minds of the average European citizen, and especially young citizens, the digitisation of government, business and everyday life - that is, of society as a whole - is linked to the growth of convenience, comfort and well-being. But is this really the case? In our paper, we compare the growth in the pace of digitalisation, as expressed by the DESI (Digital Economy and Society Index) (or its individual components) with changes in the EFI (Economic Freedom Index) (or its individual components). Our research for selected Central European countries is based on data for the period 2017-2022. The data source for the DESI index is the European Union database, and for the EFI is The Heritage Foundation database. Among the methods, we mainly used correlation analysis (Spearman's correlation coefficient) and data analysis using MS Excel tools and functions. Overall, the measurements show that the countries of the Middle East can be divided into three categories. The first category is countries where there is a strong correlation between DESI and EFI results. These countries are Austria, Slovak Republic and Slovenia. The second category is countries with a medium correlation - Poland, Czech Republic and Germany. The lowest correlation between DESI and EFI is identified in Hungary. More detailed causal analyses are presented in section 3.2 of the paper.

1. Introduction

With great expectations, the developed countries of the world are resorting to the increasing digitisation of most activities in government and public administration, in business and in the personal lives of their citizens. They are increasingly forced to cope with new and new applications that, in addition to new features, bring with them new bugs and previously unknown problems. Digitalisation in the current world view can be considered, according to (Gartners, 2024) "... the process of changing from analogue to digital form, also known as digital enablement. Said another way, **digitization** takes an analog process and changes it to a digital form ***without any different-in-kind changes to the process itself.***" And it is the second half of the definition of digitalization that represents a rather large setback in its construction and promotion. That is, simply converting existing processes from analogue to digital form is a fundamental impediment to the process. As the way of working with data changes, the logic of data processing must also change. But with it, the mindset of the people who use such technology must also change. Application programmers also have a big task in digitising government and its processes. The reliability and confidentiality of processed requests and data must be the basis of all digital services for citizens (Bokša et al, 2019). This is directly linked to proving the identity of an entity (citizen, service, device or process) and also to the confidentiality of the processed requests and subsequently the outputs. A key issue is the trustworthy identification of the entity. Central registers of state and public administration are used for this purpose. However, with the increasing number of applications and functions available to citizens, the burden on these registers - or rather on their availability and ability to respond quickly to the increasing number of queries to them - is also growing (Doucek, Hološka, Nedomová, 2022). Thus, technologically, there is a very close link to the proposed architecture of the entire e-government solution and its implementation by technical means. This is one dimension of digitalization - the **need to rethink processes**, not just their simple conversion from analogue to digital form.

Another dimension is the question of how the digitalisation of society and its processes affects citizens' behaviour and their feelings and satisfaction (Doucek, Hološka, 2019). Today's society measures these phenomena with various indices. The actual digitisation of society is measured by the DESI (Digital Economy and Society Index) (DESI, 2024), and the digitisation of companies and business activities by the CDI (Corporate Digitisation Index) (European Investment Bank, 2023). Their relationship is discussed in more detail in (Doucek, Nedomova, 2023). Another index that is the focus of this paper is the EFI (Economic Freedom Index) (EFI, 2024). The analysis of its components and its significance for the economy are dealt with in more detail in (Vrabec; Marek, 2025) in the conditions of the Czech Republic, (Karaköy, Ulutas, Karabasevic, Üre, 2023) in the conditions of the global economy. By comparing these two indices we get an idea of the impact of digitalization and the components of the DESI index on the values of the EFI index as a whole and also on its individual components. Thus, the impact of digitalization on the freedom of the economic environment. This freedom of the economic environment can also be interpreted as the degree of market environment in the country under study.

The aim of this paper is to present research on the relationship between the digitalisation of society, as represented by the Digital Economy and Society Index, and economic freedom, as expressed by the Economic Freedom Index. For the purpose of this paper, we formulated the following two research questions for selected Central European countries:

RQ1: Is the Economic Freedom Index dependent on the development of the Digital Economy and Society Index - i.e. does the digitalisation of society contribute to the growth of economic freedom?

RQ2: Which components of the DESI have the greatest impact on the economic freedom of selected Central European countries?

2. Methodology and Data Collection

To track the evolution of digitalisation and its impact on economic freedom in selected countries, we used a comparison of the Digital Economy and Society Index and the Economic Freedom Index.

The Digital Economy and Society Index consists of four components - Connection (desi_conn), Human Capital (desi_hc), Digital Public Services (desi_dps) and Integration of Digital Technology (desi_idt). For the purposes of the analysis in this paper, all components are weighted equally to calculate the total value of the annual DESI. The Economic Freedom Index consists of the following eleven components - Property Rights (PR), Judicial Effectiveness (JE), Government Integrity (GI), Tax Burden (TB), Governmental Spending (GS), Fiscal Health (FH), Business Freedom (BF), Monetary Freedom (MF), Trade Freedom (TF), Investment Freedom (IF) and Financial Freedom (FF). Their detailed analysis can be found, for example, in (Vrabec; Marek, 2025), (Karaköy, Ulutas, Karabasevic, Üre, 2023).

In our paper we analysed the period 2017-2022. The original intention was to analyse a longer time series, but DESI data for 2023 and for years prior to 2017 were not available at the time of writing. We used the Pearson correlation coefficient to determine the correlation between the variables under study. The prerequisites for its use were met in the area of assumed linear dependence, the variables were independent of each other. The condition of homoskedasticity was not met for some components of the EFI index. These were mainly the variables Financial Freedom (FF) and Investment Freedom (IF), which showed zero variance, and therefore we did not calculate the value of the correlation coefficient for them - in the tables it is marked with the symbol N/A.

The selected Central European countries for the analysis of the impact of changes in the Digital Economy and Society Index on the EFI are listed in alphabetical order as follows: Austria, Czech Republic, Germany, Hungary, Poland, Slovak Republic and Slovenia.

To answer RQ1 (section 3.1), we analysed the aggregate values of the two indices over the defined period and compared the two time series with each other using Pearson's correlation coefficient. The answer to RQ2 (section 3.2) was then much more complex. For it, we had to analyse the relationships between the individual components of the two indices under study. In this paper, due to space limitations, only selected outputs from the correlation analysis performed are presented.

For the final projection of the evaluation of all countries and the impact of the DESI index on the EFI in them, we used the following scores presented in Table 1.

Table 1. Correlation strength scores between DESI and EFI components

Point Value	Correlation Coefficient Value
3	1 - 0,7
2	0,69 - 0,4
1	0,39 - 0,1
0	0,09 - - 0,09
-1	-0,1 - -0,39
-2	-0,4 - -0,69
-3	-0,7 - -1

Source: (Authors)

3. Results and Discussion

To perform the analyses below, we used available data from the (DESI, 2024) and (EFI, 2024) databases as of mid-January 2024 for DESI on January 13, 2024 and January 12, 2024 for EFI.

3.1 Relationship between DESI and EFI basic analysis

The basic, baseline analysis was performed over the aggregate annual values of the two indices under study. The main objective was to answer RQ1 and to identify the specificities of each country under study.

RQ1: Is the Economic Freedom Index dependent on the development of the Digital Economy and Society Index - i.e. does the digitalisation of society contribute to the growth of economic freedom?

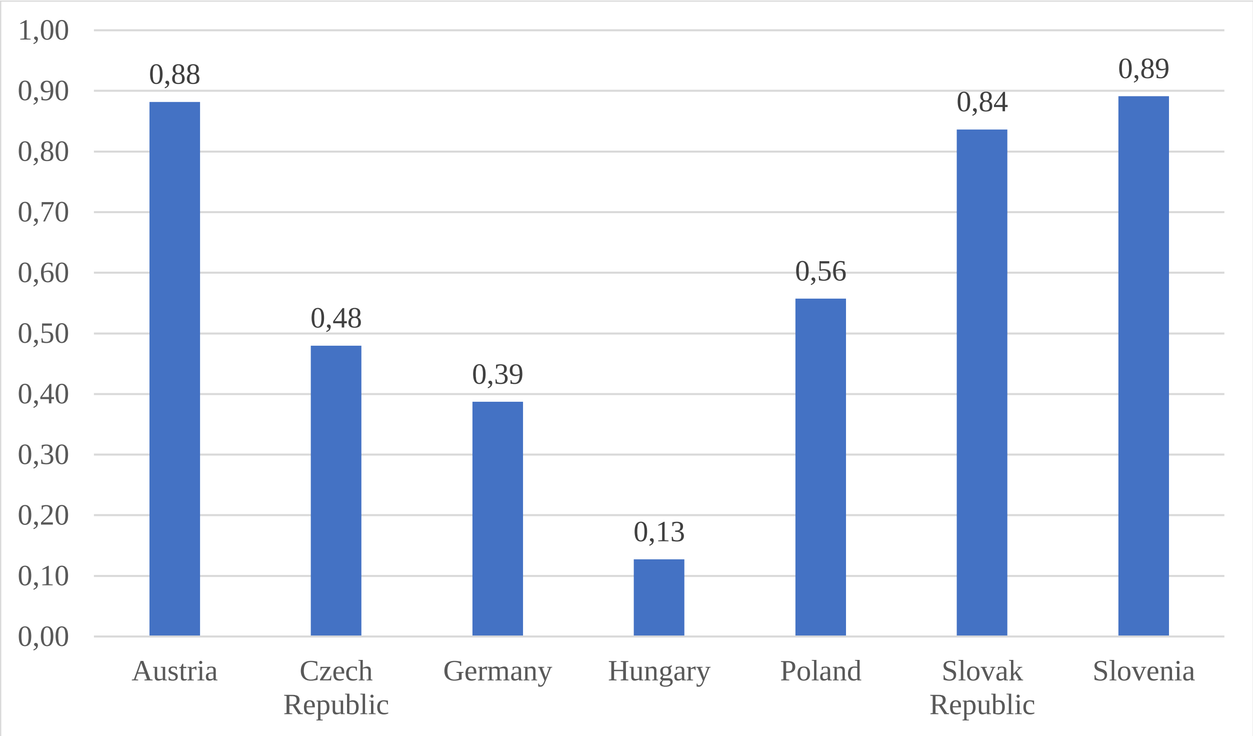


Figure 1. Comparison of correlation between DESI and EFI in 2017-2022 - aggregate values Source: own elaboration, Data: (DESI, 2024; EFI, 2024)

Note on Figure 1 - the vertical axis of the graph shows the calculated value of the correlation coefficient DESI and EFI

Figure 1 shows that there are three groups of countries in Central Europe in terms of the impact of digitalization on economic freedom. The first group are countries where the correlation between digitisation and economic freedom is very strong. These are Austria, Slovak Republic and Slovenia. For them, the correlation coefficient is higher than 0.8, indicating a very strong correlation in the period under review. The second group is a group of three countries where the observed dependence is average. Of these countries, Poland has the best results, where the correlation coefficient is above 0.5 and reaches 0.56. The other two countries in this group, the Czech Republic and Germany, have observed values below 0.5, the Czech Republic very comfortably at 0.48 and Germany at 0.39. We find this result of the analysis very surprising. However, when we analyse the annual values in more detail, we find that the DESI values are increasing over the period under review, but the EFI values are the same or even decreasing year-on-year (e.g. 2019 and 2021). A more detailed analysis is presented in section 3.2. The third group of countries includes Hungary, where only a very weak dependence between the observed phenomena was identified, with a correlation coefficient value of 0.13.

3.2 Detailed analysis of the relationship between DESI and EFI

Based on the analysis of the annual DESI and EFI values, we have classified the countries studied into three categories - strong impact of digitalization on economic freedom, weak impact and practically negligible impact. The research that was directed towards answering RQ2 then stems from these findings.

RQ2: Which components of the DESI have the greatest impact on the economic freedom of selected Central European countries?

For **the first group of countries (Austria, Slovak Republic and Slovenia)**, where the correlation between DESI and EFI is the highest, it is typical that digitalization has the most positive impact on EFI components such as Property Rights (PR), Judicial Effectiveness (JE), Government Integrity (GI) and Labour Freedom (LF) (Table1. - Table. 3.). On the other hand, components such as Tax Burden (TB), Monetary Freedom (MF) and Trade Freedom (TF) are negatively affected by digitalisation. In summary, it can be presented that the Connectivity component of the DESI has the greatest impact on the components of the EFI, both positive and negative. By this we mean that when the correlation is positive, the Connectivity component has the highest correlation; likewise, when the correlation is negative, again the contribution of the Connectivity component is the strongest. The exception is the Monetary Freedom component, where the logic is exactly the opposite.

The analysis of the relationship between DESI and EFI for Austria (Table 2) clearly shows a positive relationship of digitalization with the parameters Property Rights (PR), Judicial Effectiveness (JE), Government Integrity (GI). Surprisingly, a very strong positive correlation is found for the parameter Labour Freedom (LF). A weaker correlation is found for the component Business Freedom (BF) and a very weak correlation is found for the component Governmental Spending (GS).

On the contrary, the significantly negative correlation is particularly evident for the Tax Burden (TB) parameters, which we find strange, as we would intuitively expect a more positive correlation. The other two components of the EFI that have a relatively strong negative correlation are Monetary Freedom (MF) and Trade Freedom (TF). Here, on the other hand, we would intuitively expect a strong negative correlation.

Table 2. Correlation between ESI and EFI components - Austria

Austria	PR	JE	GI	TB	GS	FH	BF	LF	MF	TF	IF	FF
desi_conn	0,817	0,730	0,730	-0,928	0,242	-0,293	0,446	0,805	-0,316	-0,942	-0,762	N/A
desi_dps	0,734	0,484	0,873	-0,758	0,468	-0,081	0,241	0,718	-0,585	-0,841	-0,648	N/A
desi_hc	0,665	0,441	0,689	-0,710	0,350	-0,140	0,338	0,678	-0,506	-0,760	-0,568	N/A
desi_idt	0,776	0,544	0,859	-0,776	0,410	-0,143	0,295	0,750	-0,530	-0,866	-0,691	N/A

Source: authors, data:(DESI,2024; EFI,2024)

The data analysis for the Slovak Republic (Table 3) is very similar to that for Austria. Significantly positive DESI correlations on the components Property Rights (PR), Judicial Effectiveness (JE), Government Integrity (GI), medium positive correlations on Labour Freedom (LF) and Business Freedom (BF).

There is a mixed effect of the DESI components on the EFI component Fiscal Health (FH). While in Austria this correlation is slightly negative, in Slovakia two DESI components are weakly positively correlated with the EFI component Fiscal Health (FH) (desi_dps - Digital Public Services and desi_hc

- Human Capital), while the other two are weakly to very weakly negatively correlated (desi_conn – Connection and desi-idt - Integration of Digital Technology).

A very strong negative correlation is found for the components Tax Burden (TB), Monetary Freedom (MF) and Trade Freedom (TF), while a weaker negative correlation is found with the component Governmental Spending (GS).

Table 3. Correlation between ESI and EFI components - Slovak Republic

Slovak Republic	PR	JE	GI	TB	GS	FH	BF	LF	MF	TF	IF	FF
desi_conn	0,905	0,913	0,883	-0,880	-0,483	-0,312	0,439	0,419	-0,800	-0,959	N/A	N/A
desi_dps	0,784	0,740	0,898	-0,910	-0,220	0,023	0,117	0,100	-0,955	-0,814	N/A	N/A
desi_hc	0,773	0,728	0,903	-0,900	-0,206	0,046	0,095	0,078	-0,957	-0,800	N/A	N/A
desi_idt	0,777	0,758	0,842	-0,949	-0,335	-0,035	0,179	0,151	-0,903	-0,832	N/A	N/A

Source: authors, data:(DESI,2024; EFI,2024)

The analysis of the data for Slovenia (Table 4) provides a very similar view of the dependencies between the components of the two indices as in previous surveys. However, the exception is the Business Freedom (BF) component, which in the previous two analyses was slightly positively correlated with the adoption of digitalisation, whereas for Slovenia a weak negative correlation was found.

Table 4. Correlation between ESI and EFI components - Slovenia

Slovenia	PR	JE	GI	TB	GS	FH	BF	LF	MF	TF	IF	FF
desi_conn	0,890	0,818	0,726	-0,871	0,402	0,447	-0,157	0,896	-0,233	-0,978	N/A	N/A
desi_dps	0,752	0,619	0,849	-0,690	0,671	0,700	-0,479	0,869	-0,449	-0,864	N/A	N/A
desi_hc	0,708	0,530	0,875	-0,597	0,751	0,716	-0,538	0,827	-0,563	-0,811	N/A	N/A
desi_idt	0,786	0,637	0,844	-0,668	0,665	0,711	-0,492	0,904	-0,419	-0,866	N/A	N/A

Source: authors, data:(DESI,2024; EFI,2024)

The second group is represented by the countries **Czech Republic, Germany and Poland**. The results of the analysis for these countries are similar in main features to the first group. In particular, there is a strong positive correlation on the components Property Rights (PR), Judicial Effectiveness (JE), Government Integrity (GI) and also a very strong negative correlation on the components Tax Burden (TB), Monetary Freedom (MF) and Trade Freedom (TF).

The analysis for the Czech Republic (Table 5.) is interesting in that a very strong positive correlation was found for the DESI component on Property Rights (PR), Judicial Effectiveness (JE) and Business Freedom (BF). A medium positive correlation was found for the Government Integrity (GI) component. The mixed effects, especially the negative correlations of the Connectivity (desi_conn) component, are surprising for the Fiscal Health (FH) and Governmental Spending (GS) components.

Significant negative correlation was identified for the EFI components Tax Burden (TB), Labour Freedom (LF), Monetary Freedom (MF), Trade Freedom (TF) and Investment Freedom (IF).

Table 5. Correlation between ESI and EFI components - Czech Republic

Czech Republic	PR	JE	GI	TB	GS	FH	BF	LF	MF	TF	IF	FF
desi_conn	0,961	0,820	0,435	-0,904	-0,285	-0,118	0,791	-0,886	-0,822	-0,986	-0,869	N/A
desi_dps	0,892	0,601	0,607	-0,904	0,053	0,215	0,645	-0,695	-0,948	-0,865	-0,831	N/A
desi_hc	0,860	0,508	0,648	-0,873	0,126	0,244	0,579	-0,638	-0,985	-0,826	-0,786	N/A
desi_idt	0,813	0,513	0,774	-0,907	0,115	0,244	0,476	-0,594	-0,946	-0,787	-0,820	N/A

Source: authors, data:(DESI,2024; EFI,2024)

The analysis in Table 6 refers to Germany. It is interesting in that a very strong positive correlation of the DESI component to Government Integrity (GI) was found. A moderately strong positive correlation was found to the Property Rights (PR) and Judicial Effectiveness (JE) components.

The mixed effects, especially the positive correlations of the Connectivity (desi_conn) component are surprising for the Business Freedom (BF) component.

Significant negative correlation was identified for the EFI components Tax Burden (TB), Governmental Spending (GS), Monetary Freedom (MF) and Trade Freedom (TF).

Table 6. Correlation between ESI and EFI components - Germany

Germany	PR	JE	GI	TB	GS	FH	BF	LF	MF	TF	IF	FF
desi_conn	0,690	0,532	0,899	-0,914	-0,856	0,133	0,053	0,381	-0,547	-0,951	N/A	N/A
desi_dps	0,442	0,264	0,833	-0,955	-0,657	0,439	-0,253	0,592	-0,739	-0,820	N/A	N/A
desi_hc	0,603	0,450	0,958	-0,937	-0,716	0,312	-0,128	0,503	-0,738	-0,863	N/A	N/A
desi_idt	0,573	0,414	0,893	-0,976	-0,750	0,311	-0,121	0,568	-0,685	-0,892	N/A	N/A

Source: authors, data:(DESI,2024; EFI,2024)

Table 7 shows the results of the correlations for Poland. Strong positive correlations are evident for the Property Rights (PR), Government Integrity (GI) and, surprisingly, Investment Freedom (IF) components.

Significant negative correlation was identified for the EFI components Judicial Effectiveness (JE), Tax Burden (TB), Governmental Spending (GS), Labour Freedom (LF), Monetary Freedom (MF) and Trade Freedom (TF).

Table 7. Correlation between ESI and EFI components - Poland

Poland	PR	JE	GI	TB	GS	FH	BF	LF	MF	TF	IF	FF
desi_conn	0,874	-0,171	0,672	-0,961	-0,635	0,243	0,439	-0,494	-0,952	-0,913	0,737	N/A
desi_dps	0,814	-0,305	0,668	-0,981	-0,518	0,374	0,327	-0,387	-0,961	-0,856	0,814	N/A
desi_hc	0,788	-0,279	0,676	-0,990	-0,508	0,383	0,301	-0,340	-0,963	-0,858	0,806	N/A
desi_idt	0,807	-0,340	0,669	-0,978	-0,498	0,391	0,313	-0,386	-0,961	-0,840	0,831	N/A

Source: authors, data:(DESI,2024; EFI,2024)

The third group is **Hungary** (Table 8), where the identified correlation between the aggregated DESI and EFI values is the lowest. This country is characterized by a clear distribution of the influence of DESI on EFI, with a strong to very strong correlation. The EFI components Property Rights (PR), Judicial Efficiency (JE), Government Integrity (GI), Tax Burden (TB), Governmental Spending (GS), Business Freedom (BF) and Investment Freedom (IF) are positively affected by digitalization. On the other hand, the EFI parameters Fiscal Health (FH), Labour Freedom (LF), Monetary Freedom (MF) and Trade Freedom (TF) show a negative correlation, and a very strong one at that.

Table 8. Correlation between DESI and EFI components - Hungary

Hungary	**PR**	**JE**	**GI**	**TB**	**GS**	**FH**	**BF**	**LF**	**MF**	**TF**	**IF**	**FF**
desi_conn	0,958	0,442	0,642	0,894	0,623	-0,699	0,633	-0,711	-0,829	-0,905	0,516	N/A
desi_dps	0,910	0,380	0,578	0,883	0,704	-0,613	0,551	-0,659	-0,882	-0,888	0,616	N/A
desi_hc	0,819	0,350	0,776	0,975	0,603	-0,494	0,410	-0,649	-0,781	-0,788	0,326	N/A
desi_idt	0,933	0,487	0,513	0,877	0,620	-0,704	0,649	-0,661	-0,818	-0,932	0,597	N/A

Source: authors, data:(DESI,2024; EFI,2024)

4. Conclusions

The paper offers a reflection on the impact of the digitalisation of the economy and public and state administration, represented by the DESI index, and economic freedom - the de facto quality of the market environment, represented by the EFI index. The components of the DESI index represent different components of the digitalization of a society (state), while the components of the EFI index show the degree of economic (market) freedom in that state. Let's see what the answers to our research questions look like.

RQ1: Is the Economic Freedom Index dependent on the development of the Digital Economy and Society Index - i.e. does the digitalisation of society contribute to the growth of economic freedom?

The basic answers can be found in Table 9. Some conclusions surprised us a lot. The first of them concerns the influence of digitization on the growth of the centralization of the economy - this can be expressed mainly by the indicators Property Rights (PR), Judicial Effectiveness (JE) and Government Integrity (GI). Here it is clearly seen that even in states where digitization is not very developed, it significantly strengthens the functions of the central state, both in a good and a bad sense. Significantly negative correlations between digitization and Monetary Freedom (MF) and Trade Freedom (TF), on the contrary, present that digitalization has a negative effect. Their statistically established relationship can be represented as that economic freedom decreases with increasing digitization. Thus, in particular, the freedom of the labour market and the freedom of the market environment are decreasing. An interesting situation is with the **Tax Burden (TB)** component. The correlation between digitization and this EFI component is negative for practically all monitored countries except Hungary. Let's start with the definition - Tax Burden - is a measure of the tax burden imposed by government. It includes direct taxes, in terms of the top marginal tax rates on individual and corporate incomes, and overall taxes, including all forms of direct and indirect taxation at all levels of government, as a percentage of GDP (EFI, 2024). This can be interpreted as the fact that the tax burden decreases with increasing digitization. Based on my experience with practical functioning, we would rather expect that digitization will enable better and more consistent collection of taxes, for example by means of applications for mandatory digital records of sales, seamless communication during payment between companies and between economic entities and the state. This will make tax collection more efficient and overall the tax burden will increase. This is also due to the fact that increasing digitization means higher investment costs, computer technology depreciation is increasing, and the number of specialized IT department workers in the back offices of companies and public and state administration organizations is also increasing.

From the point of view of the DESI index, the component that has the greatest influence on the components of the EFI index – that is, on the freedom of the economy, is the Connection component (desi_conn). Its relatively large influence on the economic environment can be observed in practically all EFI indicators, with the exception of Fiscal Health (FH) and Labour Freedom (LF). The DESI Digital Public Services (desi_dps) component has the greatest impact on these two indicators. In this of the article, we do not distinguish positive or negative correlation, but the level of its projection according to Table 1.

Answer to **RQ2: Which components of the DESI have the greatest impact on the economic freedom of selected Central European countries?** - The overall impact of digitalisation on market freedom in the selected Central European countries studied can be seen also in Table 9 below.

Table 9. Relationship between DESI and EFI components - Projections for all countries analysed

Projection	**PR**	**JE**	**GI**	**TB**	**GS**	**FH**	**BF**	**LF**	**MF**	**TF**	**IF**	**FF**
desi_conn	**3,000**	2,143	**2,571**	-2,143	-0,429	*-0,286*	1,429	0,143	**-2,286**	**-3,000**	N/A	N/A
desi_dps	**2,857**	1,429	**2,571**	-1,429	0,429	*0,714*	0,714	0,429	**-2,714**	**-3,000**	N/A	N/A
desi_hc	**2,714**	1,571	**2,571**	-1,429	0,143	*0,429*	0,429	0,429	**-2,714**	**-3,000**	N/A	N/A
desi_idt	**2,857**	1,714	**2,714**	-1,429	0,143	*0,286*	0,571	0,571	**-2,571**	**-3,000**	N/A	N/A

Source: authors, data:(DESI,2024; EFI,2024)

A comparison of the correlations in Table 9 across all three groups shows that increasing digitization, and especially the Connectivity component, has a very strong positive effect on Property Rights (PR) and Government Integrity (GI). In addition, digitization also has a positive effect on the Judicial Effectiveness (JE) component. On the other hand, digitalization shows a very strong negative effect on the Monetary Freedom (MF) and Trade Freedom (TF) components. Digitalization also shows a stronger negative effect on the Tax Burden (TB) component.

The effects of digitalization on the other components of the EFI are not crutial and we came to the conclusions that they are almost weak, both positively and negatively across all countries surveyed in the research.

Acknowledgement

Paper was processed with support from institutional-support fund for long-term conceptual development of science and research at the Faculty of Informatics and Statistics of the Prague University of Economics and Business (IP400040) and the project IG 409033.

References

Bokša, M., Bokšová, J., Horák, J., Pavlica, K., Strouhal, J., & Šaroch, S. (2019). Digitální Česko v digitální Evropě. Škoda Auto Vysoká škola o.p.s. Available at: https://vlada.gov.cz/assets/evropske-zalezitosti/aktualne/Digitalni_Cesko_FINAL-ONLINE-VERSION.pdf

Doucek, P., & Hološka, J. (2019). Digital economy and industry 4.0. In: IDIMT-2019 Innovation and Transformation in a Digital World. Linz: Trauner Verlag Universität, pp. 33–40. Available at: https://idimt.org/wp-content/uploads/2019/08/IDIMT-2019-proceedings.pdf

Doucek, P., Hološka, J., & Nedomová, L. (2022). Management and digitalization. In: IDIMT-2022 Digitalization of Society, Business and Management in a Pandemic. Linz: Trauner Verlag Universität, pp. 35–42. DOI: 10.35011/IDIMT-2022-35.

Doucek, P., & Nedomova, L. (2023). Digital services and public administration. In: IDIMT-2023, New challenges for ICT and management. Linz: Trauner Verlag Universität, pp. 179–186. DOI: 10.35011/IDIMT-2023-179.

DESI. (2024). Shaping Europe's digital future. The Digital Economy and Society Index (DESI). Available at: https://digital-strategy.ec.europa.eu/cs/policies/desi

EFI. (2024). Index of Economic Freedom. Available at: https://www.heritage.org/index/explore?view=by-region-country-year&countryids=4&regionids=2&yearids=29

European Investment Bank. (2023). Digtalization in Europe 2021 – 2022: Evidence form the EIB Investment Survey. European Investment Bank. DOI:10.2867/76258

Gartners. (2024). Information Technology – Gartner Glossary – Digitization. Available at: https://www.gartner.com/en/information-technology/glossary/digitization

Karaköy, C., Ulutas, A., Karabasevic, D., Üre, S., & Bayrakçil, A.O. (2023). The Evaluation of Economic Freedom Indexes of EU Countries with a Grey Hybrid Mcdm Model. Journal for Economic Forecasting, 26 (1), 129-144. Available at: https://ideas.repec.org/a/rjr/romjef/vy2023i1p129-144.html

Vrabec, M., & Marek, L. (2015). Analysis of the index of economic freedom using structural models. In: Mathematical Methods in Economics 2015 (MME). Plzeň: University of West Bohemia, pp. 900–905.

SUPPORT OF DIGITAL MATURITY BY ENTERPRISE ARCHITECTURE APPROACH IN DIFFERENT SECTORS OF NATIONAL ECONOMY: A CASE OF THE CZECH REPUBLIC

Martin Lukáš, Miloš Ulman, Václav Lohr

Czech University of Life Sciences Prague
Faculty of Economics and Management
lukas@pef.czu.cz, ulman@pef.czu.cz, lohr@pef.czu.cz

DOI: 10.35011/IDIMT-2024-223

Keywords

Public administration; digital transformation; digital maturity; graphical language; enterprise architecture

Abstract

Digital transformation (DT) and digital maturity (DM) are long-term approaches that require entities to have many social-technical, organisational, and personal prerequisites to ensure transformation successfully. These complex prerequisites become increasingly important and depend on the types of entities, internal drivers, motivations for the change, skills, and maturity of modelling the enterprise architecture (EA). As understanding the current state (AS-IS) plays an important role, having tools and capabilities facilitating communication between business subject matter experts of DT via an enterprise architecture approach is essential. Little is known about how entities tackle the prerequisites of DT & DM. We present a case based on a survey of business entities in the Czech Republic's sector. While business entities from the finance and banking sectors reported the largest use of EA and modelling AS-IS state, the business entities in the energy sector show the lowest in both. Interestingly, the public entities are mostly in the middle. Support of DM in the early stage of planning by matured social-technical, organisational, and personal prerequisites determines, to some extent, the success of the DT process. Further research is needed on the barriers and challenges of DM from another aspect than what is mentioned in this article.

1. Introduction

Digital maturity (DM) is a phenomenon that has emerged along with the digital economy and Industry 4.0. Till now, there is no common definition and understanding of this process, so this research topic has plenty of room to study and explore (Aslanova & Kulichkina, 2020). Another view on the same term is DM provides evidence that firms with higher DM earn superior corporate performance. Business and industry players are asked to run the business wheel actively to survive in this digital transformation era (Firmananda et al., 2024). Digital Transformation (DT) integrates digital technology into all business areas, fundamentally changing how businesses operate and deliver customer value. Digital transformation combines multiple digitalisation projects with the intention of "customer-driven end-to-end strategic business transformation that requires organisation-level changes to core competency.” Many scholars define opportunities created by digitisation in organisational patterns and cultural barriers and transform existing business models, socio-economic cultures, and legal measures (Gunawardene, 2017). Zachman's Framework for Enterprise Architecture (ZFEA) is a descriptive, holistic representation of an enterprise that provides insights. Some scholars claim that Enterprise architecture (EA) is imperative to ensure successful business structures or business-IT alignment, or more recently with Enterprise Architecture Management (EAM), to manage required organisational transformation (Gerber et al., 2020).

Most emerging new technology, digital products and services try to penetrate the service providers and consumers without considering the fundaments of business processes (the fundaments of EA). Having considered the potential to work closely with consumers & customers and enable them wherever they live, it is a digital channel (e.g. mobile & web-based application) or physical environment (e.g. branch network, etc.) and/or combination of both quite often opening many not deeply investigated areas of how the understanding of the current state of the EA impacts DM of organisations in different sectors. Vice versa, how DM relies on the understanding of the EA.

Digital transformation (DT) has been increasingly focused on by researchers, practitioners, and politicians (Siegel & Gabryelczyk, 2021). In many sectors of the Czech Republic (finance, energy, public administration), if we look a little bit behind the shoulders back a few years of history, the initiatives focusing on laying down the solid cornerstone on how to develop and maintain EA effectively to increase DM and enhance it.

In the public sector, DT aims to bring positive, tangible changes in people's lives, which makes it different from mere information technology (IT) enabled transformation that is more focused on improving the operation of an entity through IT (Khisro, 2021).

Learning from the mistakes and shortcomings of others helps the financial, energy and public administration sectors in the Czech Republic sometimes not to repeat the same failures. Therefore, the DM of the leading players in these sectors is based on the EA approach (business architecture, business process improvement, application portfolio management, microservice architecture, robust backend data centre infrastructures, etc.). The paper aims to analyse how DM depends on the EA modelling (EAM) approach, especially how EAM affects digital transformation and DM. In particular, we focus on the analysis of the use of the EAM and its methodological tools in different sectors of the Czech national economy, the use of the graphical notation language ArchiMate for EA modelling as a basic assumption for DT, the modelling of the AS-IS state to enhance DM.

The paper is structured as follows. First, we outline the issues with the DT phenomenon, its relation to DM, and the penetration of digital technologies into many aspects of the national economies. We also illustrate how EAM creates a solid background for DT and provide more details on the Czech national level. Further, we present findings from a survey with 55 respondents spread across three different, however essential, sectors of the Czech national economy. Based on the survey, we analyse the differences in the specific needs of using the EAM approach based on the graphical notation language ArchiMate and its usage for modelling AS-IS state in mentioned sectors of the Czech national economy. We identify differences between the sectors and point out the results. In the final section, we discuss our findings and suggest further research work.

2. Literature review

2.1 Enterprise architecture methods and tools, digital transformation, and digital maturity

Enterprise architecture (EA) defines and represents a high-level view of an enterprise's business processes and IT systems, their interrelationships, and the extent to which different parts of the enterprise share these processes and systems. EA aims to define a suitable operating platform to support an organisation's future goals and the roadmap for moving toward this vision. Despite significant practitioner interest in the domain, understanding the value of EA remains a challenge (Tamm, Seddon, Shanks, Reynolds, 2011). As an organisational role, EA is positioned between IT and business strategy formulation on the one hand and project-focused solution architecting (sometimes called system architecting) on the other.

The task of EA is to translate the broader principles, capabilities, and goals defined in the strategies into systems and processes that enable the enterprise to realise these goals. In this regard, EA is a step towards enacting a strategy. In turn, EA guides solution architectures defined before specific development or implementation projects and provides the finer specifications for operationalising those systems. Its mediating role means that EA is like a strategy that aims to provide a long-term and organisation-wide vision of business processes and IT systems. Still, it describes this vision in greater detail (Tamm et al., 2011). EA is a common communication framework for experts to support achieving the enterprise's operational and developmental goals (Lukáš & Ulman, 2020).

The significant practitioner and organisational interest in DM supported by the EA approach on the one hand, as well as long-term effort to demonstrate the value of graphical notation language used commonly to develop and maintain EA as a platform for DT to get a higher ratio of DM across different sectors of Czech Republic, on the other, shows that there are many questions to be answered.

Assuming a holistic stance, EAM covers all dimensions of an enterprise (business, application, information, data, and infrastructure aspects), fosters the usage of a common language, and provides a consistent decision base to align business ultimately and IT (Aier, Riege, and Winter, 2008). The embracing nature of an EA coupled with the constantly changing environment in which its management takes place gives rise to many severe challenges. As an example, EAM might suffer from a symptom literature calls' ivory tower syndrome' (van der Raadt et al., 2008), describing the situation where there exists a disparity of stakeholders' requirements and delivered EA products.

In the past few decades, we have seen digital technologies' rising penetration of our lives. They influence many aspects of social & professional life. This type of penetration, with essential changes and impacts on lives, is often called digital transformation (DT). An evolutionary process that leverages digital capabilities and technologies to enable business models, operational processes, and customer experiences to create value (Morakanyane et al., 2017). The realignment of technology and new business models to more effectively engage digital customers at every touchpoint in the experience lifecycle (Schuchmann & Seufert, 2015).

The term "digital maturity" is closely related to digital transformation. We can say that digital maturity is the final stage of digital transformation, which companies aspire to achieve; those that have achieved such digital maturity have now witnessed important improvements in the company's operation and increased customer satisfaction (Domingues, 2015).

In general, maturity models offer a way to make an object of interest's progress towards a target state tangible across various management research disciplines – for single business units, an organisation, or whole industries. DM on the EU country level is captured regularly each year, known as the DESI index (DESI index, 2022). We conclude that the change in the DM level of a financial institution depends on the quantity and quality of current banking products and services modified (Magomaeva et al., 2020). Firms with more employees seem more strongly impacted and better prepared for the digital transformation. Concerning the industry cluster, no significant effects could be identified. A high EBIT margin predicts a slightly stronger impact and better-prepared firms. (Remane at al. 2017).

3. Methodology

Based on the empirical evidence presented above, we formulate the following research questions (RQ) that examine and search for the relationship between the EA approach and graphical language ArchiMate as a driver for communication with the EA project and DM of given sectors of the Czech Republic. We continue to investigate the area of EA and DT. Lukáš et al. (2022) found that the "EA approach sufficiently supports the preparation phase of digitalisation projects in local public administration projects?". Based on the findings, we formulate the following research questions.

RQ#1 Is there a standard graphical language, ArchiMate, that can be used to model the enterprise architecture? **RQ#2** What kind of sector is more digitally mature (and uses enterprise architecture approach and graphical language ArchiMate) than the others? **RQ#3** Are the AS-IS models of specific sectors created and maintained in the graphical language ArchiMate?

As regards the limited empirical results on DT, DM and EA challenges mentioned above and their practical relevance to industry, we have been conducting an exploratory survey across given sectors of the Czech Republic. This survey consists of 9 questions spread into three sections. The first section contained three identification questions in which respondents were asked to provide information about their position in the enterprise (job title) as well the sector (public administration, energy, finance) where the enterprise in which the respondent works acts.

The last question of the identification section focused on enterprise volume in terms of employees. The second section gathered information about EAM's support. This means that respondents were asked to estimate the complexity of EAM based on the number of people directly connected and/or acting in some EA role. The last section of the survey focused on using the graphical notation language ArchiMate within the enterprise and its practical usage for the AS-IS stage description of the enterprise. The compiled survey was accessible online.

This design, however, is a non-exhaustive set of questions covering EA challenges in the presented related works to examine the relation between EA, DT, and DM. The optimisation phase was initiated once the researchers completed the survey design process. We piloted the survey with three participants, each from a different sector (public administration, energy, and finance), and rephrased it based on the pilot persons' feedback. To receive respective and relevant information from the survey, we targeted respondents acting and working in managerial, organisational, software development, consultancy, and other related positions in a given sector of the Czech Republic. We sent over 105 survey invitations via e-mail, including a personalised cover letter. The distribution list of respondents has been compiled based on the personal network of the leader of the research (first and main author). Additionally, we published the survey in the Google Forms platform between April and June 2023.

Table 1. Respondents by Job title

Job title	#	%
IT Administrator	10	18,2
IT Analyst	4	7,3
Enterprise Architect	7	12,8
IT Specialist	4	7,3
IT Engineer	2	3,6
IT Consultant	1	1,8
IT Coordinator	7	12,7

Job title	#	%
IT Director	16	29,1
Chief Executive Officer	2	3,6
Delivery director	1	1,8
Project manager	1	1,8
Total	**55**	**100**

Table 2. Enterprise per industry sector

Industry sector	#	%
Finance/Banking/Insurance	22	40,0
Energy (Power supply & distribution)	12	21,9
Public Administration	18	32,7
Chemical industry	1	1,8
Services	1	1,8
IT Services	1	1,8
Total	55	100

Table 3. Enterprise per employees

Enterprise volume	#	%
0-19 employees	3	5,45
20-49 employees	2	3,64
50-99 employees	9	16,37
100-1001 employees	20	36,36
1001+ employees	21	38,18
Total	**55**	**100**

Table 4. Employees involved in EA

Volume of employees	#	%
1-3 employees	18	32,73
4-6 employees	17	30,9
7-10 employees	7	12,73
11+ employees	13	23,64
Total	**55**	**100**

The distribution of respondents by job title is shown in Table 1. Contrary to this, Table 2 illustrates the distribution of the industry sector of the respondents who belong to the enterprise. The volume of enterprise counted in the number of employees indicated in Table 3. These three tables belong to the Identification section of the online survey. Table 4, belonging to the second section of the survey, provides the view of personal support of EAM in the enterprise to which the respondent belongs.

4. Results

The essential findings of the survey are summarised in Table 5 and Table 6. Both represent the results of the third section of the online survey, which focused on using the graphical notation language ArchiMate in general and for AS-IS modelling. The indicators in Table 5 illustrate the usage of ArchiMate notation language for EA development/maintenance and AS-IS modelling. Unsurprisingly, two-thirds of the companies where the respondents work use the ArchiMate graphical language for EA, while one-third do not use ArchiMate. The question is, therefore, how they model and document EA. Almost two-thirds of enterprises use the notation language ArchiMate for AS-IS state modelling and documentation. Although this difference is small, it would be interesting to ascertain why companies use ArchiMate slightly more for EA development than AS-IS modelling.

Table 5. Usage of ArchiMate notation language

Type of ArchiMate usage		Yes	No	Total
EA development	#	36	19	**55**
	%	65,45	34,55	**100**
AS-IS modelling	#	35	20	**55**
	%	63,64	36,36	**100**

Table 6 illustrates the distribution of the type of use of ArchiMate by industries.

Table 6. Usage of ArchiMate notation langue per industry sector

Industry sector		**Yes**		**No**		**Total**	
		EA	**AS-IS**	**EA**	**AS-IS**	**EA**	**AS-IS**
Public Administration	#	12	12	6	6	**18**	**18**
	%	66,67	66,67	33,33	33,33	**100**	**100**
Finance/Banking/Insurance	#	9	8	3	4	**12**	**12**
	%	75,00	66,67	25,00	33,33	**100**	**100**
Energy (Power supply & distribution)	#	14	15	8	7	**22**	**22**
	%	63,64	63,64	36,34	36,36	**100**	**100**
Other sectors	#	3	3	3	3	**3**	**3**
	%	0	0	100	100	**100**	**100**

The graphical notation ArchiMate language is mostly used in the energy sector of the Czech Republic (Table 6). The public administration sector follows this, followed by the financial sector. The same is true for the creation of AS-IS states. In public administration, enterprises that use the graphical notation ArchiMate language always model AS-IS, but this is not the case for the banking sector. Not all enterprises use the graphical notation ArchiMate language model, the AS-IS state in ArchiMate.

5. Discussion and conclusion

In this discussion, we combine the theoretical and practical conclusions of researchers who have conducted research in roughly the same or similar research area as ours with the conclusions and ideas we have reached while designing, conducting, and developing the research. From an academic perspective, answering the questions above would help explain why organisations find EA useful. From a practical perspective, the answers would enable organisations to understand what benefits they might realistically expect from EA and accordingly make better-informed decisions on the appropriate level of investment in EA. Based on the empirical evidence from our research, it is obvious that EA contributes to DM through DT. The use of ArchiMate in more than 65 % of cases indicates that the graphical language can be used for EA modelling (RQ #1). Although ArchiMate is the frequent choice of EA visualization (Zhi & Zhou, 2022), it is also used for linkages with business models, creating ontologies, business model canvases, visualising operational processes in ITIL, and modelling business strategy concepts (Zhou et al., 2020). Using ArchiMate to analyse the current state of the business process and information systems (AS-IS modelling) is another use (RQ #3).

Because technical aspects drive digitisation, it would be useful to measure digital maturity with variables at all layers of a company (Schwer et al., 2018). The framework for EA, the graphical language ArchiMate, is used to compare the variables. The variables are assigned to the six layers of architecture: Strategy, Business Environment, Applications, Technology, Physical and Implementation, and Migration. The mapping between models of DM and ArchiMate is clearly shown in research conducted by (Schwer et al., 2018). Two-thirds of our sample reported using ArchiMate, indicating their digital maturity. The energy sector was most represented, followed by public administration and financial institutions (RQ #2). Ochoa-Urrego & Peña-Reyes (2021) confirm the relationship between the DM and ArchiMate layers. Their findings regarding analysing the ArchiMate dimensions used for modelling EA and DM are similar. DT of the industry involves using new digital technologies and requires the optimal organisation of business processes. Interconnected and interdependent layers are distinguished when building EA as an integrated EAM.

The multilevel structure of EAM determines the relationship between the main components of the system (Ilin et al., 2018). We also uncover and provide new opportunities for further research on the EA-DM relation that can improve understanding of other contexts and influences of EA on DM.

The further research opportunities are: **a)** enhancing the scope of the research internationally, **b)** focusing on TO-BE state and modelling in research (Hindarto, 2023), and **c)** incorporating EA aspects in the project & program management of DT initiatives.

We conclude that both empirical investigations and studies and soft methods based on exploring the understanding of how EA, its methods, and tools enhance the DM of key players in the national economy of the Czech Republic can advance the understanding of the relationship between EA and DM. Due to the requirement to approve digitisation projects with budgets over 6 million CZK in the Czech Republic, the city's countywide digitisation projects must be approved by the Department of the Chief Architect of eGovernment of the Ministry of the Interior. The project application includes an EA of the organisation, including an analysis of the current state, schemas in ArchiMate, and answers to several questions regarding the processes and IT systems operated by the organisation. Therefore, the lower-level authorities needed to strengthen their capacity and obtain the necessary competence to prepare the applications (Lukáš. et al., 2023).

5.1. The limitation of a case study

Despite its scope, we see two limitations of a case study that should be acknowledged and considered: 1) the respondents are from the Czech Republic, the case study covers specifics of the Czech Republic (not cover international conditions), and 2) the graphical notation language ArchiMate for EA modelling is de facto standard in the Czech Republic at national level (CNAP, 2024).

Acknowledgement

The results and knowledge included herein have been obtained owing to support from the following institutional grant. Internal grant agency of the Faculty of Economics and Management, Czech University of Life Sciences Prague, grant no. 2023B0005.

References

A Europe fit for the digital age. European Commission. Retrieved May 15, 2022 from https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/Europe-fit-digital-age_en.

Aier, S. & Riege, C. & Winter, R. (2008). Classification of EA scenarios, explanatory analysis. Model Archi 3(1): 14-23

Aslanova I.V. & Kulichkina A. (2020). Digital Maturity: Definition and Model. Advances in Economics, Business and Management Research, volume 138, 2nd International Scientific and Practical Conference "Modern Management Trends and the Digital Economy" (MTDE 2020). Atlantis Press. DOI 10.2991/aebmr.k.200502.073

CNAP, (2024). Czech National Architecture Plan - Vision of eGovernment of the Czech Republic, https://archi.gov.cz/en:start

DESI index (2022). Learn more about the Digital Economy and Society Index (DESI)

Firmananda, F. I., Ependi, Z. & Adeswastoto, H. (2024). Analysis and Design of Enterprise Architecture in The Implementation of Digital Transformation at Village Tourism of Kampung Patin Kampar Regency using TOGAF. Jurnal Teknik Industri Terintegrasi (JUTIN), 7(1), 130–136. https://doi.org/10.31004/jutin.v7i1.21332

Gerber, A., & Roux, P., & Kearney, C., & Merwe, A. (2020). The Zachman Framework for EA: An Explanatory IS Theory Responsible Design, Implementation and Use of ICT, 2020, Volume 12066. ISBN: 978-3-030-44998-8

Gunawardene N., (2017). Digital transformation in Sri Lanka: Opportunities and challenges in pursuit of liberal policies Asian Journal of Economics, Finance and Management 1(2): 69-78, 2019; Article no. AJEFM.105

Hafseld, K., H. J. & Hussein, B. & Rauza A., B. (2020). An attempt to understand complexity in a government digital transformation project. ISSN (print):2182-7796, ISSN (online):2182-7788, DOI: 10.12821/ijispm090304.

Hindarto, D. (2023). The Management of Projects is Improved Through Enterprise Architecture on Project Management Application Systems. International Journal Software Engineering and Computer Science (IJSECS). 3. 151-161. 10.35870/ijsecs.v3i2.1512.

Ilin, I., & Levina, A., & Borremans, A., & Kalyazina, S. (2020). Enterprise Architecture Modeling in Digital Transformation Era. 10.1007/978-3-030-57453-6_11.

Khisro, J. (2021). Strategising Digital Transformation: A Clinical Inquiry into a Swedish Public Sector Organization. Americas Conference of Information Systems AMCIS 202.

Lukáš, M., Ulman, M. (2020). Lost in Translation: Enterprise Architecture in e-Government Projects. Online. In: CEUR Workshop Proceedings volume 2797, 2020 Ongoing Research, Practitioners, Posters, Workshops, and Projects of the International Conference EGOV-CeDEM-ePart, EGOV-CeDEM-ePart 2020. CEUR-WS, 2020, pp. 279-287.

Lukáš, M.; Burešová T.; Ulman, M.; Havránek, M.; Jarolímek, J. (2023). Digital transformation at the national, regional, and local levels of public administration: A case of the Czech Republic. Online. In: IDIMT-2023 New Challenges for ICT and Management. 31st Interdisciplinary Information Management Talks. Linz: TRAUNER Verlag + Buchservice, 2023, pp. 187-193, 978-3-99151-176-2. DOI: 10.35011/IDIMT-2023-187.

Lukáš, M.; Burešová, T.; Ulman, M. (2022). A novel approach to business process analysis and optimisation based on enterprise architecture: a case of local public administration in the Czech Republic. Online. In: The Conference IDIMT 2022. Digitalization of Society, Business and Management in a Pandemic. Linz: TRAUNER Verlag + Buchservice 2022, pp. 59-66, 9783991137580. DOI: 10.35011/IDIMT-2022-59.

Magomaeva, L.R. & Galazova, S.S. & Magomaev. T.R. (2020). Assessment of the Level of DM of the Financial Institute, Depending on the Modification of the Available Banking Products&Services. DOI 10.2991/aebmr.k.201205.032. Proceedings of the 2nd International Scientific and Practical Conference on Digital Economy. P.195-201. ISBN 978-94-6239-291-5. Atlantis Press.

Morakanyane, R., Grace, A. A., & O'Reilly, P. (2017). Conceptualizing Digital Transformation in Business Organizations: A Systematic Review of Literature, 427–443. https://doi.org/10.18690/978-961-286-043-1.30.

Ochoa-Urrego, RL., Peña-Reyes, JI. (2021). Digital Maturity Models:A Systematic Literature Review. Schallmo, D.R.A., Tidd, J. Digitalization. Management for Professionals. Springer, Cham. https://doi.org/10.1007/978-3-030-69380-0_5

Remane, G. & Hanelt, A. & Wiesbock, F. & Kolbe, L. (2017). Digital Maturity in traditional industries – an exploratory analysis. 25th European Conference on Information Systems, Guimarães,Portugal. P.10. ISBN 978-989-207655.

Schuchmann, D., & Seufert, S. (2015). Corporate Learning in Times of Digital Transformation: A Conceptual Framework and Service Portfolio for the Learning Function in Banking Organisations. International Journal of Advanced Corporate Learning (IJAC), 8(1), 31–39. https://doi.org/10.3991/ijac.v8i1.4440.

Schwer, K. & Hitz, Ch. & Wyss, R. & Wirz, D. & Minonne, C. (2018). Digital maturity variables and their impact on the enterprise architecture layers. Problems and Perspectives in Management. 16. 141-154. 10.21511/ppm.16(4).2018.13.

Siegel, U. A., & Gabryelczyk, R. (2021). Exploring value streams and CSFs to foster DT in public admin. AMCIS.

Tamm, T. & P B Seddon, G Shanks, P Reynolds. (2011). Communications of the Association for Information Systems: How does enterprise architecture add value to organisations? Volume 28, Number 1, Article 10, 3-1-2011. Department of Information Systems, The University of Melbourne, Australia, 10536/DRO/DU:30153159.

Van der Raadt, B., van Vliet, H. (2008). Designing the EA Function. Quality of Software Architectures. Models and Architectures. QoSA 2008, vol 5281. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-87879-7_7.

Zhi, Q., & Zhou, Z. (2022). Empirically Modeling EA Using ArchiMate. Computer Systems Science&Engineering, 40(1).

Zhou, Z., Zhi, Q., Morisaki, S., & Yamamoto, S. (2020). A systematic literature review on enterprise architecture visualization methodologies. IEEE Access, 8.

RISK FACTORS IN IMPLEMENTATION OF INDUSTRY 4.0 FOR ERP/MES IN SMALL AND MEDIUM ENTERPRISES

Roman Danel

VŠB – Technical University Ostrava
Faculty of Economics
roman.danel@vsb.cz

VŠTE - Institute of Technology and Business in České Budějovice
Faculty of Technology
rdanel@mail.vstecb.cz

Bożena Gajdzik

Silesian University of Technology Gliwice
Faculty of Material Engineering
bozena.gajdzik@polsl.pl

Hana Štverková

VŠB – Technical University Ostrava
Faculty of Economics
hana.stverkova@vsb.cz

DOI: 10.35011/IDIMT-2024-231

Keywords

ERP; MES; pillars of Industry 4.0; IT/ICT; Smart manufacturing; risk factors

Abstract

The aim of the study is to identify risk factors in the area of using ERP/MES systems in the pillars of Industry 4.0. The outcome of the study is a matrix of risk factors assigned to specific technologies of Industry 4.0. The background of the analysis was the strong digitalization of production-logistics processes in the Fourth Industrial Revolution. Accelerated digitalization increasingly affects the small and medium-sized enterprise segment, which requires guidelines for assessing risk when investing in ERP/MES systems or improving them. The study continues with previous research on risk assessment during the implementation of ERP systems in small and medium-sized enterprises. The list of risk factors prepared by us is open-ended, and further expansion is anticipated.

1. Introduction

Risk is present in every business activity, human action, investment, and other developmental initiatives It is present at every level of management within an organization. Risk Management stands as one of the pivotal components of company management, integral to business models and decision-making processes. Managing risk is also a key aspect of the Business Continuity Plan (BCP). There exist numerous precise definitions of risk management within the corporate sphere (Dionne, 2013; Fan and Stevenson, 2018; Damayanti, 2023; Calandro, 2015; Hanafi, 2009). Following the definition of Risk Management, authors typically outline the stages of methodology, which commonly include risk analysis, risk categorization, risk assessment, and risk prevention (Roberto, 2009; Cornell, 1996; Sadgrove, 2015; Manuj & Mentzer, 2008; Marhavilas, Koulouriotis and Gemeni, 2011). Organizing the entire process procedurally involves establishing and developing various documents related to risk management.

In the context of the highly popular concept of industrial development based on Fourth Industrial Revolution technologies, known as Industry 4.0, the importance and significance of IT risk analysis are increasing. This includes the analysis of key systems supporting business processes, such as ERP (Enterprise Resource Planning) and MES (Manufacturing Execution Systems). As these systems complement each other strongly within an enterprise that aims to leverage Fourth Industrial Revolution technologies to create smart manufacturing processes, also known as smart factories, the need for robust IT risk analysis becomes even more pronounced. Industry 4.0 mainly refers to MES systems, however, it is also linked to the functions of ERP systems (planning, logistics and others). That is why we use the ERP/MES connection in the following text.

The risk analysis in the utilization of ERP/MES systems in the context of Industry 4.0 pillars such as IoT (Internet of Things), cloud, artificial intelligence, Big Data & data analytics, collaborative robots, additive manufacturing, modelling and simulation, system integration, augmented/virtual reality and digital twin, involves a systematic approach to identifying and assessing the risks associated with the operation of these systems within each pillar of Industry 4.0. (For more information on Industry 4.0 pillars, refer to publications (Erboz, 2017; Hermann et al, 2016; Lu, 2017). Therefore, in the main part of the papers, the risk was typified in ERP/MES systems in the aforementioned specified pillars. The identification of risk in ERP/MES systems within the individual pillars of Industry 4.0 represents a new research field that is continuously evolving as industrial sectors and small and medium sized enterprises are in the phase of implementing smart manufacturing. Consequently, the subject of evaluating risk associated with IT systems in the context of the set of Industry 4.0 technologies is continuously developed.

The paper consists of three sections. Section 1: ERP/MES in Industry 4.0. This section is based on a literature review and serves as background to our analysis. Section 2 is based on direct research conducted in SMEs in the Czech Republic. The research topic was the assessment of risk in implementing an ERP system. The result of our research was a risk typology matrix. ERP and MES implementations differ depending on the size of the company. Small enterprises typically lack the financial resources for large-scale and intricate ERP systems. Medium-sized enterprises often acquire systems that incorporate both ERP and MES functions. In contrast, large companies typically maintain separate ERP and MES systems, with data aggregated from the MES system being transferred to the ERP system. In the context of Industry 4.0, technologies are primarily implemented within MES systems. However, Industry 4.0 demands a more sophisticated approach, necessitating the inclusion of ERP system functionality in the solutions. Section 3, which the authors consider fundamental, pertains to the typology of risk in various technological segments of Industry 4.0 within the implementation and development of ERP/MES systems. The results presented in this section are based on the authors' expert knowledge. The final compilation in the form of a matrix of risk analysis of ERP/MES systems in the pillars of Industry 4.0 constitutes a significant added value to IT science.

2. Trends ERP/MES in the context of Industry 4.0

Industry 4.0 represents a highly dynamic convergence between IT and manufacturing companies. The following trends are frequently highlighted: Big Data, cloud, Internet of Things (IoT), simulation, digitalization, additive manufacturing, advanced control algorithm, augmented/virtual reality, artificial intelligence etc. (Hermann et al., 2016; Kagermann et al. 2013). Cloud solutions and big data sets are already relatively widely employed and utilized, serving as a kind of pivotal factor enabling ongoing transformations. But the milestone of the new trend is the IoT. The increasing availability of internet connectivity, decreasing costs of internet connection, and the growing number of devices embracing Wi-Fi technology and others are perfect for creating IoT (Chui et al., 2016; Atzori et al, 2016; Li et al., 2014). The integration of ERP/MES and IoT is also crucial worldwide. For instance, global ERP leaders such as SAP, Microsoft, Infor and Oracle are also leading in IoT. Many national strategies of leading industrialized countries devote attention to this topic. There are numerous ways to address readiness for Industry 4.0 in companies. (https://www.industrie40-readiness.de/, https://przemyslprzyszlosci.gov.pl/, https://www.firma40.cz/). IoT connects technologies, systems, processes, products, and people (Lie et al., 2014). Continuous digitalization is at the core of Industry 4.0 (Kagermann, 2015; McKinsey, 2015). Although production digitalisation began in the 1980s - over 30 years ago, today it takes on a new form. ERP systems have played a significant role in the applied concepts and platforms from the beginning. Meanwhile, MES systems emerged somewhat later. The process of digitalization is inconceivable without business applications - ERP and MES systems. ERP systems played a significant role in digitizing companies in the 1990s (N.N.A Brief History).

Recently, ERP/MES systems have responded to innovations such as IoT, social networks, mobile devices, and other process solutions towards smart initiatives. Companies believe that ERP/MES systems are crucial for supporting processes towards Industry 4.0, serving as the foundation of its architecture (Basl, 2017, Berić et al. 2018; Majstorovic et al., 2020). Why? (Berić et al., 2018). Conversely, MES systems, which are strongly developed in enterprises moving towards smart factories, answer the questions: What to? (Kletti, 2007). ERP system is at the enterprise control level and MES system at the manufacturing control level, but both are needed in an enterprise going to Industry 4. 0. The development of systems (MRP/ERP) has a long history, and the development has gone in two directions: (i) the business aspect - from inventory planning at the plant level to the entire chain (inquiry - delivery of finished product) at the company level, and (ii) the technology aspect - from software package to a client server architecture (Majstorovic et al., 2020). Historically, the ERP system went through the following stages: 1960s / IC / I level: warehouse control application; 1970s / MRP I / II level: Materials Requirements Planning, 1980s / MRP II / III level: Manufacturing Resources Planning; 1990s / ERP / IV level Enterprise Resource Planning in Integrated business activities at the organization units, 2000s / ERP II / V level Enterprise Resource Planning by Internet (Services Oriented Architecture (SOA)), 2010s / Cloud based ERP / VI level Cloud based ERP (ERP as software a service (SaaS) model), 2020s / I4.0 ERP / VII level ERP of Industry 4.0 model I4.0 concept introducing (Majstorovic et al., 2020).

In publication (Basl, 2017) there is presented the evolution of ERP systems, taking the beginning of the 1980s as the starting point. The evolution process begins with FMS (Flexibility Manufacturing System), followed by CAD/CAM and CIM (Computer Integrated Manufacturing) in the 1990s (Scheer, 1987), and subsequent enhancements in contemporary times (early 21st century), integrating ERP with CRM + SCM, further augmented by PLM (Product Lifecycle Management). The year 2010, according to the author (Basl, 2017), marks the beginning of integration between ERP and MES, where this integration receives support from the APS system (Advanced Planning and Scheduling). In industry 4.0, ERP entered the Industry 4.0 concept and evolved alongside technological solutions towards enhancing the utility of this system in Smart factories and Digital Twins (Hochmuth et al., 2017; Longo et al., 2019; Rudberg and Sandelin, 2017).

MES solutions provide real-time information about production (location: shop floor). This system is an information bridge between the planning systems used in strategic production management (such as ERP) and SCADA (Supervisory Control and Data Acquisition) production floor control. It bridges the layers of the Production Information System: strategic planning and direct execution, through proper management and control of up-to-date information related to the company's core resources (Mejía et al., 2007). Core functions of MES according organisation MESA are Planning System Interface, Data Collection, Exception Management, Work Orders, Work Stations, Inventory and Material Movement. Supporting functions of MES are Genealogy, Maintenance, Time and Attendance, Statistical Process Control, Quality Assurance, Process Data and Documentation Management. However, there is an increasing need to provide support defining and implementing an interoperability relationship between these manufacturing software and business applications such as ERP systems (Panetto and Molina, 2008).

In this current phase of evolution, ERP/MES systems require (collaborate with) the following technologies: IoT Apps, machine-to-machine communication, man-machine communication, Artificial Intelligence, and Big Data Analytics (Berić et al. 2018). The development of ERP/MES systems can also be presented in the context of data development, as did the team of researchers Majstorovic et al. (2020), from medium to Big Data and cloud computing technology. Reis and Gins in their paper (2017) present such trends in data processing: (i) from Univariate, to Multivariate, to High-Dimensional ("Mega-Variate") (ii) from Homogeneous Data Tables to Heterogeneous Datasets, (iii) From Static, to Dynamic, to Non-Stationary, (iv) from Monitoring the Mean, to Dispersion, to Correlation (v) From Unstructured to Structured Process Monitoring (Reis and Gins, 2017). In the Industry 4.0 there is a need for cooperation of computer systems with cloud computing because of Big Data and better software (Xu, 2012). The challenge for the development of ERP/MES systems are smart manufacturing (Olson et al., 2018; Kusiak, 2019; Wang et al, 2016; Gajdzik et al, 2024). In the smart manufacturing, information flows are implemented through the cloud and physical layers. The cloud layer includes models and algorithms related to: operations and configuration management, process and service models, and status monitoring. ERP covers this last layer. The physical layer includes devices and sensors at the production level. MES system is needed at this level. In this way, distributed virtual-physical systems through cloud computing realize resource sharing, managed through ERP/MES systems (Majstorovic et al., 2020). There is no universal path for implementing Industry 4.0 solutions (Gajdzik et al., 2021). The level of digitization of processes varies across industries (e. g. steel industry) (Gajdzik and Wolniak 2021a.b, Gajdzik, 2021). In addition, the development of ERP/MES systems is adapted to the types of production (Danel and Gajdzik, 2024). There are three fundamentally different types of production: discrete, continuous and batch. The differences in the implementation of Industry 4.0 in continuous production processes compared to other types of production are analysed in (Danel & Chlopecký, 2023) and (Danel, Gajdzik and Ropyak, 2023; Danel and Gajdzik, 2024). In continuous production, the implementation of new technologies is slower. The main control task is ensuring the required quality of output products in real-time during production process.

It is evident that the concept of Industry 4.0 is based on industrial integration through information technologies. This integration involves real-time or near-real-time data sharing, information exchange, and continuous communication. It also represents potential for further development of ERP/MES systems.

3. ERP systems risk analysis based on field research

At the Faculty of Economics of VŠB – TU Ostrava, there was carried out research regarding the implementation of ERP systems in small and medium-sized enterprises (Štverková, 2021). This research also included risk analysis and analysis of failed information system implementations.

Examples of unsuccessful companies in the ERP implementing include, for example, Lidl, Mall, Deutsche Post and others, who invested significant financial resources and time, but without adequate results. Lidl devoted seven years and 500 million EUR to preparing for the implementation of the SAP system, terminating its implementation in 2018. Similarly, Deutsche Post - DHL incurred the same costs in an unsuccessful attempt to implement this system in 2015 (van Marle 2015). Lidl's project was ambitious and transformational, with correspondingly high expectations (such as reducing efforts in maintaining master data, real-time key data, forecasting analysis, etc.). Therefore, the company's management considered returning to the original Wavi system (Bayer 2018).

For the implementation of ERP systems, the following lessons emerge from this:

- Implementation must not take several years,
- The digital transformation program should not be based on the paradigm of an old solution created long ago.

Many small and medium-sized enterprises enthusiastically adopt an ERP system, believing that its implementation will advance their competitiveness and support the growth of their business. However, unsuccessful ERP implementation can have major economic consequences. The effect of underestimated threats of ERP implementation can be huge, only after the operation is stopped. In order to avoid such a mistake in ERP implementation, it is important to understand the risks. (Wright and Wright 2002).

After examining all stages of ERP implementation lifecycle (planning - acquisition - implementation - usage - expansion), it was found that prior to ERP implementation, careful inspection and analysis should be conducted. In a study conducted at the Faculty of Economics, after identifying risks, the FMEA method (Failure Mode and Effect Analysis) was utilized for their assessment. FMEA is a systematic and proactive method that includes an overview of steps in the process, failure modes, causes of failure, and consequences of failure (Carbone and Tippett, 2004). The primary task of FMEA is to prevent problems associated with the system, products, and operational processes before they occur. The research involved the analysis of secondary data (based on the analysis of published data) and original empirical research among ERP implementation experts using a questionnaire method. Respondents were divided into two main categories: consultants (both functional and technical) and managers who are experienced members of ERP implementation teams. The research was two-phased. The first phase concerns general information about the respondents, their current affiliations, experiences, ERP experiences, and FMEA knowledge. The second phase includes questions related to the application of FMEA. It was found that 76 % of respondents had previous knowledge of FMEA (Štverková, 2021).

To determine the main causes of risks, it is necessary to divide them into several categories and then outline appropriate measures to prevent them. As already mentioned, risks are divided according to five different phases of ERP implementation. Table 1 presents the distribution of risks together with an assessment of their severity (S), occurrence (O) and detection (D), which is represented by the Risk Priority Number (RPN).

Table 1. Results of research focused on individual risks for all stages of ERP implementation in small and medium-sized enterprises

Stages	Risk Type	S	O	D	RPN
Planning	Lack of owner involvement	10	4	8	320
	Inefficient communication with the user	8	6	7	336
	Inappropriate financial management	9	7	8	504
Acquisition	Inadequate technology and infrastructure	9	5	7	315
	Inadequate knowledge of ERP and supplier selection	8	7	5	216
Implementation	Insufficient knowledge of project management	7	8	5	280
	Inefficient ERP project management	8	5	3	120
	Changes in the scope of the project	9	5	8	360
	Insufficient and ineffective training	7	7	6	245
Usage	Misuse in business operations	5	5	9	225
	User resistance	9	8	4	288
	Insufficient motivation of users by management	8	7	5	280
	Abuse of the ERP system	9	8	8	576
Expansion	Business process integration mismatch with supply chain partners	7	6	3	126
	Loss of business due to non-adoption of technology	7	6	6	252

RPN is calculated on following equation:

$$RPN = severity \times occurence \times detection \quad (1)$$

Where importance, occurrence and detectability are calculated as a mean value from an expert assessment in the range of 0 to 10, where 0 is the lowest and 10 is the highest (https://www.iqasystem.com/news/risk-priority-number/).

Based on the RPN, according to the respondents, the risk of misuse of the ERP system is perceived as the highest with an RPN value of 576. All three risks in Table 1 associated with the planning phase of ERP adoption, are perceived as high. In their research, the authors focused on the application of FMEA in ERP implementation in small and medium-sized enterprises. The output of the FMEA analysis is the perception of risk at each stage of implementation and the description of measures that should be accepted by key stakeholders, leading to the streamlining and simplification of the implementation process (Štverková, 2021).

4. Risk factors identification in Industry 4.0

The previous chapter presented the results of risk analysis research on the implementation of ERP systems in small and medium-sized enterprises. In the next part, we will analyse the risks from the point of view of implementing the Industry 4.0 concept technology in ERP/MES systems. The list of risk factors that we have prepared is the first stage of risk assessment in the conditions of Industry 4.0. There is a gap in the market for risk analysis in ERP/MES systems and the implementation of the Industry 4.0

concept, both from the demand side and from the system suppliers. The prepared list has an open form and we anticipate further changes to it. Small and medium-sized enterprises can use this list when making investment decisions in these systems or modifying them in cooperation with Industry 4.0 technologies. The list we mentioned is not subject to evaluation, only categorization.

The first segment we analyse concerns ERP/MES & IoT (Segment 1: S1). The following risk factors have been designated for this segment: S1.1 data transmission speed, S 1.2 reliability of data, S 1.3 quality of network, S 1.4 quality of availability (WAN network like LoRa, SigFox, etc.), S 1.5 energy consumption (stability, durability, etc.), S 1.6 data accuracy and verification, S 1.7 safety, S 1.8 security, S 1.9 recovery model, S 1.10 diagnostics S 1.11 level of maintenance. Total of risk factors in the pillar (S2) is 11. The next segment is about ERP/MES & cloud computing (S2) and following risk has been established (in total 5): S 2.1 security S 2.2 confidentiality S 2.3 connectivity, S 2.4 dependency on providers, S 2.5 risk according type of cloud services (private, public, hybrid) S 2.6 service availability, S 2.7 usability of investment. Third segment is connection ERP/MES with Artificial Intelligence (S3) and consist of S 3.1 incorrectly prepared model given wrong results, S 3.2 verification of results, S 3.3 legal aspects (who is responsible for results and decisions) – in total 3 factors. With fourth segment, ERP/MES and Big Data/Data Analytics (S4) we detected in total 8 risk: S 4.1 security, S 4.2 demands on workers' knowledge, S 4.3 data safety and integrity, S 4.4 model verification, S 4.5 usefulness of models, S 4.6 level of autonomy, S 4.7 data quality, S 4.8 legal aspects (e.g. GDPR, data ownership, …). Fifth segment (S5) connects ERP/MES with application of cobots and there are following risk factors: S 5.1 safety (cooperation with operators), S 5.2 fault-tolerant solution (in total 2). Sixth segment (S6) is about additive production (3D print) in ERP/MES with 7 risk factors: S 6.1 safety, S 6.2 environmental risk, S 6.3 material defects, S 6.4 unsuitable technology and material for the given type of production, S 6.5 workplace category from the point of view of people's health, S 6.6 insufficient and incorrect commercial production planning, S 6.7 the risk of missing integration between the use of 3D printing for the production of spare parts and maintenance planning. Seventh segment connects ERP/MES with modelling and simulation (S7): S 7.1 demands on workers' knowledge, S 7.2 reality and verification of models, S 7.3 incorrectly prepared or selected model leading to erroneous simulation results, S 7.4 insufficient connection between simulations and the decision-making system, S 7.5 insufficiently tuned algorithm. In total 5 factors. Systems integration within ERP/MES gives 8th segment (S8) with 6 risk factors: S 8.1 missing communication standards (e.g. sensor interfaces), S 8.2 misunderstanding the concept of systems leading to wrong implementation, S 8.3 the quality of the integration solution, S 8.4 different user interfaces, S 8.5 security and authorization, 8.6 demands on workers' knowledge. Last segment (S 9) is ERP/MES & augmented/virtual reality/digital twins with S 9.1 user disabilities S 9.2 the effectiveness of using virtual reality, S 9.3 the difference between real object and its virtual model, S 9.4 the effectiveness (usability by investment). There are 4 factors in total.

Table 2 presents combinations with common parts for risk factors: security (4 pillars), legal aspects (2 pillars) and demand on works knowledge (4 pillars). The analysis shows that the most factors appear in the S1 pillar - IoT and S4 - big data and analytics. In the risk analysis during ERP/MES implementation, these pillars should receive the greatest attention.

Table 2. Matrix of common risk factors in analysed pillars of Industry 4.0

Risk factors/Pillars of Industry 4.0	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**
Security	1	1		1		1			
Legal aspects (who is responsible for results and decisions/GDPR, data ownership)			1	1					
Demands on workers' knowledge				1	1		1	1	

To compare a number of risk factor in particular pillars in Industry 4.0, authors summarized the number of risk factor and fixed ranks for them in table 3.

Table 3. Ranking of pillars of Industry 4.0 according of risk factors

pillar	S1	S4	S6	S2	S8	S7	S9	S3	S5
total	11	8	7	7	6	5	4	3	2
rank	1	2	3/4	3/4	5	6	7	8	9

Based on analysis following recommendation for small and medium enterprises that want to implement or improve ERP/MES with Industry 4.0 technologies are:

- On management and operating levels in the ERP/MES systems enterprises have ensure procedures of security particular for pillars: IoT (S1), Cloud (S2), Big Data (S4), additive production (S6) (Table 2)
- Establishment of a person with responsibility for data security and for result responsibility both, in implementation of pillars AI (S3) and Big Data (S4) in ERP/MES systems (legal aspects in the Table 2).
- Ensuring continuous access to knowledge management (demands on workers' knowledge in table 2) in ERP/MES systems in pillars: Big Data (S4), cobots and autonomous robots (S5), modelling and simulation (S7) and System integration (S8).
- At the current level of Industry 4.0 development, most risk factors can be identified in the S1 (IoT) pillar (Table 3), which is due to the fact that it is an external technology that is adapted to the enterprise in the form of IIoT (Industrial Internet of Thing).
- The fewest risk factors (Table 3) were identified for the S5 pillar (cobots), due to the fact that it is an extension of existing industrial robots commonly used.

5. Conclusion

The study presents analysis that continues with previous research on risk assessment during the implementation of ERP/MES systems in small and medium-sized enterprises. The aim of the study has been analysis of risk factors within implementation of Industry 4.0 technologies in ERP/MES systems. The list of risk factors prepared by authors is open-ended, and further expansion is anticipated. Next step in research should be assessing of risk in compliant with ISO standards. Every company should assess risk into small, medium and large, taking into account internal and external conditions when implementing the pillars of Industry 4.0 in ERP/MES.

Acknowledgement

This paper has been realized due to support from the 11/040/RGJ23/0032 project provide by Silesian University of Technology, Gliwice, Poland. The project was realized during research internship of B. Gajdzik at the VŠB – Technical University Ostrava and VŠTE - Institute of technology and Business in České Budějovice.

References

Atzori L., Iera A., Morabito G. (2010). The Internet of Things: A survey. Computer Networks, 54 (15), 2787-2805

Bayer, M. (2018). eLWIS auf HANA-Basis ist tot: Lidl stoppt millionenschweres SAP-Projekt für Warenwirtschaft [online], [Accessed December 19, 2020]. Available at: https://www.cio.de/a/lidl-stoppt-millionenschweres-sap-projekt-fuer-warenwirtschaft,3545457

Basl, J. (2017). Penetration of Industry 4.0 Principles into ERP Vendors' Products and Services – A Central European Study. 11th International Conference on Research and Practical Issues of Enterprise Information Systems (CONFENIS), Shanghai, China. pp.81-90

Berić, D. Darko Stefanović, Bojan Lalić, Ilija Ćosić (2018). The Implementation of ERP and MES Systems as a Support to Industrial Management Systems. International Journal of Industrial Engineering and Management (IJIEM), 9(2), pp. 77-86. Available online at www.iim.ftn.uns.ac.rs/ijiem_journal.php

Calandro, J. (2015). A leader's guide to strategic risk management. Strategy and Leadership, 43(1), 26–35.

Carbone, T. A. and Tippet, D. (2004). Project risk management using the project risk FMEA. *Engineering Management Journal*, 16(4), 28-35. ISSN 1042-9247

Chui M., Löffler M., Roberts R. (2016). The Internet of Things. The McKinsey Quarterly, 2 (47), 1-9.

Damayanti E. S. (2023). Risk Management: In an Overview of Literature Review. Formosa Journal of Science and Technology (FJST) 2(4), 1115-1122

Danel, R. and Chlopecký, J. (2023). Implementation of Industry 4.0 in Continuous Production. 15th International Conference on Strategic Management and its Support by Information Systems, SMSIS 2023, pp. 36–43

Danel, R. and Gajdzik, B. (2024). Integrated Subsystems of Materials and Information Flow for Continuous Manufacturing of Coal and Steel. Management Systems in Production Engineering 2024(2).

Danel, R., Gajdzik, B., and Ropyak, L. (2023). Trends in Data Processing Control in Continuous Production Systems. Scientific Papers of Silesian University of Technology. Organization & Management Series. Zabrze, Poland, nr. 181, pp. 93-105. ISSN 1641-3466

Dionne, G. (2013). Risk Management: History, Definition and Critique. Risk Management and Insurance Review 16(2), 147-166, http://dx.doi.org/10.2139/ssrn.2231635

Erboz, G. (2017). How To Define Industry 4.0: Main Pillars of Industry 4.0. Conference: Managerial trends in the development of enterprises in globalization era. Slovak University of Agriculture in Nitra, Slovakia

Fan, Y. and Stevenson, M. (2018) A review of supply chain risk management: definition, theory, and research agenda. International Journal of Physical Distribution and Logistics Management, 48 (3). pp. 205-230. ISSN 0960-0035.

Gajdzik, B. (2021). Transformation from Steelworks 3.0 to Steelworks 4.0: Key Technologies of Industry 4.0 and their Usefulness for Polish Steelworks in Direct Research, European Research Studies Journal, vol. 0(3B), pp. 61-71.

Gajdzik B., Grabowska S., Saniuk S. (2021). A Theoretical Framework for Industry 4.0 and Its Implementation with Selected Practical Schedules. Energies, 14 (4), 1-24, 940

Gajdzik B. and Wolniak, R. (2021a). Digitalization and Innovation in the Steel Industry in Poland — Selected Tools of ICT in an Analysis of Statistical Data and a Case Study. Energies, 14, 3034, 1-25

Gajdzik B., Wolniak R. (2021b). Transitioning of Steel Producers to the Steelworks 4.0—Literature Review with Case Studies. Energies, 14, 4109; https://doi.org/10.3390/en14144109

Gajdzik, B., Wolniak, R., Grebski, W., Szymszal, J., Grebski M. E. (2024). Sustainability and Smart Manufacturing: The Transformation of the Steelwork Industry. London: Routledge. eBook ISBN: 9781032697611

Hanafi, M. (2009). Resiko Manajemen Edisi kedua. Yogyakarta: STIM YKPN

Hermann, M., Pentek, T., Otto, B. (2016). Design Principles for Industry 4.0 Scenarios: A Literature Review. Available at: https://www.computer.org/csdl/proceedings/hicss/2016/5670/00/5670d928.pdf [Accessed July 27, 2017].

Hochmuth, C. A., Bartodziej, C., Schwägler, C. (2017). Industry 4.0. Is your ERP system ready for the digital era? https://www2.deloitte.com/content/dam/Deloitte/de/Documents/technology/Deloitte_ERP_Industrie-4-0_Whitepaper.pdf

Kagermann et al. (2013). Recommendations for implementing the strategic initiative Industrie 4.0. Technical report, National Academy of Science and Engineering, Germany, Available at: http://www.acatech.de/fileadmin/user_upload/Baumstruktur_nach_Website/Acatech/root/de/Material_fuer_Sonderseiten/Industrie_4.0/Final_report__Industrie_4.0_accessible.pdf

Kagermann H. (2015). Change Through Digitalization – Value Creation in the Age of Industry 4.0, [in:]: H. Albach et al. (eds.), Management of Permanent Change, Springer Fachmedien Wiesbaden, 23-45

Kletti, J. (2007). Manufacturing Execution Systems - MES. Springer, ISBN 978-3-540-49743-1

Kusiak, A. (2019). Fundamentals of smart manufacturing: A multi-thread perspective, Annual Reviews in Control, Volume 47, pp. 214-220. https://doi.org/10.1016 /j.arcontrol.2019.02.001

Li Da Xu, Wu He, and Shancang Li. (2014). Internet of things in industries: A survey. IEEE Transactions on Industrial Informatics, pages 2233-2243, doi: 10.1109/TII.2014.2300753

Longo, F., Nicoletti, L., Padovano, A. (2019). Ubiquitous knowledge empowers the Smart Factory: The impacts of a Service-oriented Digital Twin on enterprises' performance, Annual Reviews in Control, 47 (2019), 221-236

Lu, Y. (2017). Industry 4.0: A Survey on Technologies, Applications and Open Research Issues. Journal of Industrial Information Integration. Available at: http://www.sciencedirect.com/science/article/pii/S2452414X17300043.

Manuj, I., & Mentzer, J. T. (2008). Global supply chain risk management strategies. International Journal of Physical Distribution & Logistics Management, 38(3), 192-223. https://doi.org/10.1108/09600030810866986

Majstorovic, V., Stojadinovic, S., Lalic, B., Marjanovic, U. (2020). ERP in Industry4.0. International Conference on Advances in Production Management Systems (APMS), Novi Sad, Serbia, pp.287-294

Marhavilas, P. K., Koulouriotis, D., Gemeni, V. (2011). Risk analysis and assessment methodologies in the work sites: On a review, classification and comparative study of the scientific literature of the period 2000-2009. Journal of Loss Prevention in the Process Industries 24, 477-523

McKinsey. (2015). Industry 4.0 how to navigate digitization of the manufacturing sector. McKinsey Digital, Available at https://www.mckinsey.de/files/mck_industry_40_report.pdf

Mejía R., López A., Molina A. (2007). Experiences in developing collaborative engineering environments: an action research approach, Computers in Industry, 58(4), 329-346, Elsevier

Olson, D., Johansson, B., Carvalho, R. (2018). Open source ERP business model framework, Robotics and Computer–Integrated Manufacturing, 50, 30–36

Panetto, H., Molina, A. (2008). Enterprise Integration and Interoperability in Manufacturing Systems: trends and issues. Computers in Industry, 59 (7), pp.641-646

Pat-Cornell, M. E. (1996). Uncertainties in risk analysis: Six levels of Treatment Reliability Engineering and System Safety 54 (1996) 95-111, Elsevier Science Limited. Printed in Northern Ireland

Reis, M. S., Gins, G. (2017). Industrial Process Monitoring in the Big Data/Industry 4.0 Era: From Detection, to Diagnosis, to Prognosis. Processes, 5, 35

Roberto, M. A. (2009). Know what you don’t know: How great leaders prevent problems before they happen. Pearson Prentice Hall 10

Rudberg, Z. and Sandelin, 0. (2017). Impact on manufacturing execution systems through the use of smart connected devices. Supervisor: Andrea de Giorgio

Sadgrove, K. (2015). The Complete Guide to Business Risk Management. Third Edition. Routledge

Scheer, A.W. (1987). CIM – Computer Integrated Manufacturing, Springer Verlag

Štverková, H. (2021). Konkurenceschopný malý a střední podnik. Monography. [In Czech]. Vědecké nakladatelství veřejnoprávních a ekonomických studií v Uherském Hradišti Vysoká škola Jagielońská v Toruni. ISBN 978-80-88328-12-4

Van Marle, G. (2015). DP-DHL finally abandons ill-fated NFE IT project and is forced to write off €345m. *The Loadstar* [online]. [2020-12-19]. Available at: https://theloadstar.com/dp-dhl-finally-abandons-ill-fated-nfe-it-project-and-is-forced-to-write-off-e345m/

Wang S., Wan J., Li D., Zhang C. (2016). Implementing Smart Factory of Industrie 4.0: An Outlook, International Journal of Distributed Sensor Networks, Article ID 3159805, 10 pages, http://dx.doi.org/10.1155/2016/3159805

Wright, S. and Wright, A. M. (2002). Information System Assurance for Enterprise Resource Planning Systems: Unique Risk Considerations. Journal of Information Systems [online]. 16(s-1), 99–113. ISSN 0888-7985

Xu X. (2012). From cloud computing to cloud manufacturing, Robotics and Computer-Integrated Manufacturing, 28(1), pp. 75–86

A CASE STUDY EVALUATING THE CURRENT LEVEL OF USE AND FUTURE POTENTIAL OF DYNAMIC SIMULATION IN CZECH COMPANIES

Petra Kašparová, Kateřina Hušková

Technical University of Liberec
Faculty of Economics
petra.kasparova1@tul.cz, katerina.huskova@tul.cz

DOI: 10.35011/IDIMT-2024-241

Keywords

dynamic simulation; decision-making; business processes; case study

Abstract

This paper aims to determine the current usage of dynamic simulation (DS) as a decision-making method in selected Czech companies. The research was based on a case study that examined the use of DS in five companies of different sizes and from various business fields. Respondents expressed the current involvement of DS in designated business processes and evaluated the potential for future application as well. Based on the literature review, processes from the following departments were selected for the research: human resources, logistics, finance and controlling and R&D and production. The results show that DS is most often used in logistics and finance regardless of company size. At the same time, it also turned out that companies use the potential in defined processes only between 15-30 percent.

1. Introduction

DS is often defined as a process of creating logico-mathematical model of real process, system or decision making process and the associated possibility to implement a large number of experiments in this model (Dyntar, 2018). It is a versatile tool that can be applied across various disciplines to improve processes by providing insights into complex dynamics, fault identification, productivity enhancement, and business process optimization.

This paper aims to verify the current level of DS usage in selected Czech companies. Since this is the introductory part of long-term research, a case study (a combination of expository and comparative case studies) was chosen as the research method. Five enterprises of different sizes were nominated. Respondents were given a short form to fill out expressing the current use of DS in specific business processes in their organizations. Based on theoretical research, processes from the following areas were selected for evaluation: human resources, logistics, finance and controlling, R&D and production. Respondents also evaluated the potential of DS and assigned a level of possible future application within defined processes. Finally, both values were compared and the current relative usage rate (Up) against future potential is determined. The results of the conducted research thus aimed to answer two research questions:

RQ1: What is the current level of use of DS in Czech companies for selected business processes?

RQ2: Where do Czech companies see the greatest potential for future use of DS?

The answers to the defined questions are intended to provide the first insights into the current use of DS in Czech companies, enabling further identification of barriers that generally prevent more complex decision-making methods from being used in practice (Kašparová, 2022). Through simulation methods organizations can gain insights into their operations and make informed decisions for improvement. Repenning (Davis et al., 2007) highlights that simulation can elucidate the outcomes of interactions among organizational and strategic processes over time, providing a comprehensive view for enhancement. This is further supported by (Zhao et al., 2021), who discuss how simulation enables situational awareness of business processes, aiding in better understanding and management of complex business dynamics.

In business processes, DS offers several advantages that can lead to improved decision-making, process optimization, and overall efficiency. By utilizing DS, organizations can gain insights into the behaviour of complex systems and processes, allowing for a deeper understanding of how different variables interact over time (Dijkman et al., 2008). This enhanced understanding can aid in identifying bottlenecks, inefficiencies, and areas for improvement within business processes. Furthermore, DS enables the simulation of new business models, facilitating the study of different strategies and their potential outcomes (Yudha et al., 2019). By simulating business models dynamically, organizations can test various scenarios, assess risks, and make informed decisions regarding business model innovation.

2. Methodology

The research in this paper is based on a case study. Qualitative data collection is a suitable research method since it provides innovative insight into the research topic at the beginning (Takahashi & Araujo, 2019). Case studies are used as a research method in the social sciences, management, medicine and other fields that focus on the detailed analysis of individual cases, which can be people, groups, events, organisations or processes (Yin, 2018).

A combination of two types of case studies was utilised in this research: exploratory study (which aims to investigate new topics or phenomena that have not yet been sufficiently explored) and comparative study (which analyses and compares two or more cases to identify similarities and differences) (Hayes et al., 2015).

According to the specific methodology of creating case studies, the entire research was divided into the following steps (Mills et al., 2010):

Formulation of study objectives and research questions: The objectives were clearly presented in the introduction, and two specific research questions were defined to achieve the desired outcomes. Goals and RQs were formulated based on the literature search and previous similar research.

Choice of cases: The selection of cases required ensuring the relevance of individual companies and their diversity. Table 1 shows the organization’s size, the business field, and the job position of the representative who filled out the questionnaire. In order to obtain accurate and relevant data, all companies were promised anonymity when publishing the results.

Table 1. Characteristics of researched companies.

Comp.	Size	Business field	Representative
A	Small, start-up (< 50 employees)	IT	IT analyst
B	Medium (< 250 employees)	Construction	CEO's office manager
C	Medium (< 250 employees)	Automotive – after market	Lean specialist
D	Large (> 250 employees)	Automotive	Logistics manager
E	Large (> 250 employees)	Logistics, shipping	Quality manager

Source: own processing

Data collection: In case studies, this may include documents, interviews, observation, or statistical data analysis. We designed a questionnaire summarising the main areas of DS application in business processes. Bellow you can find most widely used application areas and the references used as a starting point. Particular examined processes are stated in chapter 3., Table2. Chosen areas of business processes are:

- Human resources (Chowdhury et al., 2023; Park & van der Aalst, 2021; Vrontis et al., 2022);
- Finance and controlling (Acheson et al., 2017; Murphy et al., 2020; Truong, 1999);
- Logistics (Huskova & Dyntar, 2022; Jiao et al., 2018; Kovalský & Mičieta, 2017; Wang et al., 2020);
- R&D; Production (Lidberg et al., 2020; Vieira et al., 2018).

Subsequently, four levels of engagement of DS within the application of the business processes were applied (Kašparová & Michalová, 2023):

- Status 0: Zero application, no awareness of the possibility of implementing DS.
- Status 1: Sporadic use of DS for specific tasks.
- Status 2: DS as a support of decision-making in the context of set objectives. DS application established in the organisation.
- Status 3: DS as a key source of input information, automatic implementation in processes.

In order to determine the current use of DS in various business processes and to determine the potential to be optimally used within the organisation, the respondents were asked to provide their opinion for both situations: current usage and future potential. For the sake of clarity, the recorded results are divided into two parts and the potentially possible applications contain the letter 'p (potential)' for easier identification (Table 3).

Data analysis: Firstly, data were summarized in one overview to obtain a general awareness of the current and potential use of DS in business processes across industries. Furthermore, as part of the data processing, the individual levels were assigned a point value according to the individual statuses (Status 0 = 0 points, Status 1 = 1 point, etc.) so that the obtained outputs could be expressed numerically in the next phase. Finally, both values (current level and potential) were compared and the relative usage rate (Up) of future potential is determined.

Interpretation of results and formulation of recommendations: Based on the data analysis, the study's results are interpreted. In the discussion, factors that may explain the differences between the cases are identified, and their implications in relation to the defined objectives and research questions are discussed. Finally, a study summary is presented, and recommendations for practical applications are made. Limits of research and possible future directions of follow-up research are also added.

3. Research results

Selected employees were asked to evaluate the current use of dynamic simulation in particular business processes. Furthermore, their task was to estimate the potential future using DS for these processes. The results of this evaluation are shown in Table 2.

Table 2. Results of qualitative research – current and potential engagement of DS.

Possibilities of use DS / Level of engagement of DS	**Status 0**	**Status 1**	**Status 2**	**Status 3**	**Status 0**	**Status 1**	**Status 2**	**Status 3**
	Current application				**Potential – optimal application**			
Human resources								
Simulation of work processes - training	A, B, C, D, E				Bp	Ap, Cp, Ep	Dp	
Optimization of the number of workers	A, D, E	C	B			Ap	Dp	Bp, Cp, Ep
Testing innovations and changes	A, B, D, E	C			Bp	Ap	Dp	Cp, Ep
Finance; Controlling								
Simulation of investment strategies	C, D, E	A		B			Cp, Dp, Ep	Ap, Bp
Risk behaviour simulation	C, D, E	A	B				Cp, Dp, Ep	Ap, Bp
Optimization of information flow	C, D, E	A	B				Cp, Dp, Ep	Ap, Bp
Testing innovations and changes	D	A, C, E	B				Dp	Ap, Bp, Cp, Ep
Logistics								
Warehouse workflow simulation	B, D	A, C, E			Bp		Dp	Ap, Cp, Ep
Simulation of transport networks, e.g. milk-runs	B, D	A, C, E			Bp		Dp	Ap, Cp, Ep
Demand Forecasting / Inventory Management	B, D	A, C, E			Bp			Ap, Cp, Dp, Ep
Testing innovations and changes	B, D	A, C, E			Bp		Dp	Ap, Cp, Ep

Distribution strategy	B, D	A, C, E			Bp		Dp	Ap, Cp, Ep
Optimization of material / production flow	B	A, C, D, E			Bp		Dp	Ap, Cp, Ep
Layout optimization	B	A, C, D, E			Bp		Dp	Ap, Cp, Ep
R&D; Production								
Risk Analysis (FMEA)	B, C, D, E	A			Bp, Dp	Ep		Ap, Cp
Simulation of new products, technologies	B, C, D, E	A			Bp	Dp, Ep		Ap, Cp
Testing production strategies	B, C, D, E	A			Bp	Dp, Ep		Ap, Cp
Calculation of the capacity of machines and equipment	B, C, E	A, D			Bp	Ep		Ap, Cp, Dp
Operational / capacity planning of the production	B, C, E	A, D			Bp	Ep		Ap, Cp, Dp

Source: own processing

Verbal answers were converted into numerical expression according to a designation score (0-3), see Table 3.

Table 3. Results in numeric expression.

Area of using DS / Corporate setting	**Current situation**						**Potential – possible future status**						**Up (%)**
	A	B	C	D	E	**Σ**	Ap	Bp	Cp	Dp	Ep	**Σ**	
Human Resources	0	2	2	0	0	**4**	3	3	7	6	7	**26**	15,4%
Finance / Controlling	4	9	1	0	1	**15**	12	12	9	8	9	**50**	30%
Logistics	0	0	7	2	7	**16**	21	0	21	15	21	**78**	21,4%
R&D; Production	5	0	0	2	0	**7**	15	0	15	8	5	**43**	16,2%
Sum of score	**9**	**11**	**10**	**4**	**8**	**42**	**51**	**15**	**52**	**37**	**42**	**197**	**21,3%**

Source: own processing

4. Discussion

The results made it possible to answer the defined research questions. The first was determined as follows:

RQ1: What is the current level of use of DS in Czech companies for selected business processes?

As can be seen from the outputs, Czech companies currently do not use DS much. Based on the point evaluation of the four levels, DS is most often used in the logistics sector. This value is driven by company E, which primarily operates in logistics and shipping goods, where DS is nowadays necessary to maintain a competitive advantage. And logistics in the monitored companies usually generates one of the the highest costs, which is why including more complex methods to support decision-making is increasingly common. In presented evaluation, logistics is followed by the area of finance and controlling, where in the application, it is the farthest constructing company B. The distribution of funds is crucial in the construction field due to the sharply increasing inputs.

The total usage (Up) of the declared potential represents only 21.3% of the monitored companies. The potential is best used in finance and controlling (30%). In contrast, the largest reserves were determined by HR (15,4%), R&D and production (16,2%). This finding is surprising, at least in the R&D department and in production. There is, therefore, a relatively high potential here, which companies have yet to fulfil. In the future, the use in production should follow a generally wider use in logistics.

Regarding the evaluation of the overall potential in particular areas, the second research question addressed this:

RQ2: Where do Czech companies see the greatest potential for future use of DS?

All in all, the respondents strongly identified logistics as the area with the greatest potential (78). DS is commonly used in logistics for various purposes such as optimizing supply chain operations, improving warehouse layouts, determining the most efficient routes for transportation, and evaluating the impact of changes in demand or infrastructure. DS allows logistics professionals to create virtual models of their operations and test different scenarios to identify potential bottlenecks, optimize resource allocation, and improve overall efficiency (Kovalský & Mičieta, 2017; Wang et al., 2020). Therefore, the achieved evaluation in this study is not such a big surprise. By simulating real-world conditions and interactions between different components of the logistics system, DS helps in making informed decisions and reducing risks associated with changes or investments in the logistics infrastructure (Vieira et al., 2018).

Logistics in potential was followed by the area of finance and controlling (50). DS is utilized in finance for a variety of purposes, including risk management, portfolio optimization, scenario analysis, and modelling complex financial systems. Overall, DS plays a crucial role in enhancing decision-making processes, managing risks, and improving the efficiency of financial systems and operations (Murphy et al., 2020).

DS can also play a significant role in R&D and production. The potential value reached 43. DS in this field helps in the development and optimization of products and processes by simulating their behaviour over time. Since the use of the declared potential, as mentioned above, is relatively low, it is apparently necessary to promote the wide possibilities presented across research in the business environment. This includes testing different design configurations, materials, and manufacturing methods to improve product performance, reliability, and cost-effectiveness (Lidberg et al., 2020). DS is also used for capacity planning to determine the optimal capacity of production facilities and resources required to meet demand while minimizing idle time and resource shortages. By simulating different scenarios and production schedules, manufacturers can make informed decisions about capacity expansion, resource allocation, and production investments (Vieira et al., 2018).

5. Conclusion

Based on the theoretical background and the potential expressed in this study, DS can play a crucial role in group decision-making processes by providing information, feedback, and visual representations of system dynamics. This can help management teams evaluate different scenarios, assess the impact of decisions, and make informed choices based on the simulation outcomes (Škraba et al., 2007). Additionally, DS can contribute to the reinvention of business models by incorporating acquisition-based dynamic capabilities and fostering innovation (Čirjevskis, 2017).

Respondents from our research are also apparently aware of these advantages (declared high values for future potential). However, one of the sub-goals of follow-up research should be to reveal barriers to further development. One crucial advantage of simulation is, thus, the ability to provide a detailed understanding of complex systems and processes (Childs et al., 2018). The main advantage of simulation, according to Kelton and Law (Law, 2015), is the possibility to choose the right variant, as it is possible to test the proposed variants without physical implementation. This will allow companies to save costs associated with the necessity of real testing of intended innovations. Furthermore, the simulation in the model enables the acceleration or deceleration of time, and thereby it is possible to examine a given part of the process in detail, or to skip it in case it is not interesting for decision-making. Last but not least, the simulation provides a better overview of the variables in the modelled process, their connections with each other, and it further enables easier work with them (Huskova & Dyntar, 2022).

On the other side, while DS offers valuable insights into business processes, it is essential to be aware of its limitations, including potential deviations from optimal behaviour, challenges in capturing real-world complexities, uncertainties in human-based activities, and the dynamic complexity of business environments. Among other limitations of this research, we can include a case study, which only gives a picture of the narrow sample examined. In the future, it is essential to supplement presented research with other methods, such as in-depth interviews or quantitative research involving a more comprehensive representation of companies from other fields and different sizes.

References

Acheson, C., Mackle, D., Murphy, A., Higgins, P., Collins, R., Higgins, C., Butterfield, J., Darlington, J., & Tame, R. (2017). Integrating Financial Metrics with Production Simulation Models. In Advances in Manufacturing Technology XXXI (pp. 567–572). IOS Press. https://doi.org/10.3233/978-1-61499-792-4-567

Childs, A. M., Maslov, D., Nam, Y., Ross, N. J., & Su, Y. (2018). Toward the first quantum simulation with quantum speedup. Proceedings of the National Academy of Sciences, 115(38), 9456–9461. https://doi.org/10.1073/pnas.1801723115

Chowdhury, S., Dey, P., Joel-Edgar, S., Bhattacharya, S., Rodriguez-Espindola, O., Abadie, A., & Truong, L. (2023). Unlocking the value of artificial intelligence in human resource management through AI capability framework. Human Resource Management Review, 33(1), 100899. https://doi.org/10.1016/j.hrmr.2022.100899

Čirjevskis, A. (2017). Acquisition based dynamic capabilities and reinvention of business models: Bridging two perspectives together. Entrepreneurship and Sustainability Issues, 4(4), 516–525. https://doi.org/10.9770/jesi.2017.4.4(9)

Davis, J. P., Eisenhardt, K. M., & Bingham, C. B. (2007). Developing Theory Through Simulation Methods. Academy of Management Review, 32(2), 480–499. https://doi.org/10.5465/amr.2007.24351453

Dijkman, R. M., Dumas, M., & Ouyang, C. (2008). Semantics and analysis of business process models in BPMN. Information and Software Technology, 50(12), 1281–1294. https://doi.org/10.1016/j.infsof.2008.02.006

Dyntar, J. (2018). Návrh a optimalizace dodavatelských systémů s využitím dynamické simulace (Praha). FinEco.

Hayes, R., Kyer, B., & Weber, E. (2015). The Case Study Cookbook. Worcester Polytechnic Instit. https://zbook.org/read/9daf9_the-case-study-cookbook-worcester-polytechnic-institute.html

Huskova, K., & Dyntar, J. (2022). Increasing Efficiency in Inventory Control of Products with Sporadic Demand Using Simulation. Acta Informatica Pragensia, 11(2), 254–264.

Jiao, Y., Xing, X., Zhang, P., Xu, L., & Liu, X.-R. (2018). Multi-objective storage location allocation optimization and simulation analysis of automated warehouse based on multi-population genetic algorithm. Concurrent Engineering, 26(4), 367–377. https://doi.org/10.1177/1063293X18796365

Kašparová, P. (2022). Intention to use business intelligence tools in decision making processes: Applying a UTAUT 2 model. Central European Journal of Operations Research. https://doi.org/10.1007/s10100-022-00827-z

Kašparová, P., & Michalová, T. (2023). SECI Knowledge Model and Opportunities of Engaging Business Intelligence by Maturity Level: Case Study at Selected Businesses in the Czech Republic. Journal of Intelligence Studies in Business, 13(2), 6–24. https://doi.org/10.37380/jisib.v13i2.1080

Kovalský, M., & Mičieta, B. (2017). Support Planning and Optimization of Intelligent Logistics Systems. Procedia Engineering, 192, 451–456. https://doi.org/10.1016/j.proeng.2017.06.078

Law, A. M. (2015). Simulation modeling and analysis (Fifth edition). McGraw-Hill Education.

Lidberg, S., Aslam, T., Pehrsson, L., & Ng, A. H. C. (2020). Optimizing real-world factory flows using aggregated discrete event simulation modelling. Flexible Services and Manufacturing Journal, 32(4), 888–912. https://doi.org/10.1007/s10696-019-09362-7

Mills, A. J., Durepos, G., & Wiebe, E. (2010). Encyclopedia of Case Study Research: L - Z; Index. SAGE.

Murphy, A., Taylor, C., Acheson, C., Butterfield, J., Jin, Y., Higgins, P., Collins, R., & Higgins, C. (2020). Representing financial data streams in digital simulations to support data flow design for a future Digital Twin. Robotics and Computer-Integrated Manufacturing, 61, 101853. https://doi.org/10.1016/j.rcim.2019.101853

Park, G., & van der Aalst, W. M. P. (2021). Towards Reliable Business Process Simulation: A Framework to Integrate ERP Systems. In A. Augusto, A. Gill, S. Nurcan, I. Reinhartz-Berger, R. Schmidt, & J. Zdravkovic (Eds.), Enterprise, Business-Process and Information Systems Modeling (pp. 112–127). Springer International Publishing. https://doi.org/10.1007/978-3-030-79186-5_8

Škraba, A., Kljajić, M., & Borštnar, M. K. (2007). The Role of Information Feedback in the Management Group Decision-Making Process Applying System Dynamics Models. Group Decision and Negotiation, 16(1), 77–95. https://doi.org/10.1007/s10726-006-9035-9

Takahashi, A. R. W., & Araujo, L. (2019). Case study research: Opening up research opportunities. RAUSP Management Journal, 55(1), 100–111. https://doi.org/10.1108/RAUSP-05-2019-0109

Truong, D. (1999). Application of discrete event simulation to the activity based costing of manufacturing systems. International Journal of Production Economics. https://www.academia.edu/1936816/Application_of_discrete_event_simulation_to_the_activity_based_costing_of_manufacturing_systems

Vieira, A. A. C., Dias, L. S., Santos, M. Y., Pereira, G., & Oliveira, J. A. (2018). Setting an industry 4.0 research and development agenda for simulation—A literature review. https://doi.org/10.2507/IJSIMM17(3)429

Vrontis, D., Christofi, M., Pereira, V., Tarba, S., Makrides, A., & Trichina, E. (2022). Artificial intelligence, robotics, advanced technologies and human resource management: A systematic review. The International Journal of Human Resource Management, 33(6), 1237–1266. https://doi.org/10.1080/09585192.2020.1871398

Wang, Q.-Z., Chen, J.-M., Tseng, M.-L., Luan, H.-M., & Ali, M. H. (2020). Modelling green multimodal transport route performance with witness simulation software. Journal of Cleaner Production, 248, 119245. https://doi.org/10.1016/j.jclepro.2019.119245

Yin, R. K. (2018). Case study research and applications: Design and methods (Sixth edition). SAGE.

Yudha, F., Raharja, S. D., & Rizal, M. (2019). SIMULATING A NEW BUSINESS MODEL: A DYNAMIC BUSINESS MODEL APPROACH STUDY ON PT. AMM POULTRY PARTNERSHIP COMPANY. Russian Journal of Agricultural and Socio-Economic Sciences, 92(8), 3–9. https://doi.org/10.18551/rjoas.2019-08.01

Zhao, X., Yongchareon, S., & Cho, N.-W. (2021). Enabling situational awareness of business processes. Business Process Management Journal, 27(3), 779–795. https://doi.org/10.1108/BPMJ-07-2020-0331

MANAGEMENT OF MANAGERIAL COMPETENCY

Jana Holá

University of Pardubice
Faculty of Health Studies
jana.hola@upce.cz

Lukáš Čegan

University of Pardubice
Faculty of electrical engineering and informatics
lukas.cegan@upce.cz

DOI: 10.35011/IDIMT-2024-249

Keywords

Competency Management; Nursing; managerial competencies

Abstract

Competency Manager is a software application designed to manage key competencies in nursing, providing systemic support to manage managerial competencies, which are also crucial for nursing but are currently undervalued. This application was developed within a project of applied research for a specific client, a health care facility. The architecture and significance of the software application have already been published. This article focusses on creating a model for managing managerial competencies, which, unlike other competency groups, were not subject to legislative norms, but aimed to meet the client's needs. The article outlines theoretical foundations and evidence-based methods (stakeholder engagement strategy, evidence-based implementation with change agents, discussion, brainstorming, and Delphi method for review within expert panel) leading to the selection, description, and implementation of 9 core managerial competencies within the created model.

1. Introduction

The Czech Republic, like many other European and other countries, has recently suffered from a lack of general nurses, especially in acute care hospitals in inpatient wards. A more effective setting of competencies for nurses could lead to better utilisation of their work capacity and solving nurse and physician shortages in the Czech healthcare system.

The shortage of nurse's overloads existing staff and places high demands on management. Therefore, as part of the stabilisation in recent years, the salaries of general nurses and other health workers have increased significantly. General nurses perceive higher financial remuneration as a motivating factor, but not as a major factor that influences job satisfaction. Job satisfaction is considered a crucial indicator of employee retention (Coomber and Barriball, 2007; Alameddine et al., 2017) related to management, teamwork, job competence, and perceived respect for one's work.

The results of our latest research (Glajchová et al., 2021) showed that management and teamwork are two of the main factors determining job satisfaction of general nurses. Nurses associate their job satisfaction with relationships with supervisors and the leadership style used to ensure the smooth running of the department. From a sample of 2,479 general nurses (out of a population of approximately 80,000), 27.9% reported being in a leadership position (of which the majority, 53%, were line manager nurses), and only 4% reported having received managerial training.

Most nurse managers on the front line do not have the necessary management training and are – from a qualification perspective – not prepared to perform a managerial function. The education and performance of a nurse is governed by legislation. For the preparation of nurse managers, a specialised study programme is offered at the Faculty of Health faculties, Organisation, and Management in Health Care. A general nurse with a bachelor’s degree can study this specialisation. A nurse with a bachelor's degree can also obtain management qualifications in master’s Health of Administration programs (often financially unavailable) or other study programs in other types of faculties. However, the acquisition of specialised or other managerial education is not a condition to perform a managerial function in most Czech hospitals.

As part of the Competent Nurse of the 21st Century project conducted in 2019-2021, supported by the Czech Technology Agency, we developed a software application for the management of general nurses’ competencies. Support and development of competencies strengthen the identities and professionalism of nurses and are a source of satisfaction / dissatisfaction and well-being of practice nurses in the UK (Wood, 2021). The study by Numminen et al. (2016) also revealed significant connections between commitment and competency settings, fluctuation intentions, and job satisfaction. The competence model we developed (Holá et al.,2022) is based on valid legislation regulating the professional competence of general nurses and on the results of research on the fulfilment of these competences in performing the profession. It is a model for more effective use of the working capacity of general nurses, in order to increase the efficiency of the education, training, development, and performance of nursing professions, especially in competencies that are not based on legislative standards. The competence encapsulates legislative norms, acquisition of professional specialisation, and competence assigned by the organisation (administrative, organisational, managerial, mentoring, etc.). The purpose of this article is to demonstrate how to create a model for managing managerial competencies in nursing within an acute care hospital.

2. Theoretical frame

The basic starting point for nursing management competencies could be the publication Leadership Competencies for Healthcare Services Managers from 2015 by the International Hospital Federation (IHF,2023). This consortium adopted the basic competency framework as a basis for the development of healthcare management with a recommendation for use in an academic environment to train future nurses.

Another suitable source for creating a set of managerial competencies was the American Organisation of Nurse Executives (2015), and the Management Competency User Pack for Nurses and Midwives (MCUP) presented by Office for Health Management, 2023) also proved to be a suitable basis. This user pack created by more than 70 Irish hospitals, offers great inspiration for setting up management competencies based on evidence based. The material is devoted to individual areas of competence, using examples of signals of their absence (indicators of less effective performance), the authors show their importance, and define specific application possibilities. The material is very extensive and detailed for individual competencies. Recommendations for setting competencies in practice through education and development are also an important part. The authors highlight their significance and also delineate specific application possibilities. The material is extensive and detailed with respect to each competency. An important part includes recommendations for setting competencies in practice through education and development.

In their research, the authors of the MCUP identified 8 general competency areas that are the basis for effective performance at all levels of nursing management and 13 additional critical competencies for different levels of management, namely three for frontline workers, five for middle management, and five for top managers. General competencies, see below.

- Support for Evidence-Based Decision Making
- Building and maintaining relationships
- Communication and Influencing Skills
- Initiation of innovations
- Resilience and Balance
- Integrity and ethical attitudes
- Sustained personal commitment.
- Professionalism

The authors also included the following areas in the competency categories for top management, for example strategic and systemic thinking, for middle management should primarily derive from leadership area, and for front-line management, the authors identified three significant competency areas: planning and organisation, team building and leadership, management and supervision of clinical practice, and quality of service delivery. Many authors agree and their work confirms that the dominant competency in management and leadership is communication and relationship-based management. This is affirmed by the competency model mentioned above the IHF (2015) which emphasises Communication and Relationship Management as one of 5 main competency areas, and, for example, it is also evidenced by the authors of the publication "Effective Nursing Leadership: Performing" (Cziraki et al., 2018), which emphasises the importance of managerial communication in nursing.

Communication competencies listed in the Gower Handbook of Internal Communication (Wright, 2016) could be a suitable starting point. In the chapter ‘What makes a competent communicator’, communication competences are defined as a sum of knowledge, skills, and experience, which are manifested not only in rhetoric and contemporary speech, but in the general actions and behaviour of a manager. The first most important competency is Building effective relationships.

3. Methodology and results

The purpose of creating a competency model and the Competency Manager software application was to effectively set up competency management at any stage of professional performance, specifically during the adaptation period, stable performance, and during re-assessment and development through lifelong learning. The competency model itself can serve as an appendix to job descriptions in accordance with national legislation, requirements for meeting standards, certificates of safe and quality care, and requirements for human resource management of healthcare facilities in practice.

The creation of the competency model (content) and its implementation in the form of the Competency Manager software application primarily relied on compiling the best practices found in the literature review and the needs of the application guarantor. By involving management and experts from the ranks of the application guarantor, the recommended stakeholder engagement strategy (those who will support the implementation of the result) according to (Gallangher-Ford et al., 2011) and change agents (experts who created what they will work with) in the preparation and implementation of the project according to the recommendations of the JBI Evidence Implementation Manual (Porritt et al., 2020) was applied.

The technological process of creating a software application and its deployment into the intranet environment of the application guarantor can be found in the publication by Holá and Čegan (2019). A detailed description of the creation of the application and its functions according to the client's requirements is given in the publication Holá et al. (2022).

An expert panel was formed as a working group consisting of nursing care experts from (the principal applicant guarantor and opponents from the ranks of the main representatives of other guarantors. This working group (11 experienced nurse managers with more than 20 years nursing experience and a minimum of 5 years of management position experience in clinical nursing, and 2 representatives of application guarantors as supervisors from Czech Association of Nurses and certification body for quality and safety of providing care) regularly met with the aim of defining the basic framework of the competency model, its content and the framework of conditions for its implementation. The group used methods such as moderated discussion, brainstorming, and mutual interaction, as well as principles of the Delphi method for collecting and sorting data, information, and knowledge. The system design idea is based on the definition of the system according to Hall and Fagen (1968), and the advantages of systemic setting are also declared within the recommendations of JBI Evidence Implementation (Porritt et al., 2020) and recommendation by Winter (2018) in The design of an Evidence-based Competency Model which also includes education and development of the given competency, its reassessment, and adoption as part of onboarding process and its integration into the education system and human resources management standards as part of management quality standards.

After analysing hospital needs, considering all best practices as results of the literature review and discussions within the activities declared above, the expert group selected nine key managerial competencies for nursing managers (72 competencies were chosen in total), which were further elaborated; see the example of building effective relationships shown in Fig.1.

- Financial management and economy
- Regular evaluation of workers
- Application of professionalism
- Personal management
- Managerial communication and relationship building
- Leadership
- Conducting safe and quality care
- Information management
- Strategic management

Regular Employee Evaluation and Development
This competence entails the consistent and thorough evaluation of nursing staff performance, coupled with active engagement in their professional development. The supervisor demonstrates adeptness in various aspects of evaluation, including tool selection and program design. They conduct regular evaluations, meticulously documenting protocols and ensuring the continuity of the process. Moreover, they conduct evaluation interviews aimed at enhancing key competencies and fostering individual career growth. They adeptly identify strengths, weaknesses, and developmental obstacles, motivating staff to set and achieve new goals. Collaboratively, they establish conditions for further development, aligning with the employee's expectations and organizational needs. Together with the employee, they devise educational plans for the future, adjusting based on assessment outcomes. Additionally, they assign, adjust, and reassess key competencies, influencing personal allowances based on performance evaluations
Assigns/verifies by Supervisor nurse
Acquisition Conditions: A minimum of 6 months in a management position coupled with internal training
Re-evaluation Frequency: Employee evaluations are conducted per 2 years
Mandatory/Recommended Education: Training in the adaptation process, completion of an external management course, specialized education in Organization and Management in Healthcare, pursuit of an MBA or managerial higher education, and participation in an internal course facilitated by a competency mentor, preferably from HR or a training background.
Mandatory/recommended education: Training in the adaptation process, external management course, specialized education Organization and management in health care, MBA, managerial higher education, internal course led by a competence mentor - HR or trainer
Compliance with internal standards: Adaptation process, employee evaluation, manual for working with SW Competence Manager

Figure 1. Example of key managerial competency description Source: (author)

All selected competencies must be regularly evaluated and developed; just as key competencies tied to job performance. In the case of adapting a new nurse manager, emphasis is placed on acquiring each competency. The education of nurse managers is tied to the development of these key competencies.

4. Discussion

The work of the Finnish team Pihlainen, Kivinen, and Lammintakanen (2016) noted the fundamental conclusion that the creation of a framework for the development of managers and leaders in hospitals is a necessary prerequisite not only for their own development, but for that of the entire organisation. This support must reflect the assessment of their role, as well as the implementation of changes and the response to current challenges in the delivery of health care. The authors state that the basic framework of managerial competencies must include, in addition to professionally orientated competences, also general competences of management and leadership. These competencies clearly belong to the strategic development of the organisation because the fulfilment of basic managerial competencies improves the motivation and performance of managers and the performance of the entire organisation. Even in Lenburg's (1999) basic nursing competency model, managerial and leadership competencies are an essential part of the competency model for nurses. The authors provide evidence of the importance of managerial competencies for team performance and creativity based primarily on managerial communication from their study (Boies, Fiset & Gill, 2015).

From this point of view, it is very positive that the authors of the competency model succeeded in creating a set of acceptable managerial competencies for nursing care in an acute care hospital, which until now only worked with professional competencies defined by legislation for the performance of the nursing profession. Especially because nearly 80% of nurse managers in the survey by Glajchová et al. (2021) report the absence of managerial education.

Managerial competencies are the competencies of a professional, and, like other competencies, they are based on a combination of soft (emotional) and hard (knowledge-based) skills. According to Holá (2017), management communication is a tool for advocating and achieving organisational goals and therefore is a highly significant competency and an important part of the internal communication system. For this reason, it is necessary to ensure adequate development of managerial communication competencies. For several years now, the Faculty of Health Studies has been providing the Managerial Communication in Nursing course, mainly for cooperating hospitals (Holá, Moravcová, Hlaváčková, 2020) and in response to demand the faculty opened a specialised master's program in Healthcare Organization and Management for nonmedical healthcare professions in 2022.

Although MCUP (2023) identified not only eight general competence areas that are the basis for effective performance at all levels of nursing management, it also defined additional critical competencies for different levels of management, as mentioned in the previous pages, in our case the selected competencies are the same for all levels of management and are perceived more generally than specifically. The client preferred this general setting, rejecting specific competencies such as critical thinking, and wanting to set competencies that are generally accepted and feasible in his environment. For the implementation of the entire competence model, it was very important that it be acceptable for practice. Although some unimplemented competencies and their implementation certainly belong to evidence-based nursing management, they cannot be "screwed" without primary system changes.

The competency model for managerial skills was developed within a public healthcare facility and can serve as an inspiration for other facilities to establish their own specific models, with a similar or different number and content of competencies. The significance of conceptualizing managerial competencies is also supported by the authors of the study "Exploring the management competencies of nurse managers in the Greater Accra Region, Ghana" (2020). Therefore, this issue should not be overlooked in Czech healthcare, approved by the Ministry of Health.

Nursing management education should be a priority for hospitals because it influences the quality of care provided, thus affecting the well-being of the entire society (International Council of Nurses (2020).

5. Conclusions

The creation and selection of managerial competencies was one of the most difficult parts of creating the content of the competency model. Managerial competencies, in contrast to professional competencies focused on nursing performance, are not described in any legislation, and the employer should define them himself. The importance of the influence of managers on employee performance is indisputable and supported by evidence in the Czech healthcare sector, or in acute care hospitals, management, as in many nonmedical fields, is underappreciated. However, staff shortages pose many challenges to management at all levels and will become increasingly significant. The absence of managerial education and the unconscious prioritisation of managers' intuitive, no systemic setting in the unstable healthcare sector are obstacles that our healthcare sector must overcome and with which the university can significantly help in education and research. The Competence Manager application is still in trial operation; however, managerial competencies are already being reflected in the setting of the competencies of nurse managers and in their development.

Acknowledgement

The Competent Nurse for the 21st Century: The Analysis and Design of the Optimization of Nursing Education and Professional Practice supported by Technology Agency of the Czech Republic TL01000094

References

Alameddine, M., Bauer, M. J., Richter, M. & Sousa-Poza, A. (2017). The paradox of falling job satisfaction with rising job stickiness in the German nursing workforce between 1990 and 2013. Human Resources for Health, 15(55). DOI: 10.1186/s12960-017-0228-x

American Organization of Nurse Executives. (2015). AONE Nurse Executive Competencies. Chicago, IL: Author. Available from: http://www.aone.org/resources/nurse-leader-competencies.shtml

Boies K., Fiset J., Gill H. (2015). Communication and trust are key: Unlocking the relationship between leadership and team performance and creativity. Leadership Quaterly. DOI: 10.1016/j.leaqua.2015.07.007.

Coomber, B. & Barriball, K.L. (2007). Impact of job satisfaction components on intent to leave and turnover for hospital-based nurses: A review of the research literature. International Journal of Nursing Studies, 44(2):297-314. DOI: 10.1016/j.ijnurstu.2006.02.004

Glajchová, A et al. (2021) Job satisfaction of general nurses in standard and intensive care units: "The nurse is a jack-of-all-trades!". Kontakt, 2021, roč. 23, č. 4, s. 289-296

Holá, J. et al. (2022). Competency Manager as a very useful competency management tool. IDIMT-2022; Digitalization of Society, Business and Management in a Pandemic. Linz: Johannes Kepler Universität, pp. 313-321. ISBN 978-3-99113-758-0

Holá, J., Moravcová, M., & Hlaváčková, E. (2020). Communication competency: The topic of lifelong learning for nurse managers in hospitals. Kontakt, 22(1), 33-39. DOI:10.32725/kont.2019.053

Holá, J. & Čegan, L. (2019). Information system architecture for competency model. IDIMT-2019: Innovation and Transformation in a Digital World. Linz: Trauner Verlag, pp. 369-376. ISBN 978-3-99062-590-3

Numminen, O., Leino-Kilpi, H., Isoaho, H. & Meretoja, R. (2016). Newly graduated nurses' occupational commitment and its associations with professional competence and work-related factors. Journal of Clinical Nursing, 25(1-2):117-126. DOI: 10.1111/jocn.13005

International Hospital Federation (2023). Global Healthcare Leadership Competency Model. Available from: https://ihf-fih.org/what-we-do/global-healthcare-leadership-competency-model/

International Council of Nurses (2020). Nurses: A Voice to lead Nursing the Word to Health. Available from: https://www.icn.ch/resources/publications-and-reports/nurses-voice-lead-nursing-world-health

International Hospital Federation. Leadership Competencies for Healthcare Services Managers (https://ihf-fih.org/)

Lenburg c. B., klein c., abdur-rahman v., spencer t. & boyer, s. (2009) The COPA model: A Comprehensive Framework Designed to Promote Quality Care and Competence for Patient Safety. Nursing Education Perspective 30(5), pp 312–317

Ofei, A. M. A., Paarima, Y., & Barnes, T. (2020). Exploring the management competencies of nurse managers in the Greater Accra Region, Ghana. International Journal of Africa Nursing Sciences, 13. https://doi.org/10.1016/j.ijans.2020.100248

Porritt K, McArthur A, Lockwood C, Munn Z (Editors). JBI Handbook for Evidence Implementation. JBI, 2020. Available from: https://implementationmanual.jbi.global. https://doi.org/10.46658/JBIMEI-20-01

Office For Health Management (2023). Management competency user pack: for nurse and midwife managers. Office for Health Management. Available from https://www.lenus.ie/handle/10147/42677.

Raven, J. (2011). Competence, education, professional development, psychology, and socio-cybernetics. In G.J. Neimeyer (ed.). Continuing education: Types, roles, and societal impact. New York: Nova Science Publisher, Inc. ISSN 11788-3619.

Pihlainen V., Kivinen T., & Lammintakanen J. (2016). Management and leadership competence in hospitals: a systematic literature review. Leadership in Health Services. Emerald Group Publishing Ltd. DOI 10.1108/LHS-11-2014-0072.

Winter, P.B. (2018) The design of an Evidence-based Competency Model. Journal for Nurses in Professional Development. Vol 34/4, 206-211. ISSN 3169-9798

Wood, E. et al. (2021). Sources of satisfaction, dissatisfaction and well- being for UK advanced practice nurses: A qualitative study. Journal of Nursing Management, 29(5):1073-1080. DOI: org/10.1111/jonm.13245

Wright, M. (2016). Gower Handbook of Internal Communication. Farnham: Gower Publishing Ltd. 2009. ISBN: 978-1-3155-8569-7

MESSAGE BROADCAST METHODS FOR THE INTERNET OF THINGS

Jan Čapek, Dawit Dejene Bikila

University of Pardubice
Faculty of Economics and Administration
capek@upce.cz, dawitdejene.bikila@student.upce.cz

DOI: 10.35011/IDIMT-2024-257

Keywords

Internet of Things (IoT); messaging techniques; message queuing; communication models

Abstract

Ambient Intelligence and the Internet of Things (IoT) have shown noticeable growth in recent years in changing people's daily lifestyles. The Internet of Things creates a vast network of interconnected devices, sensors, and actuators that can collect, transmit, analyse, and use data to make intelligent decisions and take action. We can say that the IoT is the materialisation or manifestation of AmI that integrates this imaginary world with the real world through a common platform. IoT organises an environment which can be considered intelligent and independent, such as smart cities, things, health, or even intelligent life. The IoT communication model refers to several ways, in which IoT devices connect and share data. In this regard, we have observed from the literature that only the two methods have been commonly used for IoT messaging. Hence, in this work, we propose a messaging broadcast system using Producer-Consumer and Reader-Writer methods. A study with discussion and analysis provided and indicates the reasonability and applicability of this proposed approach in IoT-based smart environments.

1. Introduction

Ambient Intelligence (AmI) and IoT have shown a noticeable growth in recent years that changes people's daily lifestyles. AmI is an environment that uses technology to create intelligence and intelligent environments that are aware of our needs and can respond to them automatically (Anastasopoulos, M. et al., 2005; Kissoum, Y. et al., 2014). The Internet of Things (IoT) can be denoted using the following general representation: IoT = Data (or information) + Sensors + Networks + Services. We can say that the IoT is the materialisation or manifestation of AmI. This makes our physical surroundings be seamless and personalised environment.

IoT devices are usually based on affordable wireless communication interfaces through which they can communicate and transmit information to other IoT devices or a centralised system. With the rapid development of these technologies, everyday routines of life are centred on virtual space, a virtual world as indicated by Razzaq, M. A. et al. (2017). As a result, users can then shop, work, plant, or keep animals in this virtual world while physically living in the real world.

The interconnection and interaction of IoT devices have altered the way we live by providing increased flexibility and convenience through IoT applications. With the rapid growth and a huge number of devices connecting to IoT, 75 billion things are expected to be connected by 2025. As a result, a large volume of real-time data will be generated and transmitted via the Internet (Khashan & Khafajah, 2023). Without IoT, it would be difficult or impossible to create smart and intelligent environments that AmI envisions.

However, replacing human activities is difficult to fully automated way. IoT is used to organise an environment that is intelligent and independent, such as intelligent cities, things, health, or even intelligent life (Abomhara & Køien, 2014), Figure 1 presents the general concept of the IoT, including its capabilities.

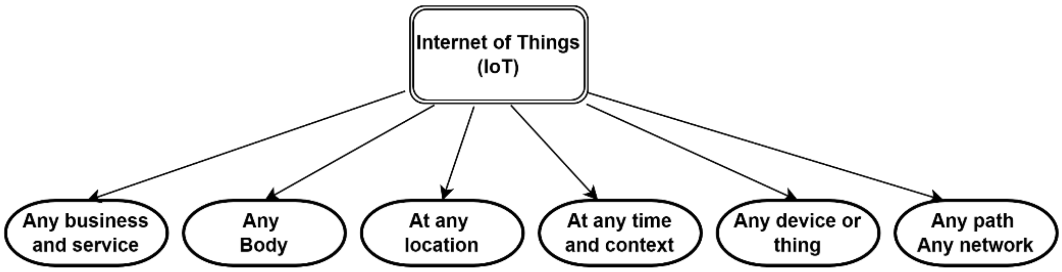


Figure 1. General concept of IoT modified Source: (Razaq M.A. et al. (2017))

Communication with IoT devices can be done via Wi-Fi, Bluetooth, or communication protocols. Hence, IoT messaging methods are widely used to enable message sharing among these devices. Consequently, several communication methods have been proposed and used frequently. However, employed methods at any layer of IoT architecture resulted in synchronisation and interoperability issues that need greater concern (Abdelouahid et. al., 2021). The main aim of this paper is to show the possibilities of message broadcast methods for IoT.

The paper is organised as follows. Section 2 presents related works regarding IoT messaging systems. Section 3 explains the background and methodology. Then, section 4 will discuss the result and analysis of the proposed work. Finally, section 5 will present the discussion and conclusion.

2. Related works

Tallberg (2020) defined message queuing systems are used to store data in IoT applications and allow devices to connect to the queueing system. These systems can improve stream processing using asynchronous processing, and decoupling of producers and consumers, another study by Nguyen et al. (2019) concluded that compatibilities, such as storage and synchronised processing, in IoT big data platforms and with other tools should be considered.

In another study by Fu et al. (2021) two modes, Push and Pull, were briefly explained. In Push mode, messages are pushed continuously to the consumer with better real-time performance, but it demands a flow control mechanism. In pull mode, message requests are based on a time interval but setting a reasonable pull interval is difficult. Continuous pull requests will impose a significant load on the queueing system. However, if the pull requests are not continuous and timely, they will cause latency problems. Thus, synchronisation will play a great role in solving such a problem.

IoT message queueing models are divided into the following categories based on their communication models: Request and response, Publisher subscriber, and polling (Domínguez-Bolaño et al., 2022). Choosing the right communication protocol that ensures security, reliability, and efficient data transmission is a challenge in IoT applications (El Ouadghiri et al., 2020). Generally, Amjad et al. (2021) discussed the two basic categories of IoT communication protocols Device to Device(D2D); protocols such as Data Distribution Service (DDS) used for independent smart device communication, and Device to Server (D2S) protocols.

Ghotbou and Khansari (2021) stated that Constrained Application Protocol (CoAP) uses a request-response architecture to exchange messages in IoT. Still, Seoane et al. (2021) reported that CoAP has no mechanism for congestion control, confirming message delivery, which results in possible errors. In another study by Amjad et al. (2021) Message Queuing Telemetry Transport (MQTT) is a protocol that uses the publisher-subscriber model. It is currently used in several IoT applications. According to Mishra and Kertesz (2020), it is suitable for resource-constrained devices with reduced delay and low bandwidth. However, it doesn't support any encryption method.

Authors da Cruz et al. (2019) introduced a gateway for the application layer in HTTP, called MiddleBridge. The MiddleBridge uses packets smaller than those sent by an IoT device. However, simultaneous one-to-many communications is not supported (Šikić, L. et al., 2020). Authors Uy and Nam (2019) noted that the Advanced Message Queueing Protocol (AMQP) provides long-lasting queues and low-latency transmission. However, this model has limitations such as high bandwidth, unsuitable for real-time systems, and not supporting automatic resource discovery (Bang et al., 2022).

Authors Lohitha and Pounambal (2023) proposed a cloud-based IoT framework by integrating push-pull and publisher-subscriber methods for real-time IoT data sharing. They stated that streaming audio and video using this model does not support synchronous end-to-end communication. Further, Longo and Redondi (2023) state that the publisher-subscriber is a single broker model. Hence, a single point of failure and managing real-time data generated by IoT is a problem. They proposed a modified distributed broker resulting in reduced latency and network traffic. But they didn't consider the synchronisation problem.

Lu and Da Xu (2018) show capabilities in messaging within IoT in the context of smart house solutions. These solutions pose a challenge for the elderly in creating a safe and secure home environment to reduce stress, fear, falls, or social isolation. Anastasopoulos et. al. (2005) indicated that IoT devices are connected to the Internet within an ambient intelligent home environment using different communication protocols. Each communication protocol shall be suitable for every use, depending on several key factors.

IoT message queuing techniques need to be interoperable, power-efficient, fast, low latency, reliable, scalable, and robust in security. Further, based on the literature overview the response-request and publisher-subscriber schemes are mostly used. Thus, this work will propose message queueing models using Producer-Consumer and Reader-Writer.

3. Materials and Methods

3.1. IoT Messaging Techniques

IoT messaging techniques are the various methods by which IoT devices can communicate and exchange data with each other. This research reliability was ensured by using model triangulation. Some of the most common IoT communication models are displayed in Table 1.

Table 1. Internet of Things Requirements and Protocol

Protocol	*Transport*	*Messaging*	*2G,3G,4G (1000's)*	*LowPower and Lossy (1000's)*	*Compute Resources*	*Security*	*Success Stories*	*Arch*
CoAP	UDP	Rqst/Rspnse	Excellent	Excellent	10Ks/RAM Flash	Medium - Optional	Utility field area ntwks	Tree
Continua HDP	UDP	Pub/Subsrb Rqst/Rspnse	Fair	Fair	10Ks/RAM Flash	None	Medical	Star
DDS	UDP	Pub/Subsrb Rqst/Rspnse	Fair	Poor	100Ks/RAM Flash +++	High-Optional	Military	Bus
DPWS	TCP		Good	Fair	100Ks/RAM Flash ++	High-Optional	Web Servers	Client Server
HTTP/ REST	TCP	Rqst/Rspnse	Excellent	Fair	10Ks/RAM Flash	Low-Optional	Smart Energy Phase 2	Client Server
MQTT	TCP	Pub/Subsrb Rqst/Rspnse	Excellent	Good	10Ks/RAM Flash	Medium - Optional	IoT Msging	Tree
SNMP	UDP	Rqst/Response	Excellent	Fair	10Ks/RAM Flash	High-Optional	Network Monitoring	Client-Server
UPnP		Pub/Subscrb Rqst/Rspnse	Excellent	Good	10Ks/RAM Flash	None	Consumer	P2P Client Server
XMPP	TCP	Pub/Subsrb Rqst/Rspnse	Excellent	Fair	10Ks/RAM Flash	High-Manditory	Rmt Mgmt White Gds	Client Server
ZeroMQ	UDP	Pub/Subsrb Rqst/Rspnse	Fair	Fair	10Ks/RAM Flash	High-Optional	CERN	P2P

Source: (Kim and David (2016))

For further investigation, according to Table 1, we focused on message broadcast methods. Table 1 shows different communication protocols for message transmission, the Publish-Subscriber and/or Request-Response techniques are used preferably. The Producer-Consumer and Reader-Writer that originate from operating systems will be discussed later.

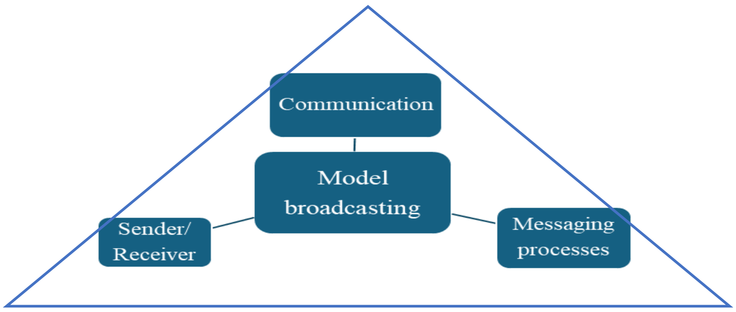


Figure 2. The triangulation model of broadcasting messages between IoTs

3.2. Publisher-Subscriber model

This messaging method involves one or more publishers, one or more subscribers, and a broker. In this technique, publishers produce messages and send them to a broker, which distributes the messages to one or more subscribers according to their interests. The Publisher-Producer model works using the following steps:

1. The publisher sends its message to the broker,
2. The broker receives the message and stores it,
3. The subscriber expresses interest by subscribing to specific topics or channels. A topic is a label or tag publishers use to categorise their messages,
4. The broker delivers messages to subscribers after it checks the subscribers' interests.

The subscriber processes the message and performs some other action based on the content of the message. This model is widely used in distributed messaging systems and event-driven architectures asynchronously. It decouples publishers and subscribers and allows them to operate independently and at different rates. The Publisher sends the messages to the broker, and then the broker forwards them to everyone who requested them (subscribers). No messages are shared among devices directly without a broker (message distributer).

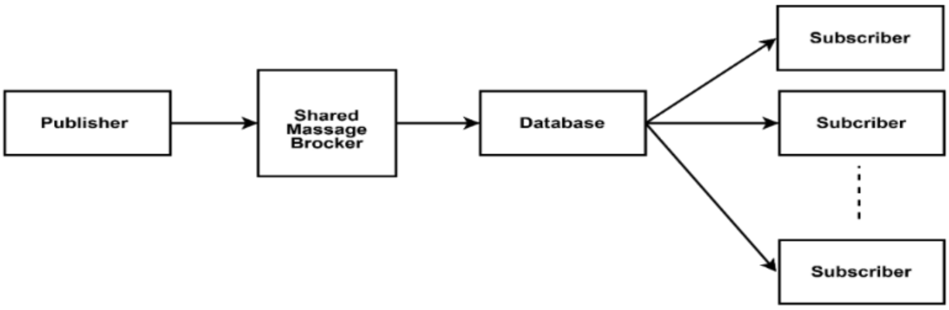


Figure 3. General communication model: Publisher-Subscriber

3.3. Request-Response Model

This model is a common way of communication between two entities. It consists of two parts, the client and the server. Messages are requested by the client and responses are generated and sent to the client by the server. The request-response messaging model works using the following steps:

1. The client sends the request to the system to trigger the communication. The request typically contains information about what the client wants the server to do or what data it needs,
2. The server processes the request and generates a response. The server may need to perform some computations, access a database or other resources, or perform some other operations to generate the response,
3. The server sends a response. The response typically contains the requested data or information about whether the requested operation was successful,
4. The client processes the response and the client may proceed with the next step.

In a request-response communication process, one sender and one receiver typically exist. This model is widely used in client-server systems, such as web applications. The request-response model allows efficient communication between the client and server, as each entity can focus on its specific roles and responsibilities.

The number of senders and receivers in a request-response communication process can be increased if multiple clients need to send requests to the same server. In this case, the server can handle multiple requests simultaneously using techniques such as multithreading or asynchronous programming.

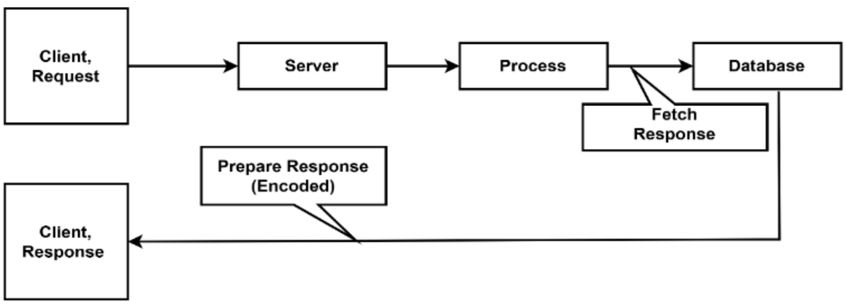


Figure 4. Request-response relationship of the general messaging model

As additional messaging techniques, methods originating from the operating system environment can be used. They are, for example, producers-consumers or writers-readers.

4. Result and Analysis

4.1 Producer-Consumer

The producer-consumer model is a common design pattern used in computer programming (operating systems). In this model, one or more "producer" devices generate data, and one or more "consumer" devices process data. The producers and consumers are decoupled so that each can operate independently and at its own pace. They are usually separate threads or processes, and communication between them is typically through the shared queue. The consumer retrieves the data or tasks from the queue and processes them.

The IoT producer-consumer model is a framework in which data is collected by IoT devices and then consumed by various applications. In this model, IoT devices act as data producers, while applications or systems act as consumers. The producer-consumer messaging model works using the following steps:

1. Date generation by IoT devices,
2. Data Transmission,
3. Data Storage,
4. Consumer applications,
5. Data Consumption.

The producer-consumer model in IoT environments can be implemented in various ways. One common approach is to use a message queue, where producers publish messages and consumers subscribe to those messages. This allows producers and consumers to operate asynchronously without requiring them to be connected at the same time.

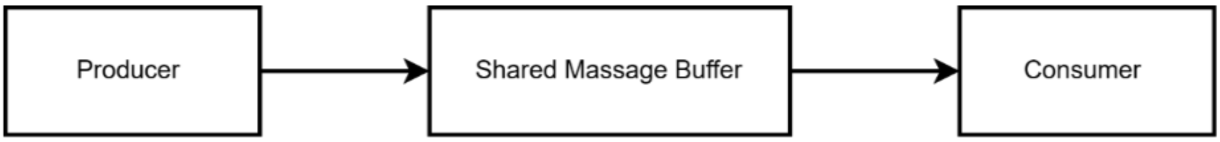


Figure 5. General Messaging Model of Producer-Consumer

4.2. Reader-Writer Model

The IoT reader-writer model is another method used in IoT environments to manage access to shared resources. Readers are entities that only read data and writers are entities that read and write data from the shared resource. The goal of this model is to ensure that multiple readers can access resources that are shared simultaneously. At a time only one writer can access the resource to prevent conflicts and ensure consistency of data.

The reader-writer messaging model works using the following steps:

1. The writer creates the message,
2. Multiple readers can access the message parallelly,
3. No readers can access the message when a writer is waiting for the resource,
4. No writer can access the resource when a reader is waiting for the message.

Two processes that share the same buffer, are not allowed to read the buffer at the same time. Such two processes are parallel mutual exclusion and sequential mutual exclusion. To implement the reader-writer model in an IoT environment, we need to synchronize the reading and writing process. This ensures that only one writer can publish messages on the topic at a time while allowing multiple readers to receive messages simultaneously. The IoT reader-writer model is a powerful tool for managing shared resources in IoT systems.

Table 2. Show how many senders and receivers one can use in each messaging method.

Basic messaging process	Senders	Receivers
Request-Response	One sender	One receiver
Publish-Subscribe	One or multiple publishers	Multiple Subscribers
Producer-Consumer	One or multiple producers	One or multiple consumers
Reader-Writer	One writer	Multiple readers

As indicated by Sethi and Sarangi (2017), the IoT is a three-layer architecture, namely the perception, network, and application layers. They are described as follows:

- The perception layer involves collecting data from various sensors, devices, and actuators. It includes devices embedded with sensors to sense, connect, and interact with the real world,
- The network layer is responsible for data messaging or transmission. If message transmission and receiving are not handled properly, a synchronisation problem occurs. This problem should be solved according to operating systems concepts semaphores, critical section, mutual exclusion, etc. (Stallings, 2021),
- The application layer defines various applications in which the IoT can be deployed. This layer is responsible for application-specific service delivery to the user.

5. Conclusion

Message queuing is a technique that plays an important role in data sharing among IoT devices in smart environments, where efficient data sharing is required. Due to the heterogeneous nature of IoT, message queuing techniques need to be interoperable, power-efficient, fast and low latency, reliable, scalable, and robust in security. The key findings of this study indicate the possibilities of IoT messaging techniques using operating system concepts Producer-consumer and reade-writer schemes. These messaging techniques allow flexibility in handling varying data generation and processing speeds, and asynchronous communication. This helps in managing resource constraints, when a fast producer generates data faster than a slow consumer can process, or vice versa. Moreover decoupling between the producer and consumer enables modularity and enabling scalable and loosely coupled architectures.

It's important to note that these messaging techniques mentioned in this article, can be used in combination or tailored to the specific requirements of an IoT system, considering factors such as device capabilities, network constraints, security considerations, and scalability needs. These messaging techniques facilitate efficient communication and data exchange between IoT devices, applications, and services, enabling seamless integration and interoperability in IoT ecosystems.

Acknowledgement

This paper was supported by grant No. SGS_2024_017, Faculty of Economics and Administration, University of Pardubice, Czech Republic.

References

Abdelouahid, R. A., Debauche, O., & Marzak, A. (2021). Internet of Things: a new interoperable IoT platform. Application to a smart building. Procedia Computer Science, 191, 511-517.

Abomhara, M., & Køien, G. M. (2014, May). Security and privacy in the Internet of Things: Current status and open issues. In 2014 international conference on privacy and security in mobile systems (PRISMS) (pp. 1-8). IEEE.

Amjad, A., Azam, F., Anwar, M. W., & Butt, W. H. (2021). A systematic review on the data interoperability of application layer protocols in industrial IoT. IEEE Access, 9, 96528-96545.

Anastasopoulos, M., Niebuhr, D., Bartelt, C., Koch, J., & Rausch, A. (2005, October). Towards a reference middleware architecture for ambient intelligence systems. In ACM conference on object-oriented programming, systems, languages, and applications.

Bang, A. O., Rao, U. P., Visconti, A., Brighente, A., & Conti, M. (2022). An IoT inventory before deployment: a survey on iot protocols, communication technologies, vulnerabilities, attacks, and future research directions. Computers & Security, 123, 102914.

da Cruz, M. A., Rodrigues, J. J., Lorenz, P., Solic, P., Al-Muhtadi, J., & Albuquerque, V. H. C. (2019). A proposal for bridging application layer protocols to HTTP on IoT solutions. Future Generation Computer Systems, 97, 145-152.

Domínguez-Bolaño, T., Campos, O., Barral, V., Escudero, C. J., & García-Naya, J. A. (2022). An overview of IoT architectures, technologies, and existing open-source projects. Internet of Things, 20, 100626.

El Ouadghiri, M., Aghoutane, B., & El Farissi, N. (2020). Communication model in the Internet of Things. Procedia Computer Science, 177, 72-77.

Fu, G., Zhang, Y., & Yu, G. (2020). A fair comparison of message queuing systems. IEEE Access, 9, 421-432.

Ghotbou, A., & Khansari, M. (2021). VE-CoAP: A constrained application layer protocol for IoT video transmission. Journal of Network and Computer Applications, 173, 102855.

Kim, R., & David, L. (2016),” Internet of Things Requirements and Protocol”, IEEE Standards University E-magazine.

Khashan, O. A., & Khafajah, N. M. (2023). Efficient hybrid centralized and blockchain-based authentication architecture for heterogeneous IoT. Journal of King Saud University-Computer and Information Sciences, 35(2), 726-739.

Kissoum, Y., Kerraoui, S., & Boughaouas, M. L. (2014). Smart Home for Elderly: Modeling and Simulation. In ICAASE (pp. 148-155).

Lohitha, N. S., & Pounambal, M. (2023). Integrated publish/subscribe and push-pull method for cloud based IoT framework for real-time data processing. Measurement: Sensors, 27, 100699.

Longo, E., & Redondi, A. E. (2023). Design and implementation of an advanced MQTT broker for distributed pub/sub scenarios. Computer Networks, 224, 109601.

Lu, Y., & Da Xu, L. (2018). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103-2115.

Mishra, B., & Kertesz, A. (2020). The use of MQTT in M2M and IoT systems: A survey. IEEE Access, 8, 201071-201086.

Nguyen, C. N., Lee, J., Hwang, S., & Kim, J. S. (2019). On the role of message broker middleware for many-task computing on a big-data platform. Cluster Computing, 22, 2527-2540.

Razzaq, M. A., Gill, S. H., Qureshi, M. A., & Ullah, S. (2017). Security issues in the Internet of Things (IoT): A comprehensive study. International Journal of Advanced Computer Science and Applications, 8(6), 383.

Seoane, V., Garcia-Rubio, C., Almenares, F., & Campo, C. (2021). Performance evaluation of CoAP and MQTT with security support for IoT environments. Computer Networks, 197, 108338.

Sethi, P., & Sarangi, S. R. (2017). Internet of Things: Architectures, protocols, and applications. Journal of Electrical & Computer Engineering.

Šikić, L., Janković, J., Afrić, P., Šilić, M., Ilić, Ž., Pandžić, H., ... & Džanko, M. (2020, September). A comparison of application layer communication protocols in IoT-enabled smart grid. In 2020 International symposium ELMAR (pp. 83-86). IEEE.

Stallings, W. (2021). Operating systems: internals and design principles. Pearson.

Tallberg, S. (2020). A Comparison of Data Ingestion Platforms in Real-Time Stream Processing Pipelines.

Uy, N. Q., & Nam, V. H. (2019, December). A comparison of AMQP and MQTT protocols for Internet of Things. In 2019 6th NAFOSTED Conference on Information and Computer Science (NICS) (pp. 292-297). IEEE.

DIGITAL TRANSFORMATION AND DIGITAL BUSINESS MODELS

ORGANIZATIONAL ASPECTS OF DIGITAL TRANSFORMATION IN SLOVENIAN ENTERPRISES

Andreja Pucihar, Mirjana Kljajić Borštnar, Gregor Lenart

University of Maribor
Faculty of Organizational Sciences
andreja.pucihar@um.si, mirjana.kljajic@um.si, gregor.lenart@um.si

DOI: 10.35011/IDIMT-2024-269

Keywords

Digital transformation; organizational aspects; digital culture; enterprises; Slovenia

Abstract

This paper investigates organizational aspects that affect the digital transformation of Slovenian enterprises. We conducted a survey with 131 participating companies to gather data on this topic. The findings suggest that while basic digital skills are relatively widespread among the enterprises, advanced digital skills are less common. Respondents generally rated the digital culture positively, noting its emphasis on innovation, open communication, collaboration, and autonomy. However, less than half of the surveyed enterprises have a formal strategy for developing digital competencies, and only about a third are actively working on advanced digital skills for business model innovation. Given the connection between advanced digital skills, technology adoption, and innovation, enterprises should address these gaps by investing in skill development. This would enhance their digital transformation efforts and, in turn, their overall competitiveness.

1. Introduction

Digital transformation (DT) is an increasingly important for ensuring the competitiveness of enterprises. DT is defined as the process in which enterprises use digital technologies (often their combination) to simplify operations, increase efficiency, change the way employees work, establish improved relationships with customers, suppliers, and partners, innovate business models, products, and services (Jeansson & Bredmar, 2019; Pucihar et al., 2021; Vial, 2019; Warner & Wäger, 2019). Digital technologies have become a strategic resource for enterprises in recent years (Nadkarni & Prügl, 2020; Pucihar et al., 2021; Zammuto et al., 2007). It is important to understand that in order to successfully and effectively exploit the opportunities offered by DT, organizations must establish appropriate organizational capabilities in addition to technological infrastructure, which include the necessary competencies and knowledge of employees (Pucihar et al., 2021), different leadership styles (Foerster-Metz et al., 2018; He et al., 2023; Ko et al., 2022; Pucihar et al., 2021; Vijay Gurbaxani & Debora Dunkle, 2019), and a different organizational culture (digital culture) (Li et al., 2018; Liu et al., 2011; Pucihar et al., 2021). Without these and a proper strategy, enterprises will not be able to capture the advantages of DT (Li et al., 2018; Pucihar, 2020; Vijay Gurbaxani & Debora Dunkle, 2019). In order to plan for DT, it is important for enterprises to understand their current state of DT. In this paper, we address organizational aspects of DT, which are often overlooked in research of DT. Research was done among enterprises in Slovenia in collaboration with Chamber of Commerce and Digital Innovation Hub Slovenia (DIHS). For the purpose of research, we designed a questionnaire and conducted a survey among 131 enterprises. Results are interesting for researchers and practitioners, as well as for decision makers, which can provide appropriate measures to accelerate DT of the economy and society.

2. Organizational Aspects of Digital Transformation

Organizational capability is defined as the ability of an enterprise to execute coordinated tasks using organizational resources to achieve goals. Organizational capabilities play an important role in achieving the goals of DT. A crucial aspect of organizational capabilities is dynamic capabilities, which enable enterprises to respond to rapid changes in the environment and contribute to gaining a competitive advantage (Konopik et al., 2022). Previous research has shown the importance of digital competencies, creativity, and innovation capabilities as important organizational aspects in DT (El Sawy et al., 2016; Muehlburger & Koch, 2019; Pucihar et al., 2021). Furthermore, organizational culture is crucial in fostering innovation, creativity, an entrepreneurial mindset, learning from mistakes, encouraging the generation of new ideas, accepting risk and experimentation, competitiveness, management support for change, and constructive conflict management as central organizational values (Bärenfänger & Otto, 2015; Kane et al., 2015; Muehlburger & Koch, 2019). Collaboration within the enterprise and involving stakeholders outside the enterprise in the innovation process (open innovation) also significantly contribute to the digital culture of an enterprise (Chesbrough, 2006; El Sawy et al., 2016; Hylving, 2015; Muehlburger & Koch, 2019). A suitable organizational culture is an essential foundation for successful DT (Hanelt et al., 2020; Nadkarni & Prügl, 2020). In recent years, in addition to business strategy, enterprises have also been establishing a digitalization strategy to plan their DT (Hanelt et al., 2020; Kane et al., 2015; Pucihar et al., 2021). The role and support of management are of high importance in setting and achieving organizational goals, which also applies to planning and implementing digital transformation (Kane et al., 2015; Muehlburger & Koch, 2019; Pucihar et al., 2021). The increasing opportunities brought by digital technologies and the dynamics of the business environment dictate different roles for management. It is important for management to establish a common understanding that digital technologies represent a strategic resource for the enterprise and that they establish a leader for the area for their effective utilization. In practice, we can observe different roles under the common denominator of digital leadership (e.g., digitalization manager, digital transformation manager, innovation and digital transformation manager, etc.) (El Sawy et al., 2016; Hanelt et al., 2020; Muehlburger & Koch, 2019; Nadkarni & Prügl, 2020; Pucihar et al., 2021; Tijan et al., 2021).

3. Research methodology

For the purpose of the study, we prepared a questionnaire, which was developed in collaboration between researchers from the Faculty of Organizational Sciences, University of Maribor, representatives of the Slovenian Chamber of Commerce - ICT Horizontal Network, DIHS and other experts in the field. The questionnaire consists of 43 questions. The survey was conducted in 2022 and 2023. The content of the questionnaire was tested with 10 respondents. The questions relate to information about the respondent and the enterprise, as well as of organizational and technological aspects of DT. In this paper, we focus on organizational aspects. The questionnaire was created using the 1ka tool, which allows for the online establishment of a questionnaire. The link to the questionnaire was made available publicly. Invitations were sent through mailing lists from participating stakeholders.

4. Results

4.1 Data on surveyed companies

The survey included 131 enterprises from various industries. The largest percentage was represented by ICT enterprises (28%), followed by enterprises from other diverse business activities (16%), manufacturing activities (8%), professional, scientific, and technical activities (8%), and other activities (8%), agriculture (6%), and construction (5%). The largest percentage of participants were micro-companies (56%), followed by small (26%), medium-sized (10%), and large (8%) companies. The survey was mostly answered by directors (49%), followed by department managers (20%), IT specialists (9%), IT managers (8%), and employees in other positions.

4.2 Strategy of digital transformation and its implementation

In half of enterprises, the responsibility for leading DT is of director, in 14% of enterprises they have dedicated role of head of digitalization, and in 10% the head of information technology. Other enterprises mentioned other responsible individuals such as the head of marketing, project manager, project leader, CEO, and development director. Only 22% of surveyed enterprises have a formally defined DT strategy, more than half of enterprises do not have formally defined DT strategy (54%), and in 24% of surveyed enterprises, the strategy is currently being developed. Half of the enterprises are implementing DT.

4.3 Digital competences

More than half of enterprises have a strategy for education, training, and developing digital competencies (41%). Almost half of the enterprises (48%) believe that their employees have advanced knowledge and skills, which they continuously upgrade. Basic digital competencies are present in 35% of enterprises. In 17% of enterprises, employees do not have sufficient digital competencies. Regarding providing education, the majority of enterprises enable employee participation in conferences, seminars, and workshops (55%), followed by self-education (44%), and regular internal employee trainings (such as knowledge transfer, mentoring, etc.) (37%). 31% of enterprises organize regular trainings with external providers and provide employees with time and financial resources for education.

Enterprises were also asked about the types of digital competencies they develop. In more than half of the enterprises, they develop general digital literacy such as working with data and basic software solutions (58%), communication and collaboration through digital channels (56%), and creating digital content (50%). This is followed by the development of competencies in information and cyber security (47%), keeping up with the latest digital trends (45%), and problem solving using digital technologies (43%). Enterprises place less emphasis on developing competencies for innovation and digitalization of business models (37%) and IT infrastructure management (28%).

4.4 Employee feedback on the adoption of new technologies

In 66% of enterprises, employees are involved in the planning and implementation of new technologies. In 14% of enterprises, employees are informed about the process of implementing new technologies, while in 21% of enterprises, employees are only informed about the innovations after they have been introduced.

In less than half of the surveyed enterprises, employees embrace new digital technologies and are ready to work with them (48%). In approximately one-fifth of enterprises, employees are enthusiastic about new technologies and are willing to participate in their implementation. A small percentage of surveyed enterprises have employees who show no interest in adopting new digital technologies (6%). The most common reasons for employees rejecting new technologies are lack of awareness of opportunities, fear of job loss, lack of knowledge, additional workload from training and job tasks.

4.5 Digital culture

Survey respondents rated the digital culture in enterprises very positively. The prevailing opinion is that enterprises are constantly seeking new opportunities brought by digitalization (70%), encouraging open communication (74%), promoting innovation and learning from mistakes (68%). Leadership makes decisions based on data and the opinions of other leaders, taking into account the views of employees (64%). Enterprises also promote and reward collaboration among employees (60%), who are dedicated to their work and are willing to work outside of regular working hours if needed (66%). They are also autonomous in their work (67%).

5. Discussion and Conclusions

The survey results showed that in half of the enterprises, the responsibility for leading DT lies with the director. Only 14% of enterprises have a designated department head for DT, which reduces the opportunities for comprehensive utilization of DT opportunities. Interestingly, more than half of the enterprises do not have an established DT strategy, while only approximately one-fifth of enterprises have formally defined it. Half of the surveyed enterprises stated that they are already implementing DT. The area of DT encompasses the entire organization. It involves the identification of opportunities for using digital technologies and planning and allocating resources to achieve organizational goals. The digital strategy connects and aligns business strategy with other organizational strategies, including IT strategy, which must be coordinated with each other to achieve results (Levstek et al., 2022; Pucihar et al., 2021).

Despite the fact that competitiveness today is based on innovation and the ability to leverage digital technologies (Hanelt et al., 2020; Pucihar et al., 2021), where knowledge plays a crucial role, we find that less than half of the enterprises have a strategy for education, training, and development of digital competences. The most popular methods of employee education are conferences, seminars and workshops, as well as self-learning. Enterprises primarily develop general digital literacy, which includes working with data and basic software solutions, communication and collaboration through digital channels, and creating digital content. Only about one-third of enterprises develop more advanced digital competences for innovating business models. The data is comparable to the data from the Statistical Office of the Republic of Slovenia, which found that almost one-third of enterprises lack the necessary staff or skills needed for DT (SURS, 2021). In the future, more attention will need to be given to the development of advanced digital competences in order to establish innovation culture, which is a key to competitiveness.

The respondents rated the level of established digital culture in enterprises relatively high. However, most of questionnaires were completed by managerial employees, and their responses to the questions may be subjective. Another limitation of this study is that the study was conducted on a relatively small sample of enterprises that followed the invitation through e-mailing lists. To better understand the organizational aspects and their impact on DT, it would be beneficial to conduct in-depth case studies of successful DT in enterprises.

Acknowledgment

The University of Maribor, Faculty of Organizational Sciences, and the authors would like to thank the Public Agency for Research Activities of the Republic of Slovenia for their financial support (research program No. P5-0018).

References

Bärenfänger, R., & Otto, B. (2015). Proposing a capability perspective on digital business models. Proceedings of the 2015 IEEE 17th Conference on Business Informatics, CBI '15, 17–25.

Chesbrough, H. (2006). Open Innovation: Researching a New Paradigm. Oxford: Oxford University Press.

El Sawy, O. A., Kræmmergaard, P., Amsinck, H., & Vinther, A. L. (2016). How LEGO built the foundations and enterprise capabilities for digital leadership. MIS Quarterly Executive, 15(2), 141–166.

Foerster-Metz, U. S., Marquardt, K., Golowko, N., Kompalla, A., & Hell, C. (2018). Digital Transformation and its Implications on Organizational Behavior. Journal of EU Research in Business, 2018, 1–14. https://doi.org/10.5171/2018.340873

Hanelt, A., Bohnsack, R., Marz, D., & Marante, C. A. (2020). A Systematic Review of the Literature on Digital Transformation: Insights and Implications for Strategy and Organizational Change. https://doi.org/10.1111/joms.12639

He, Z., Huang, H., Choi, H., & Bilgihan, A. (2023). Building organizational resilience with digital transformation. Journal of Service Management, 34(1), 147–171. https://doi.org/10.1108/JOSM-06-2021-0216

Hylving, L. (2015). Competing Values in the Era of Digitalization. Proceedings of the 48th Hawaii International Conference on System Sciences 2015 (HICSS), 4161–4170.

Jeansson, J., & Bredmar, K. (2019). Digital Transformation of SMEs: Capturing Complexity. In A. Pucihar, M. Kljajić Borštnar, R. Bons, J. Seitz, H. Cripps, & D. Vidmar (Eds.), 32nd Bled eConference. Humanizing technology for a sustainable society (pp. 523–541). University of Maribor Press.

Kane, G. C., Palmer, D., Philips, A. N., Kiron, D., & Buckley, N. (2015). Strategy, Not Technology, Drives Digital Transformation - Becoming a Digitally Mature Enterprise. MIT Sloan Management Review and Deloitte University Press.

Ko, A., Fehér, P., Kovacs, T., Mitev, A., & Szabó, Z. (2022). Influencing factors of digital transformation: management or IT is the driving force? International Journal of Innovation Science, 14(1), 1–20. https://doi.org/10.1108/IJIS-01-2021-0007

Konopik, J., Jahn, C., Schuster, T., Hoßbach, N., & Pflaum, A. (2022). Mastering the digital transformation through organizational capabilities: A conceptual framework. Digital Business, 2(2).

Levstek, A., Pucihar, A., & Hovelja, T. (2022). Towards an Adaptive Strategic IT Governance Model for SMEs. Journal of Theoretical and Applied Electronic Commerce Research, 17(1), 230–252. https://doi.org/10.3390/jtaer17010012

Li, L., Su, F., Zhang, W., & Mao, J. Y. (2018). Digital transformation by SME entrepreneurs: A capability perspective. Information Systems Journal, 28(6). https://doi.org/10.1111/isj.12153

Liu, D. Y., Chen, S. W., & Chou, T. C. (2011). Resource fit in digital transformation: Lessons learned from the CBC Bank global e-banking project. Management Decision, 49(10), 1728–1742. https://doi.org/10.1108/00251741111183852

Muehlburger, M., & Koch, S. (2019). A Framework of Factors Enabling Digital Transformation. 1–10.

Nadkarni, S., & Prügl, R. (2020). Digital transformation: a review, synthesis and opportunities for future research. Management Review Quarterly. https://doi.org/10.1007/s11301-020-00185-7

Pucihar, A. (2020). The digital transformation journey: content analysis of Electronic Markets articles and Bled eConference proceedings from 2012 to 2019. Electronic Markets, 30(1). https://doi.org/10.1007/s12525-020-00406-7

Pucihar, A., Marolt, M., Lenart, G., & Vidmar, D. (2021). Digitalna preobrazba in njeno stanje v organizacijah v Sloveniji. In Znanstveno-raziskovalni trendi na področju digitalne preobrazbe (pp. 9–44). University of Maribor, University Press. https://doi.org/10.18690/978-961-286-509-2

Pucihar, A., Mohar Bastar, K., & Lenart, G. (2022). Organizacije v Sloveniji na poti digitalne preobrazbe. Znanstveno-Raziskovalni Izzivi Na Poti Digitalne Preobrazbe, 1–20. https://doi.org/10.18690/um.fov.6.2022.1

SURS. (2021). Digitalno podjetništvo, podrobni podatki, Slovenija, 2021 Podjetja v vzhodni Sloveniji izkazujejo nižji digitalni indeks.

Teece, D. J. (2018). Business models and dynamic capabilities. Long Range Planning, 51(1), 40–49. https://doi.org/10.1016/j.lrp.2017.06.007

Tijan, E., Jović, M., Aksentijević, S., & Pucihar, A. (2021). Digital transformation in the maritime transport sector. Technological Forecasting and Social Change, 170. https://doi.org/10.1016/j.techfore.2021.120879

Vial, G. (2019). Understanding digital transformation: A review and a research agenda. The Journal of Strategic Information Systems, 28(2). https://doi.org/10.1016/j.jsis.2019.01.003

Vijay Gurbaxani, & Debora Dunkle. (2019). Gearing Up For Successful Digital Transformation. MIS Quarterly Executive, 18(3). https://aisel.aisnet.org/misqe/vol18/iss3/6

Warner, K. S. R., & Wäger, M. (2019). Building dynamic capabilities for digital transformation: An ongoing process of strategic renewal. Long Range Planning, 52(3). https://doi.org/10.1016/j.lrp.2018.12.001

Zammuto, R. F., Griffith, T. L., Majchrzak, A., Dougherty, D. J., & Faraj, S. (2007). Information Technology and the Changing Fabric of Organization. Organization Science, 18(5), 749–762. https://doi.org/10.1287/orsc.1070.0307

TECHNOLOGICAL, ORGANIZATIONAL AND ENVIRONMENTAL FACTORS FOR DIGITAL AND SUSTAINABLE DEVELOPMENT

Doroteja Vidmar, Andreja Pucihar

University of Maribor
Faculty of Organizational Sciences
doroteja.vidmar@um.si, andreja.pucihar@um.si

DOI: 10.35011/IDIMT-2024-275

Keywords

Digital technologies; TOE factors, digitalization, digital transformation, sustainable development

Abstract

Digitalization and sustainability are crucial for enterprises to maintain competitiveness and address changing market, societal and regulatory demands. This study explores the factors influencing the adoption of digitalization and sustainability within business models, focusing on enterprises in Slovenia. A survey of 198 companies was conducted to assess the impact of technological factors, organizational culture, and business environment conditions on strategic changes in business models. Results indicate that a majority of Slovenian enterprises are prepared for digital transformation, with more than half ready for changes in business models linked to digitalization and sustainable development. These findings offer valuable insights for both enterprise-level decision-makers and policymakers, highlighting the importance of ongoing adaptation to digital technologies and sustainability. However, the study's limitations include a focus on managerial perspectives, suggesting the need for further research involving a broader range of employees and exploring other countries' practices.

1. Introduction

Digitalization and the ever-changing local and global environment are presenting new challenges for enterprises. Environmental, social, political, healthcare, and other challenges have repeatedly shaken the societies in recent years. Enterprises aiming to maintain their competitive advantage need to continuously adapt to changing conditions. To thrive amidst customer demands, competition, and regulatory pressures, strategic use of digital technologies is needed (Kane et al., 2017, 2018). Digital technologies play a pivotal role in revolutionizing business practices through information management, automation, and transformation. This powerful combination of digital capabilities can drive positive changes, benefiting the economy, environment, and society. Digital technologies open new opportunities for enterprises (Vial, 2019), but prior research shows that most implementations of new digital technologies in practice is challenging and often fail (Andriole, 2021; Defossez et al., 2020). Enterprises need to prepare for the implementation and effective use of digital technologies by ensuring appropriate IT infrastructure, skilled employees, as well as that the enterprise is fostering an enabling organizational culture that is conducive to change and ensuring that they have sufficient human and financial resources allocated to implement digital technologies (Defossez et al., 2020). Successful implementation of digital technologies is still challenging for many enterprises, especially SMEs (Tijan et al., 2021).

The business environment is constantly evolving, and policymakers recognize the vital importance of governing sustainability in a digitalized environment. Digital technologies have evolved into a cornerstone of innovation, enabling new ways to create value, deliver services, and manage resources efficiently (Jose et al., 2019; Bressanelli et al., 2018; Camacho-Otero et al., 2018). The transformative power of digital technologies in not just in reducing costs but also in optimization of resource utilization and distribution. Such innovative digital business models illustrate how data harnessed through digital technologies, such as user behavior patterns and usage data, can be harnessed to uncover and mitigate excess resource consumption and wastage (Evans et al., 2017; Hanelt et al., 2017; Hildebrandt et al., 2018).

Digital and sustainability transformation are of strategic importance for the European Union (EU), which is also reflected in newly adopted legislation, projects and initiatives promoted and funded by both the EU and Slovenia (European Commission, 2017). To promote sustainable development and digitalization in enterprises, it is crucial to understand the factors that influence the introduction of change in business operations, taking into account the differences between SMEs (Marolt et al., 2020; Pucihar et al., 2019).

In this paper, we discuss the impact of technological factors, organizational culture and business environment factors on the adoption of strategically important changes (of both sustainability and digitalization) in business models. For this purpose, we have prepared and conducted a survey among SMEs and large enterprises operating in Slovenia.

The survey covered technological factors, organizational culture and business environment factors that are important for strategic changes in enterprises' operations. We based the included factors on the existing literature and preliminary interviews that were conducted in several Slovenian enterprises (Vidmar, 2021).

The results provide important information about factors influencing on digital transformation for practitioners and decision makers and fill the gap in the existing literature, which lacks different perspectives in research of digital and sustainable transformation.

2. Methodology

For this research, we employed a survey approach. The survey was conducted in SMEs and large enterprises operating in Slovenia. The data was collected using an online survey system 1ka. Surveys are often used to obtain data from a larger sample of respondents, and the results allow for generalization based on analysis (Emerald Publishing, 2021).

The questionnaire consisted of 4 sets of 19 questions in total – a set of demographic questions of the enterprise and individual respondent and 3 sets of questions focusing on the overall state of technological factors, organizational culture and business environment affecting the enterprise. The responses to questions on the impact of technological factors, organizational culture and business environment factors were measured using a 5-point Likert scale to measure the level of agreement with the statements.

For data collection, we used an online tool 1ka (1ka - En Klik Anketa, 2023). 1ka allows us to design a questionnaire (including different types of questions, with limitations and caveats), collect responses from participants, and partially prepare and export the data for analysis. Contact details of Slovenian enterprises were obtained from the Business Registrar of Slovenia (PRS) (AJPES EPRS - Poslovni Register Slovenije, 2023). Invitations to participate in the survey were sent to all available e-mail addresses enterprises via e-mail. A reminder was sent four days later. Once the data collection was completed, a statistical analysis of the data was carried out.

3. Results

Invitations to participate in the survey were sent to the email addresses of 3,567 small, medium-sized and large enterprises in Slovenia. The sample consisted of 2,900 small enterprises (81,30%), 491 medium-sized enterprises (13,77%) and 176 large enterprises (4,93%). We received 198 completed questionnaires, representing a response rate of 5.6%. Some questions in the survey were not answered by all participants, which is indicated by a non-response rate (marked with n/a).

For a few categorical questions, where the options included "other", we have added a text box for clarification. Where possible, responses in the other category were subsequently coded using a description provided by the participant.

3.1 Demographic data of the participants and enterprises

Among the 198 participating enterprises, 128 (64.6%) were small, 46 (23%) medium-sized and 23 (12%) large, with 1 missing value (0.5%).

Of the 198 participating enterprises, 17.2% are engaged in manufacturing, 15.2% in trade, maintenance and repair of motor vehicles, 12.6% in construction, 12.1% in professional, scientific and technical activities and 11.1% in information and communication activities. The frequency for the other categories was below 10, or 5%, and 6 enterprises could not be classified based on the additional explanation and the standard classification of operations ("Standardna Klasifikacija Dejavnosti – SKD 2008," 2008), and are therefore listed in the category other activities.

The majority of the respondents were management representatives, with 41.9% of the respondents being the CEO, 17.2% of the respondents being the CIO, 6.6% of the respondents being chief of strategic projects, 2.5% of the respondents being chief of R&D, and 0.5% of the respondents being in charge of sustainability. Additionally, 12.6% were in charge of other departments, and 2.5% were either owners or proxies. The rest of the participants were professional staff from different departments, or their job titles could not be determined from the description. 1 of the participants did not respond to this question.

3.2 Technological, organizational, and business factors

Agreement with the statements on the impact of technological, organizational, and business factors was measured on a 5-point Likert-type scale with possible response values ranging from 1 (strongly disagree) to 5 (strongly agree). For each question, the frequencies and percentages of individual responses and the total number of responses to the question are presented below. Respondents were asked to evaluate the overall state of the enterprise, and as such evaluate the overall state of all employees in the enterprise they represent.

Table 1 presents the questions related to technological factors and their impact on the enterprise's performance. The employees in the majority of the enterprises included in this study are open to the use of new technologies, with a slightly lower, but still good, level of employee knowledge of the use of modern technologies. The technological infrastructure in most studied enterprises is suitable for upgrading to new technologies, and enterprises are largely aware of the importance of new technologies in opening new business opportunities.

More than 70% of enterprises are continuously investing in modern technologies to develop their business. The data show a strong willingness of enterprises to adopt modern technologies to bring about changes in their business.

Table 1. Technological factors

Answer	Strongly disagree	Disagree	Nor agree, nor disagree	Agree	Strongly agree	Total
Employees are open to the use of modern technologies						
Frequency	1	8	21	92	53	175
Percentage	1%	5%	12%	53%	30%	100%
Employees have the necessary skills to use modern technologies						
Frequency	4	22	51	66	31	174
Percentage	2%	13%	29%	38%	18%	100%
Existing technology in the organization allows for upgrading to modern technology						
Frequency	0	8	12	99	55	174
Percentage	0%	5%	7%	57%	32%	100%
Using modern technologies opens up new business opportunities						
Frequency	0	5	6	68	94	173
Percentage	0%	3%	3%	39%	54%	100%
We continuously invest in modern technologies to develop our business, products and services						
Frequency	1	7	33	70	63	174
Percentage	1%	4%	19%	40%	36%	100%

Source: (author)

Table 2 presents organizational factors and their impact on the enterprise's business performance.

The majority of the enterprises surveyed have a culture favourable to introducing change into their business, although almost a quarter of enterprises are undecided about introducing change, and a smaller proportion of enterprises are not favourable to introducing change into their business.

Similarly, in most of the enterprises surveyed, the management of the enterprises encourages the introduction of change into the business, promotes clear communication with employees and involves them in strategic decision-making.

Most enterprises have sustainability and digitalisation objectives defined in their strategy, as well as allocated financial and human resources to implement change.

The data in Table 2 show that the organizational culture in more than half of the enterprises surveyed is well prepared for introducing change in their business.

Table 2. Organizational factors

Answer	Strongly disagree	Disagree	Nor agree, nor disagree	Agree	Strongly agree	Total
The organizational culture is supportive of introducing business change						
Frequency	2	21	40	79	33	175
Percentage	1%	12%	23%	45%	19%	100%
The leadership promotes change in the business						
Frequency	1	6	23	80	65	175
Percentage	1%	3%	13%	46%	37%	100%
The leadership promotes clear communication with employees						
Frequency	0	5	27	83	60	175
Percentage	0%	3%	15%	47%	34%	100%
The employees have the opportunity to shape strategic decisions						
Frequency	0	11	39	96	29	175
Percentage	0%	6%	22%	55%	17%	100%
The sustainability objectives of the enterprise are defined in its strategy						
Frequency	1	11	36	86	41	175
Percentage	1%	6%	21%	49%	23%	100%
The digitalisation objectives of the enterprise are defined in its strategy						
Frequency	3	12	41	75	44	175
Percentage	2%	7%	23%	43%	25%	100%
Enterprise has allocated resources (financial, human resources) to implement change						
Frequency	2	17	50	78	28	175
Percentage	1%	10%	29%	45%	16%	100%

Source: (author)

Table 3 presents the factors related to the enterprise's business environment and their impact on the enterprise's business performance. Table 8, which presents data on the impact of business environment factors on the implementation of changes in the business of the enterprises studied, shows that more than half of the enterprises feel pressure to implement changes in their business from competitors, business partners and customers, while at least 30% of the enterprises remain neutral to the pressure from the external environment, and to a lesser extent the enterprises do not feel any pressure from the external environment.

The impact of regulation on the implementation of sustainability and digital change is slightly lower. A third of enterprises perceive the impact of regulation as neutral, while 41% of enterprises are encouraged to implement sustainable and digital change by EU and national incentives.

60% of enterprises have already obtained certifications confirming that their business complies with various standards, and 62% of enterprises feel that their business is heavily influenced by existing legislation.

Overall, the participating enterprises showed lower agreement with the statements on the impact of the business environment on the introduction of changes to their business compared to the statements on the impact of technology and organizational culture.

Table 3. Business environment factors

Answer	Strongly disagree	Disagree	Nor agree, nor disagree	Agree	Strongly agree	Total
The enterprise is feeling pressured to make changes by competitors						
Frequency	2	27	53	69	24	175
Percentage	1%	15%	30%	39%	14%	100%
The enterprise is feeling pressured to make changes by business partners						
Frequency	1	24	66	66	18	175
Percentage	1%	14%	38%	38%	10%	100%
The enterprise is feeling pressured to make changes by customers						
Frequency	1	19	58	78	18	174
Percentage	1%	11%	33%	45%	10%	100%
National and EU incentives for sustainable business transformation (e.g. sustainability, digital) influence the enterprise's activities						
Frequency	11	33	57	53	20	174
Percentage	6%	19%	33%	30%	11%	100%
The enterprise ensures that its operations comply with certifications or standards (e.g. ISO, OHSAS standards, certifications such as Family Friendly Business, etc.)						
Frequency	10	23	37	61	44	175
Percentage	6%	13%	21%	35%	25%	100%
The enterprise's operations are heavily regulated or subject to legislation						
Frequency	8	21	37	64	44	174
Percentage	5%	12%	21%	37%	25%	100%

Source: (author)

4. Conclusions

In this paper, we presented the results of a study that used a survey questionnaire to monitor the impact of organizational, technological, and business environment factors related to the implementation of digital technologies and their impact on changes of business models in the of small, medium-sized, and large enterprises. 198 companies from Slovenia participated in the study.

The data obtained show that more than half of the enterprises are well prepared for the implementation of modern digital technologies and for the introduction of changes in their business models related to digitalisation and sustainable development.

Based on the results, we estimate that enterprises operating in Slovenia are aware of the opportunities of modern digital technologies to develop new business opportunities.

The organizational culture in the participating enterprises is also well prepared for the implementation of new digital technologies and business changes. There is a slight lag in the readiness of employees for implementation of digital technologies and their digital skills. Surprisingly, the data suggest that enterprises are slightly less affected by customers, competitors, and business partner, while strategizing changes of their operations. To a slightly higher degree, enterprises are burdened by the impact of regulation or are adapting their business to existing legislation.

As many as 60% of companies already have in place certifications of compliance with various standards, and around 70% of companies have defined sustainability and digitalisation objectives as part of their business strategy.

A clear sustainability and digitalisation strategy, combined with the introduction of certificates, a high level of incentives from management and clear communication from management to employees, indicate that the implementation of digital technologies will continue in the coming years.

Based on the results, we estimate that the participating enterprises are aware of the importance of adapting their business to the changing needs of the market, both in terms of technology and organizational culture. Most of the participating enterprises are ready to implement changes in the areas of digitalisation and sustainable development, and the data also suggest that these enterprises will further develop their efforts in the future.

Results are important for decision makers on the level of enterprise and state, who can gain important insights into the current situation and perceptions of enterprises related to readiness for implementation of digital technologies and sustainable development. However, in the survey, the respondents were mainly managers and thus results reflect their perceptions. Therefore, this is one of the limitations of this research. To avoid this, case studies (or a survey) could be conducted with interviews with more employees from different businesses and hierarchies in the enterprises. Research could also be conducted in other countries.

Acknowledgement

This research was supported by the Slovenian Research Agency: Program No. P5-0018—Decision Support Systems in Digital Business and postgraduate research funding for Young Researchers No. 1000-20-0552.

References

1ka - en klik anketa. (2023). Https://Www.1ka.Si/d/Sl.

AJPES ePRS - Poslovni Register Slovenije. (2023). Https://Www.Ajpes.Si/Prs/.

Andriole, S. (2021, March 25). 3 Main Reasons Why Big Technology Projects Fail – & Why Many Companies Should Just Never Do Them. Forbes. https://www.forbes.com/sites/steveandriole/2021/03/25/3-main-reasons-why-big-technology-projects-fail---why-many-companies-should-just-never-do-them/?sh=62295fe4257c

Bressanelli, G., Adrodegari, F., Perona, M., & Saccani, N. (2018). Exploring How Usage-Focused Business Models Enable Circular Economy through Digital Technologies. Sustainability, 10, 639.

Camacho-Otero, J., Boks, C., & Nilstad Pettersen, I. (2018). Consumption in the Circular Economy: A Literature Review. Sustainability, 10, 2758.

Defossez, K., McMillan, M., & Vuppala, H. (2020, November 25). How to avoid large technology-program failures McKinsey. McKinsey Digital. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/managing-large-technology-programs-in-the-digital-era

Emerald Publishing. (2021). How to... design a survey. Https://www.Emeraldgrouppublishing.com/Archived/Research/Guides/Methods/Survey_design.Htm

European Commission. (2017). A concept paper on digitisation, employability and inclusiveness. The role of Europe. (Issue May).

Hanelt, A., Busse, S., & Kolbe, L.M. (2017). Driving business transformation toward sustainability: Exploring the impact of supporting IS on the performance contribution of eco-innovations. Information Systems Journal, 27, 463–502.

Hildebrandt, B., Hanelt, A., & Firk, S. (2018). Sharing Yet Caring: Mitigating Moral Hazard in Access-Based Consumption through IS-Enabled Value Co-Capturing with Consumers. Business & Information Systems Engineering, 60, 227–241.

Jose, C., Jabbour, C., De Sousa Jabbour, A.B.L., Sarkis, J., & Filho, G. (2019). Unlocking the circular economy through new business models based on large-scale data: An integrative framework and research agenda. Technological Forecasting and Social Change, 144, 546–552.

Kane, G. C., Palmer, D., Phillips, A. N., Kiron, D., & Buckley, N. (2017). Achieving Digital Maturity. In MIT Sloan Management Review (Issue Summer).

Kane, G. C., Palmer, D., Phillips, A. N., Kiron, D., & Buckley, N. (2018). Coming of Age Digitally: Learning, Leadership, and Legacy. https://s3.amazonaws.com/marketing.mitsmr.com/offers/DL2018/59480-MITSMR-Deloitte-Digital-Report-2018.pdf

Marolt, M., Zimmermann, H., Znidarsic, A., & Pucihar, A. (2020). Exploring Social Customer Relationship Management Adoption in Micro, Small and Medium-Sized Enterprises. Journal of Theoretical and Applied Electronic Commerce Research, 15(2), 38–58.

Pucihar, A., Lenart, G., Kljajić Borštnar, M., Vidmar, D., & Marolt, M. (2019). Drivers and Outcomes of Business Model Innovation—Micro, Small and Medium-Sized Enterprises Perspective. Sustainability, 11, 344.

Standardna klasifikacija dejavnosti – SKD 2008. (2008). In https://www.uradni-list.si/files/RS_-2008-017-00559-OB~P001-0000.PDF

Tijan, E., Jović, M., Aksentijević, S., & Pucihar, A. (2021). Digital transformation in the maritime transport sector. *Technological Forecasting and Social Change*, *170*. https://doi.org/10.1016/j.techfore.2021.120879

Vial, G. (2019). Understanding digital transformation: A review and a research agenda. The Journal of Strategic Information Systems, 28(2), 118–144. https://doi.org/10.1016/J.JSIS.2019.01.003

Vidmar, D. (2021). Vpliv informacijskih tehnologij na trajnostno uspešnost organizacij: thesis. [D. Vidmar]. https://dk.um.si/IzpisGradiva.php?id=79148

Vidmar, D., Marolt, M., & Pucihar, A. (2021). Information Technology for Business Sustainability: A Literature Review with Automated Content Analysis. Sustainability, 13(3), 1192. 10.3390/su13031192

IS RISK-AVERSION LINKED TO USE OF DEAL SITES?

František Sudzina, Antonín Pavlíček

Prague University of Economics and Business
Faculty of Informatics and Statistics
frantisek.sudzina@vse.cz, antonin.pavlicek@vse.cz

DOI: 10.35011/IDIMT-2024-283

Keywords

Deal sites; web portals; risk-aversion

Abstract

This study explores the influence of personality traits, specifically risk-aversion, on the usage of deal sites. Deal sites, characterized by their time-limited, discounted offers, play a significant role in the e-commerce sector by influencing consumer behavior and purchasing decisions. The research focuses on understanding how risk-averse individuals interact with these platforms, hypothesizing that their cautious spending habits significantly impact their engagement with such sites. Through quantitative analysis of user data and personality assessments, the study aims to identify a link between risk-aversion and deal site usage patterns. Existing literature suggests that risk-aversion affects the frequency and type of deals pursued by consumers; this study did not identify any significant relationship though.

1. Introduction

Deal sites have become popular, offering users access to discounts, promotions, and special deals on various products and services for over 15 years now. Meanwhile, deal sites have become a significant element in the e-commerce landscape, serving as platforms where businesses offer products or services at significantly reduced prices for a limited time. These sites, including well-known ones like Groupon and LivingSocial, cater to a wide array of consumer needs ranging from travel packages to restaurant deals. Their business model is primarily based on the concept of 'flash sales,' a strategy that creates a sense of urgency among consumers to capitalize on an offer before it expires. This model not only helps businesses clear out inventory and attract new customers but also stimulates consumer spending by making shopping more engaging and time-sensitive.

Understanding user behavior on deal sites is crucial for several reasons. Firstly, it enables businesses to tailor their offers more effectively. By analyzing how different users respond to various types of deals, companies can optimize their promotions to match consumer preferences (Krasnova et al., 2013) (Drossos et al., 2015), thereby increasing the conversion rates and profitability of their campaigns. Additionally, understanding these patterns helps in segmenting the customer base more accurately, allowing for more personalized marketing efforts that can lead to higher customer satisfaction and loyalty.

Moreover, the psychological drivers behind why consumers flock to deal sites are complex and varied. Some users are motivated by the thrill of finding a bargain, while others may be driven by budget constraints, and yet others might be influenced by the social proof of seeing others make purchases. Analyzing these underlying factors can provide deeper insights into consumer behavior and help predict future purchasing patterns. This understanding can be particularly beneficial in designing user interfaces and marketing messages that resonate well with target audiences, enhancing user engagement and retention.

Finally, deal sites accumulate vast amounts of data on consumer behaviour, preferences, and demographics (Hofacker et al., 2016), which can be leveraged to forecast market trends and influence stock management and pricing strategies. In the competitive e-commerce environment, where understanding and anticipating consumer needs is paramount, the insights gained from deal site usage can provide a significant strategic advantage. Therefore, thorough research into the factors influencing deal site usage not only benefits the platforms themselves but also provides broader benefits to the e-commerce ecosystem, influencing everything from marketing strategies to product development.

Literature research

Understanding the factors that influence the usage of deal sites is essential for marketers and businesses looking to reach their target audience effectively. To better understand the usage of deal sites, it is important to analyze the influence of demographic factors on users' engagement with these platforms. Demographic factors such as gender, age, social status, marital status, and economic status can all potentially influence the usage of deal sites.

Research has shown that the demographic characteristics of users play a significant role in their motivation to use deal sites. For example, gender can be a significant factor. (Sudzina, 2016) Women may be more inclined to use deal sites due to their tendency to be more price-conscious and value-driven in their shopping behavior. Age is another demographic factor that can influence deal site usage. Younger individuals, particularly millennials and Gen Z, are more likely to use deal sites as they are generally more tech-savvy and open to trying new online platforms. Social status and economic status are also important factors to consider. (Zhang et al., 2019) Individuals with lower social and economic status may be more motivated to use deal sites as they are looking for ways to save money and find affordable deals. Marital status can also play a role in deal site usage. Married individuals may be more likely to use deal sites as they are often responsible for managing household expenses and are more motivated to find ways to save money for their families. Furthermore, research has also suggested that personality traits may play a role in the usage of deal sites (Krasnova et al., 2013) (Mangiavacchi et al., 2020).

For instance, risk-aversion and deal site usage have been linked (Wu & Chang, 2007) (Wu & Ke, 2015) (Bhatnagar et al., 2000) (Seo & Moon, 2016). Individuals who are more risk-averse may be more inclined to use deal sites as a way to mitigate the potential risks associated with purchasing products or services at full price. In addition to demographic factors, other factors such as attitudes towards deal sites and self-confidence in using social media effectively can also influence usage.

The role of risk aversion in influencing online behaviour, with a particular focus on consumer characteristics and their impact on e-transactional activities has been studied profoundly in academic literature. Both scholars and professionals has highlighted the significance of perceived risk in shaping online consumer behaviour. (Youn & Lee, 2009) (Nawi et al., 2019) (Youn & Lee, 2009) (Zhang et al., 2020) Other factors such as financial, performance, social, time, psychological, and privacy risks have been found to negatively impact consumers' intention to engage in e-commerce activities, but risk aversion seems to be amongst the most important. The study by Featherman and

Pavlou (2003) revealed that perceived risk is a crucial determinant in explaining barriers to online shopping, as consumers weigh the potential losses against the benefits of purchasing goods or services through the internet (Herrero et al., 2009). Similarly, Nepomuceno et al. found that various consumer characteristics, including psychographic, behavioural, and demographic factors, serve as antecedents to online security concerns, which in turn affect e-transactional behaviour (Youn & Lee, 2009).

Interestingly, the role of trust has been identified as a moderating variable in the relationship between risk perception and online purchasing attitudes. The presence of a reputable agent may mitigate the risks associated with online transactions, highlighting the importance of building trust in the vendor for consumers to accept any potential risks.

Furthermore, online reviews can serve as a valuable data source to predict purchasing behaviour, as they provide insights into consumer perceptions and risk assessments. This underscores the need for e-marketers to carefully curate and manage their online reputation, as it can significantly influence consumer decision-making.

Ultimately, the existing literature suggests that the interplay between risk aversion, consumer characteristics, and trust considerations plays a critical role in shaping online human behaviour. (McCole et al., 2010) (Herrero et al., 2009) (Youn & Lee, 2009) (Zhang et al., 2020) Consumers' willingness to participate in e-commerce is closely linked to their assessment of the potential risks involved, which is shaped by a range of individual factors, including psychographic, behavioural, and demographic considerations (Youn & Lee, 2009).

Risk-aversion and deal sites

The influence of risk-aversion on the usage of deal sites presents an interesting area of study within the realm of consumer behavior in e-commerce. Risk-averse individuals tend to be cautious with their spending decisions, preferring options that minimize potential loss (Wu & Chang, 2007) (Santana & Parigi, 2015). This psychological trait influences how such individuals perceive the benefits and potential pitfalls of engaging with deal sites, where offers are often time-sensitive and come with varying degrees of product information and vendor reliability. The main research question, therefore, revolves around how this inclination towards avoiding risk affects the likelihood and frequency with which risk-averse consumers use deal sites compared to their less cautious counterparts.

Understanding the impact of risk-aversion on deal site usage can shed light on broader consumer behavior trends and help tailor marketing strategies accordingly. For instance, if risk-averse users are found to be hesitant to use deal sites due to fears of product quality or the legitimacy of unusually large discounts, deal sites could improve transparency and provide more comprehensive product descriptions to mitigate these concerns. Conversely, if these individuals are drawn to deal sites as a means to stretch their budget while minimizing financial exposure, then highlighting the security measures and satisfaction guarantees could be effective ways to increase their engagement. This research could help delineate the specific features and types of deals that best attract and retain risk-averse shoppers, enhancing the overall effectiveness of deal sites.

Furthermore, examining how risk-aversion influences deal site usage also contributes to a better understanding of the segmentation within the deal site market. By identifying and analyzing the nuances in behavior among different segments of risk-averse consumers, businesses can create more customized experiences that cater to the unique needs and preferences of these groups. Whether it involves curating the types of deals offered, adjusting the marketing approach, or redesigning the user experience, such insights are invaluable for optimizing consumer satisfaction and loyalty in a highly competitive market space. This approach not only benefits the consumers by providing them with deals that are more aligned with their risk tolerance but also enhances the business model of deal sites by fostering a more engaged and diverse user base.

2. Methodology

Data were collected using an online questionnaire on the web platform 1ka.si. There was no monetary incentive for respondents. The sample size was 529 (289 men and 240 women).

The aim of the analysis is to test a link between risk-aversion and type of engagement with deal sites. The dependent variable was risk-aversion. It was measured on a 1-10 Likert scale where 1 meant risk-loving, and 10 meant risk-aversion. The independent variable was use of deal sites. It was measured on a scale from 1 to 4 where 1 meant yes, often; 2 meant yes, sometimes, 3 meant no, but I think about it; and 4 meant no.

One-way ANOVA (analysis of variance) will be used for the analysis of whether there is any difference in risk-aversion between respondents with varying use of deal sites in order to be able to uncover a relationship even if it is not monotonous. IBM SPSS 27 will be used for the calculations.

3. Results

Results from one-way ANOVA testing whether there is any difference in risk-aversion between respondents with varying use of deal sites is provided in Table 1.

Table 1. Analysis of variance

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3.617	3	1.206	0.269	0.848
Within Groups	2356.421	525	4.488		
Total	2360.038	528			

Source: (authors)

There is no significant difference in risk-aversion between respondents with varying use of deal sites. Authors are confident that the risk-taking scale had a potential to uncover a relationship if there was one because in (Sudzina, Pavlíček, 2024) this scale was used, and after a re-analysis of these shared-mobility data, F was equal to 10.466, and the significance was under 0.001.

To get a better overview of the data at hand, averages and standard deviations are provided in Table 2.

Table 2. Summary statistics

	Mean	Std. Deviation
Yes, often	5.70	2.191
Yes, sometimes	5.54	2.004
No, but I think about it	5.72	2.224
No	5.70	2.173
Total	5.64	2.114

Source: (authors)

As it can be seen in Table 2, average risk-aversion is virtually the same for three of four groups of respondents. Infrequent users of deal sites have a marginally higher risk-aversion, but it is not significant given standard deviations.

4. Conclusion

Deal sites function as another channel. They allow companies to sell to customers for whom the official price is high, thus increasing revenue. For some potential customers, the price is not high per se; it is rather high considering when they buy a certain product or a service for the first time, and they perceive price in relation to the unknown quality of the product or of the service. The purchase is perceived less risky when the price is lower. On the other hand, some potential customers may avoid using deal sites because they may fear a risk that a product or a service at a lower price may not be of adequate quality, or that they will not manage to use a coupon on time, especially when it comes to a service. The aim of the paper was to test whether one of these two risk perspectives is stronger, while being open also to a U-shaped or an inverse U-shaped relationship. The analysis of variance did not yield a significant result, and also simply looking at the actual average values that differ at most by 0.18 on a 1-10 Likert scale, it can be summed up that these two risk perspectives cancel each other out.

Acknowledgement

Paper was processed with support from institutional-support fund for long-term conceptual development of science and research at the Faculty of Informatics and Statistics of the Prague University of Economics and Business (IP400040) and the project IG 409033.

References

Bhatnagar, A., Misra, S., & Rao, H R. (2000). On risk, convenience, and Internet shopping behavior. Communications of the ACM. https://dl.acm.org/doi/10.1145/353360.353371

Drossos, D., Maragoudakis, M., & Kokkinaki, F. (2015). Buying Behavior on Daily-Deal Sites: The Role of Face Value, Product Involvement, Information, and Website Quality. Journal of Internet commerce. https://www.tandfonline.com/doi/full/10.1080/15332861.2015.1011568

Herrero, Á., Bosque, I R D., & Sánchez, M M G D L S. (2009, March 1). The influence of perceived risk on Internet shopping behavior: a multidimensional perspective. Routledge, 12(2), 259-277. https://doi.org/10.1080/13669870802497744

Hofacker, C F., Malthouse, E C., & Sultan, F. (2016). Big Data and consumer behavior: imminent opportunities. Journal of consumer marketing. https://www.emerald.com/insight/content/doi/10.1108/JCM-04-2015-1399/full/html

Krasnova, H., Veltri, N F., Spengler, K., & Günther, O. (2013). “Deal of the Day” Platforms: What Drives Consumer Loyalty? Business & information systems engineering. https://link.springer.com/article/10.1007/s12599-013-0268-2

Krasnova, H., Veltri, N F., Spengler, K., & Günther, O. (2013). “Deal of the Day” Platforms: What Drives Consumer Loyalty?. https://link.springer.com/article/10.1007/s12599-013-0268-2

Mangiavacchi, L., Piccoli, L., & Rapallini, C. (2020). Personality Traits and Household Consumption Choices. The B.E. journal of economic analysis & policy. https://www.degruyter.com/document/doi/10.1515/bejeap-2020-0189/html?lang=en

McCole, P., Ramsey, E., & Williams, J. (2010, September 1). Trust considerations on attitudes towards online purchasing: The moderating effect of privacy and security concerns. Elsevier BV, 63(9-10), 1018-1024. https://doi.org/10.1016/j.jbusres.2009.02.025

Nawi, N C., Mamun, A A., Hamsani, N H., & Muhayiddin, M N. (2019, January 24). Effect of Consumer Demographics and Risk Factors on Online Purchase Behaviour in Malaysia. Multidisciplinary Digital Publishing Institute, 9(1), 10-10. https://doi.org/10.3390/soc9010010

Santana, J., & Parigi, P. (2015). Risk Aversion and Engagement in the Sharing Economy. https://doi.org/10.3390/g6040560

Seo, S., & Moon, S. (2016). Decision-making styles of restaurant deal consumers who use social commerce. https://www.emerald.com/insight/content/doi/10.1108/IJCHM-06-2015-0319/full/html

Sudzina, F. (2016). Do Gender and Personality Traits Influence Frequency of Use of Deal Sites. https://aisel.aisnet.org/bled2016/29/

Sudzina, F., & Pavlíček, A. (2024). Adoption of sustainability-oriented innovation: Case of shared mobility in the Czech Republic. In: 8th International Scientific Conference: Challenges in the turbulent economic environment and organizations' sustainable development. University of Maribor Press, accepted.

Wu, W., & Chang, M S. (2007). THE ROLE OF RISK ATTITUDE ON ONLINE SHOPPING: EXPERIENCE, CUSTOMER SATISFACTION, AND REPURCHASE INTENTION. Social behavior and personality. https://www.ingentaconnect.com/content/sbp/sbp/2007/00000035/00000004/art00003;jsessionid=4ni0dk9sh0bb6.x-ic-live-02

Wu, W., & Ke, C. (2015). An Online Shopping Behavior Model Integrating Personality Traits, Perceived Risk, and Technology Acceptance. Social behavior and personality. https://www.ingentaconnect.com/content/sbp/sbp/2015/00000043/00000001/art00008;jsessionid=2hnl2jubwwhob.x-ic-live-02

Youn, S., & Lee, M. (2009, January 1). The determinants of online security concerns and their influence on e-transactions. Inderscience Publishers, 5(3), 194-194. https://doi.org/10.1504/ijima.2009.026370

Zhang, J., Zheng, W., & Wang, S. (2020, February 3). The study of the effect of online review on purchase behavior. Emerald Publishing Limited, 4(1), 73-86. https://doi.org/10.1108/ijcs-10-2019-0027

Zhang, W., Tian, X., & He, W. (2019). Information Seeking and Online Deal Seeking Behavior. Journal of global information management. https://www.igi-global.com/gateway/article/235372

SOCIAL MEDIA AND THE ROLE OF AI

AI IN SOCIAL MEDIA: HARNESSING INNOVATION AMID ETHICAL AND PRIVACY CHALLENGES

Antonín Pavlíček

Prague University of Economics and Business
Faculty of Informatics and Statistics
antonin.pavlicek@vse.cz

DOI: 10.35011/IDIMT-2024-291

Keywords

Aartificial intelligence; social media; innovations; ethics, privacy; challenges

Abstract

This paper explores the transformative potential of artificial intelligence (AI) in social media, highlighting both its promising benefits and significant challenges. AI applications such as content personalization, sentiment analysis, and automated moderation are examined to illustrate how they enhance user experience and operational efficiency on social media platforms. These innovations enable more engaging, relevant, and safe online interactions while providing valuable insights for businesses and marketers. However, the paper also addresses the darker aspects of AI integration, including privacy concerns, ethical dilemmas, and the risk of misinformation. The creation of "filter bubbles" and the spread of deepfakes are identified as key issues that can undermine public trust and social cohesion. By fostering transparency, informed consent, and stringent data protection, stakeholders can maximize AI's benefits while mitigating its risks. This balanced discourse aims to guide the responsible evolution of AI in social media, emphasizing the need for continuous innovation and ethical vigilance.

1. Introduction

In recent years, there has been a significant evolution in social media platforms alongside an increasing reliance on artificial intelligence to facilitate digital interactions. This trend has redefined how individuals and businesses engage with each other online, shaping new opportunities and challenges in the digital landscape. One of the key areas where AI has made a profound impact is in the realm of social media, where its applications span content personalization, sentiment analysis, and automated moderation, among others. (Kalinová, 2022)

The integration of AI in social media holds immense potential, enabling platforms to deliver a more personalized and engaging user experience. AI-powered algorithms can analyze user preferences, and social interactions to curate content tailored to individual interests, enhancing overall user satisfaction. The personalization of social media content through AI-driven algorithms has been lauded for its ability to deliver more relevant and engaging material to users.

Other potential of AI in social media lies in its ability to enhance the user experience by tailoring content, streamlining interactions, and improving overall platform functionality. However, the integration of AI also raises critical concerns regarding ethical considerations, privacy, and security. (AI Roadmap - 4. Major Findings, 2019) By leveraging machine learning and natural language processing, AI-powered platforms can analyze user preferences, behavior patterns, and social interconnections to curate personalized content and feeds. (Hicham et al., 2023) (Gao et al., 2023)

2. Literature research

AI has shown promise in increasing user engagement and time spent on social media platforms. However, the opaque nature of these algorithms has raised concerns about the potential for filter bubbles, echo chambers, and the amplification of biases and misinformation. (Ananny & Crawford, 2016) The use of AI in sentiment analysis, for example, can assist in the identification of harmful content, such as hate speech or cyberbullying, enabling more effective content moderation. Nevertheless, the accuracy and nuance of these systems remain limited, potentially leading to the over-censorship of legitimate expression or the failure to detect more subtle forms of problematic content, or on the other hand, not understanding the sarcasm. (Jia et al., 2024)

As already mentioned, the integration of AI into social media platforms has significant implications for privacy and security. (Ananny & Crawford, 2016) The extensive data collection and profiling enabled by AI-driven personalization raise concerns about user privacy and the potential for misuse of personal information. (Trattner et al., 2021) Additionally, the use of AI in automated content generation, such as the creation of deepfakes or the manipulation of images and videos, poses a threat to the authenticity and trustworthiness of online information, potentially undermining the integrity of social discourse. (Jungherr, 2023) (Gupta, 2020)

Addressing these challenges will require a multifaceted approach that balances the potential benefits of AI in social media with the necessary safeguards to protect user privacy, mitigate the spread of misinformation, and ensure the responsible deployment of these technologies. (Vinuesa et al., 2020) This may involve the development of more transparent and accountable AI systems, the implementation of robust data protection policies, and the establishment of ethical guidelines to govern the use of AI in social media. (Leavy, 2022) (AI Roadmap - 4. Major Findings, 2019) (Hicham et al., 2023)

As AI continues to shape the social media landscape, it is crucial that researchers, policymakers, and industry stakeholders collaborate to address these challenges and harness the technology's potential in a manner that prioritizes user well-being, democratic principles, and societal stability. (Schiff et al., 2020) By striking a careful balance between innovation and responsible deployment, the transformative power of AI in social media can be harnessed to create a more informed, engaged, and equitable digital ecosystem. (AI Roadmap - 4. Major Findings, 2019) (Vinuesa et al., 2020) (Stop talking about tomorrow’s AI doomsday when AI poses risks today, 2023)

3. AI Innovations Affecting Social Media

The integration of artificial intelligence (AI) in social media has revolutionized the way platforms operate and interact with users. AI innovations have significantly enhanced the user experience by personalizing content, improving engagement metrics, and optimizing advertising strategies. These advancements have also introduced new tools and techniques that streamline content moderation, facilitate customer service, and provide deeper insights into user behaviour. However, while AI has brought numerous benefits to social media, it has also introduced a range of challenges, including ethical dilemmas, privacy concerns, and security risks. This section lists the most impactful AI innovations in social media:

- **Content Personalization Algorithms** - AI-driven recommendation engines that curate personalized content feeds based on user preferences, behaviors, and interactions.
- **Sentiment Analysis Tools** - AI systems that analyze user comments, posts, and reviews to determine the overall sentiment (positive, negative, or neutral) and gauge public opinion.
- **Automated Moderation Systems** - AI-powered moderation tools that automatically detect and remove harmful or inappropriate content, including hate speech, nudity, and violent content.
- **Chatbots and Virtual Assistants** - AI-based chatbots that provide customer service, support, and engage with users in real-time on social media platforms.
- **Predictive Analytics** - AI systems that analyze past data to predict future trends, user behaviors, and content virality, aiding in strategic decision-making to social media managers.
- **Deepfake Technology** - AI tools capable of creating hyper-realistic fake videos and images, which have significant implications for misinformation and privacy, which can be highly toxic in social media environment.
- **Ad Targeting and Optimization** - AI algorithms that analyse user data to deliver targeted advertisements, optimizing ad spend and improving conversion rates for businesses.
- **Influencer Identification and Analysis** - AI tools that identify and analyze social media influencers, helping brands to partner with the most impactful personalities.
- **Image and Video Recognition** - AI technologies that recognize and tag people, objects, and scenes in images and videos, enhancing content searchability and engagement.
- **Fraud Detection and Prevention** - AI systems that detect and prevent fraudulent activities, such as fake accounts, spamming, and phishing on social media platforms.
- **Social Listening and Monitoring** - AI-powered tools that monitor social media platforms for brand mentions, trends, and user feedback, providing valuable insights for businesses.
- **Content Creation and Curation** - AI-driven tools that assist in the creation and curation of content, including automated video editing, article generation, and image enhancement.
- **Voice Recognition and Interaction** - AI technologies that enable voice commands and interactions, making social media more accessible and user-friendly.
- **Augmented Reality (AR) Filters and Effects** - AI-driven AR tools that create interactive and engaging experiences through filters and effects on platforms like Instagram and Snapchat.
- **Behavioral Analysis** - AI systems that analyse users’ behavioural patterns to predict future actions, tailor user experiences, and improve platform design.
- **Trend Analysis and Forecasting** - AI tools that identify emerging trends and forecast their potential impact on social media dynamics and user engagement.

- **Content Translation and Localization** - AI-powered translation tools that enable real-time translation and localization of content, making social media more inclusive and global.
- **Emotion Detection** - AI technologies that analyze facial expressions and tone of voice in videos and audio clips to detect user emotions, enhancing content recommendations and user experience.

4. Challenges and Limitations of AI in Social Media

While the integration of AI in social media offers significant opportunities, it also presents a range of pressing challenges and limitations that must be addressed. One of the primary concerns is the issue of algorithmic bias, where AI-driven content curation and moderation can amplify existing societal biases and inequalities (Ferreira, 2023). Depending on the training data and underlying algorithms, AI systems may perpetuate discrimination against certain groups or marginalize their voices and perspectives. (Vinuesa et al., 2020) This can lead to the creation of echo chambers and filter bubbles, where users are exposed to a limited range of viewpoints, further entrenching polarization and undermining the diversity of online discourse.

Additionally, the opaque and complex nature of AI algorithms used in social media platforms can make it challenging to understand the decision-making processes and ensure transparency and accountability. (Vinuesa et al., 2020) (Schwartz et al., 2022) (Hicham et al., 2023) (Vinuesa et al., 2020) (Mello et al., 2023) Without adequate oversight and auditing mechanisms, the risk of unfairness, discrimination, and the manipulation of user experiences remains high.

The extensive data collection and profiling enabled by AI-driven personalization also raise significant privacy concerns. The aggregation and analysis of user data, including personal preferences, browsing history, and social connections, can expose individuals to potential misuse or unauthorized access, undermining user trust and autonomy.

4.1 Deepfakes and Misinformation

Positive Consequences: Deepfake technology, which uses AI to create realistic but artificial images, videos, and audio, has potential positive applications in entertainment, education, and creative industries. For instance, filmmakers can use deepfakes to create stunning visual effects or resurrect historical figures, while educators can develop engaging learning materials. This technology also enables innovative advertising and marketing campaigns, capturing audience attention with highly realistic and personalized content.

Negative Consequences: Despite its potential, deepfake technology poses significant risks, particularly in the context of social media. The ability to create convincing fake content can lead to the spread of misinformation, fraud, and malicious manipulation, undermining public trust in media and institutions. Deepfakes can be used to create false narratives, damage reputations, and even incite violence by portraying individuals saying or doing things they never did. The challenges in detecting and combating deepfakes require advanced technical solutions and regulatory measures to prevent their misuse.

Moreover, the use of AI in automated content generation poses a threat to the authenticity and trustworthiness of online information. The creation of highly convincing deepfakes, where AI-generated images, videos, or audio can be used to misrepresent individuals or events, can undermine the integrity of social discourse and erode public trust. (Stop talking about tomorrow's AI doomsday when AI poses risks today, 2023) This challenge is exacerbated by the rapid advancements in generative AI models, which can produce increasingly realistic and hard-to-detect synthetic media. (Living in a brave new AI era, 2023)

The proliferation of AI-generated misinformation, whether in the form of deepfake, or “just” fake news articles, manipulated visuals, or automated social media posts, can have far-reaching consequences for democratic processes, public safety, and social cohesion. (Epstein et al., 2023)

The ability of malicious actors to exploit these technologies to sow discord, influence public opinion, and undermine the credibility of legitimate sources of information poses a significant threat to the well-being of individuals and communities. (Stop talking about tomorrow’s AI doomsday when AI poses risks today, 2023)

4.2 Content Personalization Algorithms

Positive Consequences: Content personalization algorithms have improved – even revolutionized – the user experience on social media platforms by tailoring content to individual preferences and behaviors. This innovation enhances user engagement and satisfaction by presenting users with content that is relevant to their interests, thereby increasing the amount of time they spend on the platform. For businesses, personalized content helps target advertising more effectively, leading to higher conversion rates and improved ROI. Additionally, users benefit from discovering content and communities that align with their tastes and preferences, fostering a more engaging and enjoyable online environment. (Aiolfi et al., 2021)

Negative Consequences: Despite its benefits, content personalization raises significant concerns about the creation of "filter bubbles" and "echo chambers," where users are exposed primarily to information that reinforces their existing beliefs and opinions. This can lead to a polarized society, where individuals become less exposed to diverse perspectives and more entrenched in their views. Moreover, the use of personal data for content personalization raises privacy concerns, as it often involves extensive data collection and analysis, sometimes without explicit user consent. This can lead to potential misuse of personal information and a loss of user trust in social media platforms. (O’Reilly et al., 2024)

Furthermore, the algorithmic amplification of political content on social media platforms can contribute to the spread of misinformation and the manipulation of public opinion – and even radicalization of the society. Personalization algorithms may inadvertently prioritize sensational, emotionally-charged, or politically-charged content, which can then be further amplified through user engagement and sharing.

4.3 Sentiment Analysis

Positive Consequences: Sentiment analysis leverages AI to assess public opinion by analyzing user-generated content on social media. This technology enables businesses, governments, and organizations to gain valuable insights into public sentiment, allowing them to make informed decisions and respond promptly to emerging trends and issues. For instance, companies can use sentiment analysis to gauge customer satisfaction, improve products and services, and enhance customer engagement. In the political arena, sentiment analysis helps identify voter concerns and preferences, enabling more responsive and targeted campaign strategies.

Negative Consequences: The use of sentiment analysis also presents several challenges, including the potential for misuse and ethical concerns. AI algorithms used in sentiment analysis can sometimes misinterpret sarcasm, slang, and cultural nuances, leading to inaccurate assessments of public sentiment. Additionally, there is a risk that sentiment analysis could be exploited for manipulative purposes, such as spreading propaganda or misinformation to sway public opinion. Privacy concerns also arise as sentiment analysis often involves the collection and processing of vast amounts of personal data, raising questions about consent and data protection. (Razali et al., 2021)

Another significant challenge is the potential for AI-powered sentiment analysis to amplify existing biases and widen already existing social divisions.

4.4 Automated Content Moderation

Positive Consequences: Automated content moderation uses AI to detect and remove harmful or inappropriate content from social media platforms, enhancing the safety and well-being of users. This innovation significantly reduces the prevalence of hate speech, cyberbullying, and explicit content, creating a safer online environment. Automated moderation also helps platforms comply with legal requirements and community standards more efficiently, reducing the burden on human moderators and allowing for quicker responses to violations.

Negative Consequences: However, automated content moderation is not without its drawbacks. AI algorithms can sometimes fail to accurately distinguish between harmful content and legitimate expression, leading to the wrongful removal of content or even censorship. This can stifle free speech and limit the diversity of viewpoints available on social media. Additionally, bad actors may develop ways to circumvent automated systems, rendering them less effective over time. The reliance on automated moderation also raises concerns about the transparency and accountability of content moderation decisions. (Singhal et al., 2023)

Moreover, the use of AI in this context raises ethical questions about the impact on marginalized communities, as algorithmic biases can lead to disproportionate targeting or censorship of certain groups.

4.5 Chatbots and Virtual Assistants

Positive Consequences: Chatbots and virtual assistants powered by AI have changed customer service on social media by providing instant, 24/7 support to users. These tools enhance user experience by offering quick responses to queries, resolving issues efficiently, and providing personalized recommendations. For businesses, chatbots reduce operational costs associated with customer support and can improve customer satisfaction and retention by ensuring timely and consistent service.

Negative Consequences: Deployment of chatbots and virtual assistants can lead to challenges, such as the loss of human touch in customer service interactions. Users may feel frustrated when dealing with bots that lack the empathy and understanding of human agents, particularly for complex or sensitive issues. Additionally, poorly designed chatbots can result in miscommunications and inadequate problem resolution, potentially damaging a company's reputation. Security and privacy concerns also arise as chatbots handle personal and sensitive information, requiring robust safeguards to prevent data breaches and misuse. (Caldarini et al., 2022)

4.6 Influencer Marketing AI

Positive Consequences: AI-driven influencer marketing tools have enhanced the efficiency and effectiveness of campaigns by identifying the most suitable influencers for brands, analyzing engagement metrics, and predicting campaign outcomes. This allows companies to maximize their return on investment by partnering with influencers who align with their target audience and brand values. AI also streamlines the process of tracking campaign performance and making data-driven adjustments, leading to more successful and impactful marketing efforts.

Negative Consequences: However, AI-driven influencer marketing also introduces potential challenges, such as the commodification of personal relationships and authenticity. The reliance on algorithms to identify influencers and measure success can sometimes overlook the nuanced, human aspects of influencer-follower relationships. Additionally, the pressure on influencers to meet algorithmic criteria can lead to inauthentic content creation and a loss of genuine engagement with their audience. There are also concerns about data privacy, as influencer marketing often involves extensive data collection and analysis. (Bishop, 2021)

4.7 Predictive Analytics

Positive Consequences: Predictive analytics powered by AI enables social media platforms and businesses to anticipate user behavior, trends, and preferences, allowing for proactive decision-making. This technology helps companies optimize their marketing strategies, improve customer experiences, and stay ahead of competitors by leveraging data-driven insights. Predictive analytics also supports content creators in tailoring their posts to maximize engagement and reach, fostering a more vibrant and dynamic online community.

Negative Consequences: Despite its advantages, predictive analytics raises issues related to user autonomy and privacy. The use of AI to predict and influence user behavior can lead to concerns about manipulation and the erosion of free will, as individuals may be subtly steered towards certain actions or choices. Privacy concerns are also paramount, as predictive analytics relies on the collection and analysis of large amounts of personal data, often without explicit user consent. This can result in a loss of trust and potential legal and ethical challenges for social media platforms and businesses. (Mühlhoff, 2021)

4.8 Recommendation Systems

Positive Consequences: Recommendation systems use AI to suggest content, products, and connections to users based on their preferences and behavior. This enhances user engagement by helping individuals discover new and relevant content, fostering a more personalized and enjoyable experience on social media platforms. Businesses benefit from increased visibility and sales as recommendation systems effectively match their offerings with interested users, driving higher conversion rates and customer satisfaction.

Negative Consequences: However, recommendation systems can also contribute to issues such as information overload and the amplification of biased or harmful content. By continuously presenting users with similar types of content, these systems can limit exposure to diverse perspectives and reinforce existing biases. Additionally, there is a risk of users becoming overwhelmed by the sheer volume of recommendations, leading to decision fatigue and decreased satisfaction. The algorithms behind recommendation systems can also be exploited to spread misinformation or manipulate user behavior, posing significant ethical and security concerns. (Smith et al., 2022)

5. Mitigation Strategies and Ethical Considerations

A multifaceted approach is required to address the negative consequences of AI in social media, involving collaboration between technology companies, policymakers, and civil society organizations. One crucial step is to embed democratic values and societal objectives into the design and deployment of AI systems, ensuring they serve the greater good rather than merely optimizing for individual user engagement or profits. (Huang, 2024)

This can be achieved by developing AI systems with explicit ethical principles and oversight mechanisms to prevent the amplification of biases, misinformation, and social divisions. (Elliott et al., 2021)

Recommendation algorithms, for instance, should be designed to promote diverse perspectives and viewpoints, rather than reinforcing echo chambers. Similarly, sentiment analysis tools should be trained to recognize and mitigate the propagation of harmful content, rather than exacerbating existing prejudices. (Jia et al., 2024)

Robust data governance frameworks and privacy safeguards are also crucial to address the ethical and security concerns arising from the extensive data collection and analysis undertaken by AI systems in social media.

6. Conclusion

The rapid advancement of AI technology in the social media landscape presents both opportunities and significant challenges. While AI-driven tools have enhanced the efficiency and personalization of marketing, content recommendation, and user engagement, they have also introduced new ethical dilemmas and societal risks.

The commodification of influencer-follower relationships, the potential for manipulation through predictive analytics, and the amplification of biases and misinformation through recommendation systems are just a few of the pressing issues that must be addressed (Elliott et al., 2021)

Artificial intelligence has undeniably transformed social media, bringing forth a multitude of advancements that enhance user experiences, streamline business operations, and foster innovative content creation. From personalized content recommendations that keep users engaged, to sophisticated sentiment analysis that provides real-time insights into public opinion, AI has proven to be a powerful tool in the social media landscape. These innovations have enabled platforms to better understand and serve their users, creating more dynamic and interactive online communities. Businesses, in particular, have reaped significant benefits from AI-driven marketing and customer service solutions, which have optimized engagement and improved overall efficiency.

However, the integration of AI into social media is not without its challenges. The very technologies that offer substantial benefits also pose significant risks and ethical dilemmas. Issues such as privacy invasion, data security, and the creation of echo chambers and filter bubbles are pressing concerns that need to be addressed. The misuse of AI for manipulative purposes, such as spreading misinformation through deepfakes or exploiting predictive analytics to influence user behavior, highlights the darker side of this technological advancement. These challenges necessitate a critical examination of how AI is deployed and managed within social media platforms.

To mitigate these risks, it is essential to develop robust ethical frameworks and regulatory measures that guide the use of AI in social media. Transparency in algorithmic decision-making processes, informed user consent, and stringent data protection policies are crucial components in ensuring that AI is used responsibly and ethically. Collaboration between technology developers, policymakers, and civil society organizations can foster a balanced approach that maximizes the benefits of AI while minimizing its potential harms. Ongoing dialogue and research are needed to adapt to the rapidly evolving landscape and proactively address emerging issues.

Acknowledgement

Paper was processed with support from institutional-support fund for long-term conceptual development of science and research at the Faculty of Informatics and Statistics of the Prague University of Economics and Business (IP400040) and the project IG 409033.

To appropriately credit the AI tools used in the writing process the author acknowledges the use of Google Translate for assistance in translating some formulations into English. Additionally, the author used Grammarly for proofreading and improving the formulation of the manuscript.

References

AI Roadmap - 4. Major Findings. (2019, August 12). https://cra.org/ccc/ai-roadmap-major-findings/

Aiolfi, S., Bellini, S., & Pellegrini, D. (2021, May 14). Data-driven digital advertising: benefits and risks of online behavioral advertising. https://doi.org/10.1108/ijrdm-10-2020-0410

Ananny, M., & Crawford, K. (2016, December 13). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. SAGE Publishing, 20(3), 973-989. https://doi.org/10.1177/1461444816676645

Bishop, S. (2021, January 1). Influencer Management Tools: Algorithmic Cultures, Brand Safety, and Bias. https://doi.org/10.1177/20563051211003066

Caldarini, G., Jaf, S., & McGarry, K. (2022, January 15). A Literature Survey of Recent Advances in Chatbots. https://doi.org/10.3390/info13010041

Elliott, K., Price, R M., Shaw, P., Spiliotopoulos, T., Ng, M., Coopamootoo, K P L., & Moorsel, A V. (2021, June 1) . Towards an Equitable Digital Society: Artificial Intelligence (AI) and Corporate Digital Responsibility (CDR). Springer Science+Business Media, 58(3), 179-188. https://doi.org/10.1007/s12115-021-00594-8

Epstein, Z., Hertzmann, A., Herman, L., Mahari, R., Frank, M R., Groh, M., Schroeder, H., Smith, A., Akten, M., Fjeld, J., Farid, H., Leach, N., Pentland, A., & Russakovsky, O. (2023, January 1). Art and the science of generative AI: A deeper dive. Cornell University. https://doi.org/10.48550/arxiv.2306.04141

Ferreira, C G. (2023, January 1). A short review of the main concerns in A.I. development and application within the public sector supported by NLP and TM. Cornell University. https://doi.org/10.48550/arxiv.2308.02042

Gao, B., Wang, Y., Xie, H., Hu, Y., & Hu, Y. (2023, October 1). Artificial Intelligence in Advertising: Advancements, Challenges, and Ethical Considerations in Targeting, Personalization, Content Creation, and Ad Optimization. SAGE Publishing, 13(4). https://doi.org/10.1177/21582440231210759

Gupta, A. (2020, January 1). Montreal AI Ethics Institute's Response to Scotland's AI Strategy. Cornell University. https://doi.org/10.48550/arxiv.2006.06300

Hagerty, A., & Rubinov, I. (2019, January 1). Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence. Cornell University. https://doi.org/10.48550/arxiv.1907.07892

Hicham, N., Nassera, H., & Karim, S. (2023, September 7). Strategic Framework for Leveraging Artificial Intelligence in Future Marketing Decision-Making. 2(3), 139-150. https://doi.org/10.56578/jimd020304

Huang, S. (2024, March 18). A Roadmap to Democratic AI. https://cip.org/research/ai-roadmap

Jia, C., Lam, M., Chau, M T., Hancock, J T., & Bernstein, M S. (2024, April 17). Embedding Democratic Values into Social Media AIs via Societal Objective Functions. Association for Computing Machinery, 8(CSCW1), 1-36. https://doi.org/10.1145/3641002

Jungherr, A. (2023, July 1). Artificial Intelligence and Democracy: A Conceptual Framework. SAGE Publishing, 9(3). https://doi.org/10.1177/20563051231186353

Kalinová, E. (2022, December 31). USAGE OF ARTIFICIAL INTELLIGENCE ON SOCIAL MEDIA IN EUROPE. https://doi.org/10.33543/1202330333

Leavy, S. (2022, January 1). Inclusive Ethical Design for Recommender Systems. Cornell University. https://doi.org/10.48550/arxiv.2209.13021

Living in a brave new AI era. (2023, November 20). Nature Portfolio, 7(11), 1799-1799. https://doi.org/10.1038/s41562-023-01775-7

Mello, R F., Freitas, E L S X., Pereira, F D., Cabral, L D S., Tedesco, P., & Ramalho, G. (2023, January 1). Education in the age of Generative AI: Context and Recent Developments. Cornell University. https://doi.org/10.48550/arxiv.2309.12332

Mühlhoff, R. (2021, July 31). Predictive privacy: towards an applied ethics of data analytics. https://doi.org/10.1007/s10676-021-09606-x

O'Reilly, T., Strauss, I., & Mazzucato, M. (2024, January 1). Algorithmic attention rents: A theory of digital platform market power. https://doi.org/10.1017/dap.2024.1

Razali, N A M., Malizan, N A., Hasbullah, N A., Wook, M., Zainuddin, N M., Ishak, K K., Ramli, S., & Sukardi, S. (2021, December 1). Opinion mining for national security: techniques, domain applications, challenges and research opportunities. https://doi.org/10.1186/s40537-021-00536-5

Schiff, D., Ayesh, A., Musikanski, L., & Havens, J C. (2020, October 11). IEEE 7010: A New Standard for Assessing the Well-being Implications of Artificial Intelligence. https://doi.org/10.1109/smc42975.2020.9283454

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P B. (2022, March 15). Towards a standard for identifying and managing bias in artificial intelligence. https://doi.org/10.6028/nist.sp.1270

Singhal, M., Ling, C., Paudel, P., Thota, P., Kumarswamy, N., Stringhini, G., & Nilizadeh, S. (2023, July 1). SoK: Content Moderation in Social Media from Guidelines to Enforcement, and Research to Practice. https://doi.org/10.1109/eurosp57164.2023.00056

Smith, J J., Jayne, L., & Burke, R. (2022, September 13). Recommender Systems and Algorithmic Hate. https://doi.org/10.1145/3523227.3551480

Stop talking about tomorrow's AI doomsday when AI poses risks today. (2023, June 27). Nature Portfolio, 618(7967), 885-886. https://doi.org/10.1038/d41586-023-02094-7

Trattner, C., Jannach, D., Motta, E., Meijer, I C., Diakopoulos, N., Elahi, M., Opdahl, A L., Tessem, B., Borch, N T., Fjeld, M., Øvrelid, L., Smedt, K D., & Moe, H. (2021, December 20). Responsible media technology and AI: challenges and research directions. Springer Nature, 2(4), 585-594. https://doi.org/10.1007/s43681-021-00126-4

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S D., Tegmark, M., & Nerini, F F. (2020, January 13). The role of artificial intelligence in achieving the Sustainable Development Goals. Nature Portfolio, 11(1). https://doi.org/10.1038/s41467-019-14108-y

TURBULENT CHANGES ON THE X NETWORK AND THEIR INFLUENCE ON THE ACTIVITY OF POLITICIANS

Jana Syrovátková

Faculty of Informatics and Statistics
Prague University of Economics and Business
jana.syrovatkova@vse.cz

DOI: 10.35011/IDIMT-2024-301

Keywords

Twitter; social network X; Musk; politicians

Abstract

We have been studying the social network X (Twitter) and the activity of European politicians on it since 2020. After Elon Musk took over the Twitter social network in the fall of 2022, many fundamental changes took place on it. Against the backdrop of these changes, we continue to explore how European politicians are using the social network, now renamed X, for their own promotion and marketing. The text researches how individual politicians are active online and how they build their networks of fans and followers. We find that, although it cannot be said that after the huge increase in the time of covid, the use of politicians continues to rise, the number of fans who like politicians' pages is gradually increasing. It is still maintained that politicians from countries with higher GDP are more active than politicians from countries with lower GDP. Since this is a discussed topic, the study will continue in the following years to verify the trends.

1. Introduction

1.1 Political marketing on social networks

Social networks have long served as a place where many people meet. It is so completely natural that even political marketing does not forget about social networks. Among the earliest studies on political marketing on social networks is The New Campaign: Social Networking Sites in the 2008 Presidential Election, which compared the activity of Barack Obama and his opponent John McCain on the social networks Facebook and MySpace. In its conclusion, the study confirmed that social networks are an effective tool for reaching a large number of voters at a relatively low cost (Payne, 2009).

Safiullah et al. (2017) showed that political parties more active on social networks were able to achieve more seats in elections.

1.2 Changes on Twitter – X network

The social network Twitter was founded in 2006. In 2022, Elon Musk announced that he was interested in acquiring Twitter. Although he later withdrew his intention, after a series of legal disputes, Musk finally bought Twitter for the originally agreed amount of 44 billion dollars. (Pavlíček, 2010; Vanian, 2022; Britannica, 2024). Elon Musk became the director of the company, and from the fall of 2022, Twitter began to gradually transform.

The main and most significant change was the change of the name Twitter to X. In this text, if I continue to use the word Twitter, it is surveys and data before 2022. If it is comparisons or more recent studies or data, I use X.

What is also key to this study is the change in the principle of authentication of social network users. Account verification is only in the form of paying a subscription. If the account is verified, it gets a colored badge. A gray mark appears to indicate those profiles that represent a government official or multilateral organization. Among government persons, we can understand here heads of state, presidents or members of governments and parliaments (X Help Center, 2024; Slížek, 2023). This makes it easier for researchers to recognize that this is indeed a verified profile of a publicly known politician.

1.3 Politicians activity on the X network

One of the well-known cases involving politicians and social networks is the use of Twitter by Donald Trump. According to critics, he spread misinformation via Twitter. It all escalated after he failed in the 2020 elections and started spreading suspicions of election rigging. He used Twitter to rouse voters so much that protesters stormed the Capitol. Twitter's response was to block his account. Elon Musk, referring to the fight against censorship, returned the account to him again (Fung, 2021).

Although in Europe the social network X is not as popular as, for example, in the USA, according to Grayling (2021), who compared individual countries in terms of the number of posts, more than 2/3 of their content is shared by MEPs on the social network X. The remaining 1/3 is shared on social networks from Meta, most of it on Facebook. And this despite the fact that more parliamentarians have an account on Facebook than on the X network. The most active were parliamentarians from Spain, Poland and members of the European Parliament. On the other hand, MPs from Slovakia, Serbia and Bulgaria had the fewest contributions. When we compare it with total X users, leading countries based on number of X users as of April 2024 is USA (106 mio users), the leading EU countries are United Kingdom (24,3 mio users), Germany (15,42 mio users), France (15,22 mio users), Spain (11,43 mio users) and Netherlands (8,2 mio users) (Statista, 2024).

According to Larsson (2015), the age of the politician is also one of the important factors in how much they use the social network Twitter. Haman and Školník (2021) made available a database on the activity of members of the European Parliament on the social network Twitter. The most active were the politicians of France, Ireland, the Netherlands, Spain and the UK. On the contrary, the parliamentarians of Bulgaria and Romania were the least active.

If we look at what politicians use social network X for, we find, for example, that in Spain they use it to share political promises and proposals (López-Meri et al., 2017). In Germany, on the other hand, they mainly used it to comment on current political events and left the campaign on Facebook (Stier et al., 2018).

With the purpose of the topic of the article, i.e. among other things, the text of Barberá and Zeitzoff (2018), which showed that the better the level of democracy in a given country, the greater the increase in the use of Twitter by the head of state, is related to examining the differences between states with higher and lower GDP. Conversely, the greater the threat of social unrest in the country, the more likely politicians were to be active on Twitter (Barberá; Zeitzoff, 2018).

1.4 Previous research by the authors

The first study (Pavlíček, Syrovátková, 2022) examined only data from 2020 and compared whether there are significant differences between the activity of politicians according to geographical location (northern, southern, western and eastern states of Europe according to the UN region) and whether there are differences between the activity of politicians according to GDP. The study ultimately confirmed the hypothesis that there are more active politicians in countries with higher GDP per capita. Western Europe was the most active region in contrast to Eastern Europe, which fared the worst of all regions (Pavlíček; Syrovátková, 2022).

The second study focused on comparing activity in 2020 and 2022. Results between regions were also compared. It also turned out that there are differences between years 2020 and 2022 only in Tweets and Likes, not significant difference in number of Followers and Following. (Syrovátková; Korčák, 2023).

2. Methodology

2.1. Data collection

Since 2020, we have been regularly downloading data about leading EU politicians from the social network X (formerly Twitter). We will always take individual EU states, search for their actual official leader (heads of state – the president, the King, the Queen), the actual prime minister, ministers, the Speaker of Parliament and the leaders of the 2 strongest opposition parties (missing in 2022). We will then search if and what kind of profile they have on the social network X and put into the database the names of the actual politicians and their Twitter/X accounts for possibility of control the data. In the first year, the data was subsequently downloaded manually, in subsequent years the Apify scraper was already used, which downloaded information about individual accounts – the number of posts, likes, followers and following. As the United Kingdom was part of the EU in 2020, when we started our analyses, we still download UK-related data as well.

The data is then stored in an excel database – with the download date, the name of the downloaded account, the name and function of the policy and specific numbers.

2.2. Data differentation

For comparing and analyzing the data we differ countries into 2 groups. We have used Volume indicators GDP per capita in 2022 from Eurostat (2024). We take GDP per capita in EU as 100. Than we take first 14 countries (with UK) as countries with high GDP – it has volume 94 and more and second 14 countries as countries with low GDP – with volume 90 and less.

Countries with high GDP per capita are: Austria, Belgium, Cyprus, Denmark, Finland, France, Germany, Ireland, Italy, Luxembourg, Malta, Netherlands, Sweden, United Kingdom.

Coutries with low GDP per capita are: Bulgaria, Croatia, Czech Republic, Estonia, Greece, Hungary, Latvia, Lithuania, Poland, Portugal, Romania, Slovakia, Slovenia, Spain.

3. Results

In 2020, 188 politicians had Twitter accounts, of which 131 were Ministers and The Head of the Country, in 2022 this number increased to 155 and for 2024 it increased to only 158. In total, including other functions, 228 were followed in 2024 politicians' accounts. It is interesting that, of the Prime Ministers, only the Minister of Slovenia has not an account on the X network. The Head of the Country we could not find 3 politicians. Anyway, the X network seems to be increasingly popular among politicians, and more than 80% of high-profile politicians have an X network account. The number of politicians by individual function is in Table 1.

Table 1. Number of politicians with X (Twitter) account

Politician function	Year of the data			Total
	2020	2022	2024	
1. Politician of the Opp.	21	Not taken	27	**48**
2. Politician of the Opp.	21	Not taken	24	**45**
Head of the Parliament	15	Not taken	20	**35**
Minister for Culture	14	19	19	**52**
Minister for Education	17	18	19	**54**
Minister for Finance	19	21	24	**64**
Minister for Foreign Affairs	21	26	25	**72**
Minister of Justice	16	23	19	**58**
Prime Minister	27	23	27	**77**
The Head of the Country	17	25	25	**66**
Total	**188**	**155**	**228**	**571**

Source: (author)

3.1. Changes in the activity between 2022 and 2024

The main question of the research was how the changes from Twitter to X had an effect on the activity of politicians on this network. For further analyses, we exclude data that is missing from 2022 – i.e. we do not take into account opposition politicians and Heads of the Parliament. Table 2 shows that, apparently due to the covid years, the activity of politicians in 2020-2022 was significantly higher than it is now, but their profiles have significantly more likes and even more followers.

Table 2. Total number of Posts, Likes, Followers, Following during years

	2020	2022	2024	Total
Tweets/posts	505 360	1 061 790	999 091	**3 952 734**
Likes	152 987	520 842	943 958	**2 080 780**
Followers	14 626 637	35 621 304	42 393 941	**119 537 742**
Following	96 490	146 791	136 495	**507 053**

Source: (author)

To verify whether the increase is statistically relevant, we tested two hypotheses

H_1: the average number of likes in 2022 and 2024 is the same against the alternative that in 2024 the number of likes was higher.

H_2: the average number of followers in 2022 and 2024 is the same against the alternative that in 2024 the number of followers was higher.

Due to the lower averages for posts and following, we did not test these two hypotheses.

The F-test proved that the data likes do not have equal variances, therefore a t-test with inequality of variances was used, for H2 the hypothesis of equal variances could not be rejected, therefore a t-test with equality of variances was used. Hypothesis H1 is rejected, at the 0.05 significance level. Hypothesis H2 could not be rejected. The results are clearly shown in Table 3.

Table 3. F and t-test p-values

	F-test p-value	**t-test p-value**
H1	0	0.017
H2	0.111	0.707

Source: (author)

3.2. Differences between states with high and low GDP

For the test of differences between states with high and low GDP we will test these hypothesis:

H_1: the average number of posts of politics from the states with high GDP in 2024 is the same like for the politics with low GDP against the alternative that politics from the states with had more posts.

H_2: the average number of likes of politics from the states with high GDP in 2024 is the same like for the politics with low GDP against the alternative that politics from the states with had more likes.

H_3: the average number of followers of politics from the states with high GDP in 2024 is the same like for the politics with low GDP against the alternative that politics from the states with had more followers.

H_4: the average number of following of politics from the states with high GDP in 2024 is the same like for the politics with low GDP against the alternative that politics from the states with had more following.

A summary of the data is in Table 4.

Table 4. means, F and t-test p-values for GDP

	High GDP mean	**Low GDP mean**	**F-test p-value**	**t-test p-value**
H1 – posts	7 630	7 541	**0**	0,969
H2 – likes	7 400	4 187	**$3{,}6 * 10^{-06}$**	**0,035**
H3 – followers	345 176	119 944	**$3{,}8 * 10^{-35}$**	**0,031**
H4 – followings	1 056	747	0,265	**0,045**

Source: (author)

The F-test proved that the data likes do not have equal variances for the H1, H2 and H3, therefore a t-test with inequality of variances was used, for H4 a t-test with equal variances was used. Hypothesis H1 is rejected, at the 0.05 significance level. Hypothesis H2, H3 and H4 could be rejected at the 0.05 significance level. So we can´t say that politics from the states with high GDP are more active in the number of posts than politics from the states with low GDP. What is interesting is that although the number of posts is not significantly higher, numer of likes, followers and followings is higher (at the 0.05 significance level).

4. Conclusion and discussion

It turns out that despite the massive changes on the social network X, this social network is still actively used both by politicians and ordinary users. While politicians are only maintaining the increase that occurred during covid, users are increasingly interested in the activities that take place on X and significantly more like politicians' accounts.

This work contains hypotheses about global changes in the activity on the politicians accounst and about differences between countries with high and low GDP.

We have found that although the activity of the politicians is a little stagnating (in the number of tweets, following) and we can´t say, that the number of followers is growing, we can say, that users much more likes the politicians accounts.

However, countries with a lower GDP still lag behind countries with a higher GDP, both in the number of accounts that politicians have on X and in their activity.

Acknowledgement

Paper was processed with support from institutional-support fund for long-term conceptual development of science and research at the Faculty of Informatics and Statistics of the Prague University of Economics and Business (IP400040) and the project IG 409033.

References

Barberá, P. & Zeitzoff, T. (2018). The New Public Address System: Why Do World Leaders Adopt Social Media? International Studies Quarterly. Volume 62, Issue 1, Pages 121–130. https://doi.org/10.1093/isq/sqx047

Editors of Encyclopaedia Britannica. (2024). X | Company, History, Twitter, Elon Musk, & Uses. [online]. https://www.britannica.com/topic/Twitter

Eurostat (2024), File: Volume indices per capita, 2020-2022, (EU=100)V2.png – Statistics explained. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=File:Volume_indices_per_capita,_2020-2022,_(EU%3D100)V2.png

Fung, B. (2021). Twitter bans President Trump permanently. CNN Business. CNN [online]. https://www.cnn.com/2021/01/08/tech/trump-twitter-ban/index.html

Haman, M. & Školník, M. (2021). Politicians on Social Media. The online database of members of national parliaments on Twitter. Profesional de la información, v. 30, n. 2, e300217. https://doi.org/10.3145/epi.2021.mar.17

Larsson, A.-O. (2015). The EU Parliament on Twitter - Assessing the permanent online practices of parliamenta-rians. Journal of information technology and politics, v. 12, n. 2, pp. 149-166. https://doi.org/10.1080/19331681.2014.994158

López-Meri, A. Marcos-García, S. & Casero Ripollés, A. (2017). What do politicians do on Twitter? Functions and communication strategies in the Spanish electoral campaign of 2016. El Profesional de la Informacion. 26, 795-804. https://doi.org/10.3145/epi.2017.sep.02.

Pavlíček, Antonín. (2010). Nová média a sociální sítě. Praha: Oeconomica. ISBN 978-80-245-1742-1.

Pavlíček, A. & Syrovátková, J. (2022). Use of Twitter as an effective communication tool - case study on EU. In: Hradec economic days. Hradec Králové: University of Hradec Králové, s. 624--630. ISBN 978-80-7435-855-5.

Payne, A. (2009). The New Campaign: Social Networking Sites in the 2008 Presidential Election. Honors College Capstone Experience/Thesis Projects. [online]. Paper 204. http://digitalcommons.wku.edu/stu_hon_theses/204.

Safiullah, M., Pathak, P., Singh, S., Anshul, A. (2017). Social media as an upcoming tool for political marking effectiveness. Asia Pacific Management Review. roč. 22, č. 1, s. 10–15. ISSN 1029-3132.

Slížek, D. (2023). X (Twitter) nabízí dvě nové úrovně předplatného. Lupa.cz [online]. https://www.lupa.cz/aktuality/x-twitter-nabizi-dve-nove-urovne-predplatneho/.

Statista. (2024). Statista. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Syrovátková, J. & Korčák, J. (2023). Comparation of EU politics activity on Twitter (years 2020 and 2022). In: IDIMT-2023, New challenges for ICT and management, 31st Interdisciplinary informations management talks [online]. Linz: Trauner Verlag, Buchservice GmbH, s. 265--271. ISBN 978-3-99151-176-2. https://idimt.org/wp-content/uploads/2023/08/IDIMT-2023-proceedings.pdf

Vanian, J. (2022). Twitter is now owned by Elon Musk — here’s a brief history from the app’s founding in 2006 to the present. CNBC [online]. [cit. 2024-02-26]. https://www.cnbc.com/2022/10/29/a-brief-history-of-twitter-from-its-founding-in-2006-to-musk-takeover.html

X help center. (2024). About profile labels [online]. [cit. 2024-02-28]. https://help.twitter.com/en/rules-and-policies/profile-labels

USAGE OF CHATGPT WITH SPECIFIC PERSONAS AS A REPLACEMENT FOR REAL-LIFE RESPONDENTS FOR QUESTIONNAIRES

Jiří Korčák, Jana Syrovátková

Faculty of Informatics and Statistics
Prague University of Economics and Business
jiri.korcak@vse.cz, jana.syrovatkova@vse.cz

DOI: 10.35011/IDIMT-2024-309

Keywords

ChatGPT, large language models, AI, questionnaire)

Abstract

The rapid evolution of artificial intelligence has brought about innovative tools for data collection, among which ChatGPT, state-of-the-art language model, stands out. This study explores the feasibility of using ChatGPT, equipped to simulate detailed personas, as substitute for human respondents in questionnaires. Leveraging the capabilities of ChatGPT can potentially overcome traditional challenges associated with survey methodologies such as participant recruitment, cost, and time constraints. By programming ChatGPT to mimic specific personas, researchers can obtain high-quality data reflective of diverse respondent profiles without the logistical complexities of managing human participants. This paper presents comparative analysis of responses from human participants and those generated by ChatGPT, providing insights into the reliability and efficiency of using AI-driven methods in social research. The results suggest that ChatGPT can effectively emulate human-like responses, thereby supporting its use as an innovative tool in data collection processes.

1. Introduction

The rapid advancement of artificial intelligence has driven significant interest in automating tasks traditionally performed by humans, including data collection. Surveys and questionnaires, long-standing tools in social research, face challenges such as declining participation rates and high costs. This paper explores the use of chatbots, specifically ChatGPT, to address these challenges.

Chatbots, such as ChatGPT, offer scalability, cost efficiency, and high respondent satisfaction by simulating human conversation through text. These characteristics make chatbots a promising alternative for data collection. However, it is crucial to consider limitations such as the reproduction of typical personas without generating novel insights and the challenge of determining the representativeness of AI-generated responses.

The potential of chatbots in data collection is further supported by studies that have explored their application in various domains such as mental health (Abd-Alrazaq et al., 2021). ChatGPT - a program capable of generating human-like responses based on conversational input, could be considered a perfect tool to simulate question and response stages of a questionnaire collection (Kurban & Şahin, 2024).

This study examines the feasibility of using ChatGPT to simulate detailed personas as substitutes for human respondents in questionnaires, addressing both the advantages and limitations of this approach

2. ChatGPT technology.

GPT-4, the latest iteration from OpenAI, has been introduced to the public, showcasing significant advancements in AI text generation. Officially described as "A state-of-the-art AI that scales from a fine-tuned chatbot to a generative model capable of producing human-like text," this encapsulates the dual functionality of the model, which is central to its application in this study and similar projects which try to replace or enhance human while using chatbots (Alipour et al., 2024).

ChatGPT-4 marks a notable progression in bot technology, designed to handle both simple chat interfaces and complex, scalable text interactions. This evolution addresses some of the inconsistencies observed with its predecessor, GPT-3 (3.5), which, while effective in general conversation, often struggled with more nuanced dialogues such as role-playing or maintaining dynamic character exchanges. These challenges stemmed from GPT-3's broader focus, which wasn't solely on chat applications, underscoring the necessity for specialized testing to evaluate its performance across various conversational contexts (Alipour et al., 2024) (Floridi & Chiriatti, 2020).

2.1. Overview of ChatGPT Technology

ChatGPT, formally known as Generative Pre-trained Transformer, represents a significant advancement in language generation technology. Developed by OpenAI, this model was first introduced in 2018 and is a part of a lineage of models that have continually pushed the boundaries of text-based artificial intelligence capabilities. Initially building on the foundation laid by GPT-2, which itself was notable for its near-human text generation capabilities, ChatGPT has evolved considerably.

GPT-2, which was trained on a diverse array of web texts up to medium-length snippets, used perplexity as a measure of accuracy and was considered state-of-the-art at the time of its release. However, in comparison to its successor, GPT-3, which was unveiled in 2020, it might look like small baby trying to formulate its first words. GPT-3 marked a substantial scale-up in model architecture and training depth, employing 175 billion parameters compared to GPT-2's 1.5 billion. GPT-3's training utilized an autoregressive method with a language-modeling objective across a broad dataset, enabling enhanced capabilities in areas such as translation and text completion. (Brown et al., 2020).

With the release of GPT-4, OpenAI has taken another leap forward. This latest model builds on the strengths of GPT-3, offering further refined language understanding and generation abilities that enhance both the complexity and the subtlety of the AI’s textual outputs. GPT-4 boasts improvements in training techniques and model architecture, ensuring even greater efficacy in generating human-like text with minimal input data, thus broadening its potential applications in the AI field. This model's development was driven by the feedback and findings from its predecessors, aiming to achieve more nuanced and contextually accurate outputs (Malvankar et al., 2023)

3. Possible benefits of using ChatGPT for questionnaires

ChatGPT as an AI that can provide close and relevant answers to every question asked and which also can provide the ability to mimic a specific persona, can be beneficial to researchers. Researchers using ChatGPT need to prepare a set of questions complete with question logic. After that, the researcher just needs to run the AI, and ChatGPT will act like a real respondent depending on the given persona. This can save a lot of time because the researcher does not need to look for the right respondent and does not need to give instructions again from the basic functions of a respondent. In addition, compared to real-life respondents, ChatGPT has a cost efficiency value that is so high because the cost needed is only for the preparation of questions and the cost of running the AI, which is certainly still cheaper than paying a real respondent for each filled questionnaire (Lee et al., 2023).

For researcher using ChatGPT could be a great convenience – a 24/7 accessible conversational partner which can answer complex questionnaires is more convenient and easier to use than other data input methods or conversational agents or any other survey method. This may also increase participation and cooperation, and decrease the amount of break off in real life respondent studies using questionnaires (Lund & Wang, 2023).

This „hybrid-model" of using real life respondents and ChatGPT can aslo bring multiple advantages. Compared to paper-based and most online questionnaires, respondents do not have to read and understand each question, and then carefully select an answer. Input errors can be reduced by the agent prompting to roleplay the respondent when an unclear answer is given or when checking for incomplete responses. This enables on one side, a respondent to complete a questionnaire quickly and with little conscious mental effort, in which the quality of their responses may be compromised. On the other hand for researcher it can generate deeper insights into studied topic, as the questionnaire can consist from only handful of questions and chatbot can then help with furthering the research to deeper topics The conversational agent may be able to simulate a particular person, allowing the researcher to select the persona most similar to the respondent (Kennedy, 2023).

3.1 Increased accessibility

The availability of participants for research studies, particularly for interviews and questionnaires, can be a significant challenge due to time constraints and the need to find respondents with specific expertise or personal backgrounds. Interview-based studies are particularly time-consuming for both the participant and the researcher. However, the use of ChatGPT-generated respondents offers a potential solution to these challenges, as researchers can specify the persona of the conversational partner by providing relevant information, thereby enabling the generation of a conversational partner with a specific background. This approach can be particularly useful for running studies, as it provides access to conversational partners from around the world.

3.2 Time and cost efficiency

The implementation of ChatGPT in research methodologies, specifically in questionnaire design and execution, offers significant time and cost efficiencies. By leveraging ChatGPT as a conversational agent, researchers can drastically reduce the time typically required to conduct surveys (Sop, 2024). Traditional methods, such as paper-based or direct interviews, demand considerable preparation, distribution, and data entry efforts, which are substantially minimized when using an AI-driven approach (Gan, 2023). Furthermore, ChatGPT's ability to operate continuously without human oversight reduces labor costs and eliminates the constraints of geographical and temporal limitations.

This automation not only accelerates the data collection process but also enhances the scalability of studies, allowing researchers to handle larger datasets with fewer resources (Cooper, 2023).

4. Methods

4.1 Data collection

The basis of the data are the results from the questionnaire, where we collected data in from October to November 2020. It was a part of the bigger survey focused on social networks and privacy.

4.2 ChatGPT

To simulate respondents, we utilized the AI model ChatGPT. The model was designed to replicate human responses to our survey questions by first emulating a respondent profile based on previously gathered real-life data and then providing answers that such a profile would likely give.

The data from preceding research was split: the first segment contained responses to 53 coded information concerning demographic information, social media behavior, and IT-related practices. The second segment addressed 'deeper' questions (15) concerning behaviors on social media, such as friend networks, news sharing, and trust in the information found online, and also Facebook friends count and few other metrics.

Using the ChatGPT's Custom GPT feature with this simple configuration:

"After I receive information about respondent, I behave like it and first thing I do is give back answers to questionnaire:"

Followed by structure of the questions and answers, we synthesized five AI personas through a zero-shot prompting technique in one new ChatGPT chat window. These personas provided answers to the second set of 'deeper' questions as they were configured by the setting structure in custom GPT. Responses were given in both textual and numerical codes for comparison purposes. The responses were generated in the same structure as the actual responses were in.

To create the personas, we used detailed demographic and behavioral data. Each persona was defined by parameters such as age, gender, social media usage, and IT-related practices. The personas were designed to reflect a particular respondent from our study.

4.3 Research questions

For comparison data, we took ChatGPT's answers to questions about respondents' relationship with social media, whether they consider it a reliable source for getting news, how much they are influenced by friends and in turn influence friends. Furthermore, to what extent they share news, do they believe the news they come across on social media and whether the number of likes or comments is a sign of the credibility of the news. There were also questions about verifying messages on social media or sharing without checking the truth. Respondents expressed their agreement or disagreement with the statements on a five-point scale. Everyone answered 15 questions of this type.

We are trying to determine whether ChatGPT and human respond independently - or whether there is some relationship between human and ChatGPT responses.

4.4 Statistical hypothesis and methods

At the 0.05 significance level, we test the hypothesis that ChatGPT and human responses are not correlated. That is we test H0: $\rho = 0$ against H1: $\rho \neq 0$.

To validate the test, we use the correlation coefficient of the bivariate normal distribution. So we use a test statistic of the form:

$$T = \frac{R_{12}\sqrt{n-2}}{\sqrt{1-R_{12}{}^{2}}} \qquad (1)$$

where R12 is the sample correlation coefficient (calculated by Excel – Data Analysis – Correlation) and n is the sample range. We then compare the test statistic with the critical values of the student's t-distribution (according to statistical tables).

5. Results

5.1 Descriptive statistics

We have taken 5 personas and 15 parameters each. Each takes a value from 1 to 5. Thus, we obtained a total of 75 values (5 persons, 15 parameters each) from 1 to 5 each from human and AI. Table 1 shows the number of times each pair of responses from AI and human were shown. For example, in the first row and column labeled 2, we see that a total of 3 times the human answered 2 (agree) and the AI thought that the person would answer 1 (strongly agree). Another example is value 6 in first cell of the table represents those 6 times human wrote 1 and the AI also wrote 1.

This table presents a comparative analysis of responses from AI-generated personas and human respondents. The total count represents the aggregated answers across different questions, highlighting areas of alignment and divergence between AI and human responses.

Table 1. Comparation AI and human answers

		Human					
		1	**2**	**3**	**4**	**5**	**Total**
AI	**1**	6	3	1			10
	2		3	3	5		11
	3		2	8	7	6	23
	4		4	9	7	9	29
	5		1			1	2
	Total	6	13	21	19	16	75

Source: (author)

5.2 Statistic test

Using Excel Data Analysis we get the sample correlation coefficient $R_{12} = 0.5092$. Test statistics T we get the T = 5,055. As soon as we test H_0: $\rho = 0$ against H_1: $\rho \neq 0$, we take the critical value for student t-distribution $t_{0.\acute{\imath}75}(73) = 1.99$.

Because test statistics T > $t_{0.75}(73)$ we can reject the hypothesis that ChatGPT and human responses are not correlated at the 0.05 significance level.

5.3 Specific data analysis

Upon examining the specific answers provided by ChatGPT and comparing them with those from human respondents, several key findings emerged regarding their alignment and divergence:

- **Significant Divergence:** The largest divergence was observed in responses to the statement, "I do not trust any news I encounter on social media." Here, ChatGPT agreed four times, whereas human respondents disagreed three times and provided one neutral response. Conversely, where ChatGPT strongly disagreed, the human respondent agreed. The total absolute difference between AI and human responses was 10.
- **Notable Variance:** Another significant variance was found in responses to "I publish (re-post, re-tweet) interesting news that my friends or family posted on social media." In this case, ChatGPT's responses were predominantly neutral (disagreeing to strongly disagreeing) four times, and once disagreeing, while the human respondent (the same as in the previous question) agreed. The total absolute difference here was 9.
- **General Agreement:** For all other questions, the maximum sum of absolute differences in responses between AI and humans across five questions was 5, meaning that on average, ChatGPT's responses deviated by one degree per respondent per question.
- **Minimal Discrepancy:** The smallest discrepancy was noted in response to "If I think news is fake, I do not share it, no matter how interesting it might be." Here, ChatGPT assumed that all respondents would strongly agree, yet two respondents simply agreed.
- **Agreement on News Credibility:** Responses to "The more people who like and share a news item on social media, the more trustworthy it is," showed that ChatGPT disagreed or strongly disagreed, while in two instances, human respondents expressed more significant disagreement than AI.
- **Consistency in Less Variance:** Questions such as "Social media is a reliable source for news," "I consider social media a trustworthy news source if my friends/family use it as a news source," "If I am sure that the news I read on social media is true, I share it so my followers can also read it," and "I do not share news on social media even if I am convinced it is true," showed a total absolute difference in deviations between AI and humans of only 3. This indicates that in at least two questions, there was alignment, followed by differences of only 1 or 2 points.

These findings suggest that, at least in some cases, AI can very accurately predict responses to survey questions, closely mirroring human answers in the context of social media behavior and perception.

5.4 Population Variance

Given the small sample size of five personas, it is important to consider the potential impact of this limitation on our findings. The high degree of agreement between AI and human responses may be influenced by the limited variability within the sample. Future studies should expand the number of personas to better capture the diversity of human respondents and ensure the robustness of the conclusions drawn from the data.

6. Conclusion and Discussion

This study has explored the efficacy and reliability of utilizing ChatGPT as a substitute for human respondents in collecting questionnaire data. The findings reveal that the responses from ChatGPT exhibit a significant correlation with those from human respondents, confirming the initial hypothesis that there would be some correlation between the two sets of data. This correlation suggests that ChatGPT can mirror human-like interactions to a degree that they statistically resemble actual human responses in the context of social media behavior and perception.

To further enhance the accuracy and relevance of the responses generated by ChatGPT, it may be beneficial to refine the parameters used to define its personas. Currently, the model operates with a set of 53 predetermined responses encompassing demographics, social media behavior, and IT-related practices. Expanding these parameters to include more detailed data, such as responses from the second segment of our questionnaire which includes metrics like the number of Facebook friends and specific social media behaviors, could potentially refine the personas' depth and realism. By enriching the input data, ChatGPT could develop more nuanced understanding and responses, thereby improving its simulation of human-like interactions. This adjustment could lead to increase of the statistical correlation with human responses.

The implications of these results are profound for the field of data collection via surveys and questionnaires. Using ChatGPT could revolutionize the way researchers collect data, offering a cost-effective, efficient, and scalable method that mitigates many of the limitations associated with traditional survey methods, such as participant recruitment challenges and the high costs of data collection.

The integration of ChatGPT into survey methodologies presents notable technological benefits, including scalability and the ability to simulate diverse respondent profiles, which can enrich data collection significantly. However, this approach requires careful consideration of ethical and practical issues, such as ensuring transparency about the use of AI, and maintaining data privacy and security. Despite these advantages, ChatGPT may not fully replicate the depth of human emotional responses, which could be crucial for certain studies.

Future research should further validate ChatGPT's use across different domains and with varied demographic groups to assess its generalizability. Continued refinement of AI tools is essential to ensure their responsible and effective application in understanding complex human and societal behaviors. Building on the findings of this study, several avenues for future research can be pursued. Longitudinal studies should examine the consistency and reliability of AI-generated responses over time by programming ChatGPT to maintain persistent personas across multiple survey phases. Developing scenarios where ChatGPT interacts in complex dialogues, mimicking group discussions or debates, can help understand group dynamics and decision-making processes. Additionally, testing the ability of ChatGPT to emulate responses from diverse cultural backgrounds accurately will assess its cross-cultural validity, tailoring the AI to reflect cultural nuances in communication styles and values. Exploring how well ChatGPT can generate responses that appropriately convey emotional nuances is crucial for fields like psychology and marketing. Finally, conducting comparative studies involving ChatGPT and other AI models will evaluate strengths and weaknesses in different survey contexts, further integrating AI tools into the toolkit of social scientists and researchers.

Acknowledgement

Paper was processed with support from institutional-support fund for long-term conceptual development of science and research at the Faculty of Informatics and Statistics of the Prague University of Economics and Business (IP400040) and the project IG 409033.

References

Abd-Alrazaq, A., Alajlani, M., Ali, N., Denecke, K., Bewick, B., & Househ, M. (2021). Perceptions and opinions of patients about mental health chatbots: scoping review. Journal of Medical Internet Research, 23(1), e17828. https://doi.org/10.2196/17828

Alipour, H., Pendar, N., & Roy, K. (2024). ChatGPT Alternative Solutions: Large Language Models Survey. https://doi.org/10.5121/csit.2024.140514

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … & Amodei, D. (2020). Language models are few-shot learners. https://doi.org/10.48550/arxiv.2005.14165

Cooper, G. (2023). Examining science education in chatgpt: an exploratory study of generative artificial intelligence. Journal of Science Education and Technology, 32(3), 444-452. https://doi.org/10.1007/s10956-023-10039-y

Floridi, L. and Chiriatti, M. (2020). Gpt-3: its nature, scope, limits, and consequences. Minds and Machines, 30(4), 681-694. https://doi.org/10.1007/s11023-020-09548-1

Gan, R. (2023). Chatgpt's performance before and after teaching in mass casualty incident triage. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-46986-0

Kennedy, D. (2023). Adding a „Group Advisory Layer" to Your Use of Generative AI Tools Through Structured Prompting: Using Personas for Advisory Boards, Task Forces, Mastermind Groups, and Other Collections of Personas to Assist in Evaluations, Assessments, Recommendations, Decision-making, and much more (Including Law-related Examples) (SSRN Scholarly Paper 4570860). https://doi.org/10.2139/ssrn.4570860

Kurban, C. F., & Şahin, M. (2024). Findings and Interpretation. In The Impact of ChatGPT on Higher Education (pp. 93–131). Emerald Publishing Limited. https://doi.org/10.1108/978-1-83797-647-820241006

Lee, V. V., van der Lubbe, S. C. C., Goh, L. H., & Valderas, J. M. (2023). Harnessing ChatGPT for thematic analysis: Are we ready? (arXiv:2310.14545). arXiv. https://doi.org/10.48550/arXiv.2310.14545

Lund, B. and Wang, T. (2023). Chatting about chatgpt: how may ai and gpt impact academia and libraries?. Library Hi Tech News, 40(3), 26-29. https://doi.org/10.1108/lhtn-01-2023-0009

Malvankar, K., Fallon, E., Connolly, P., & Flanagan, K. (2023). Performance Optimization for Transformer Models on Text Classification Tasks. 2023 International Conference on Emerging Techniques in Computational Intelligence (ICETCI), 105–111. https://doi.org/10.1109/ICETCI58599.2023.10330958

Sop, S. (2024). What if chatgpt generates quantitative research data? a case study in tourism. Journal of Hospitality and Tourism Technology, 15(2), 329-343. https://doi.org/10.1108/jhtt-08-2023-0237

SOCIAL MEDIA CONTENT MODERATION, CENSORSHIP AND AI DETECTION EVASION TECHNIQUES

Lukáš Malec

Prague University of Economics and Business
Faculty of Informatics and Statistics
mall05@vse.cz

Jan Lešetický

Pan-European University
Faculty of Business
jan.leseticky@peuni.cz

DOI: 10.35011/IDIMT-2024-317

Keywords

Social media; Facebook; detection evasion techniques; harmful content; moderation systems

Abstract

Social media platforms have deployed a variety of advanced tools to moderate content and maintain safe environments for users. However, as AI-driven moderation tools have evolved, so have the strategies of malicious content creators. This article discusses the history of media content moderation and, in conclusion, identifies the following 10 specific methods used to evade detection by AI systems: content alteration, blurry images, contextual reasoning, external or malicious links unchecked, extra spacing, misleading captions, missed by AI detection, image alteration, language obfuscation, and visual text embedding. This article underscores the ongoing battle between content moderation efforts and malicious content creators. The dynamic nature of this struggle highlights the necessity for continuous improvement and innovation in AI technologies, complemented by human oversight.

1. Introduction

With technological advances in the last century and the introduction of the terminology "interoperability," information technology and systems started playing a prominent role in most, if not all, of the existing conceptual and practical models (Motta and Fowler, 2016). A clear example could be observed in the application of information technology to facilitate media and hence be the primary tool in moderation and/or censorship, and in the last decade, the application of platform governance by artificial intelligence and algorithm-enhanced content moderation.

Social media platforms today possess a multitude of tools and methods to moderate content, ensuring that their spaces remain safe and respectful for users. These tools include sophisticated algorithms and artificial intelligence (AI) systems designed to detect and remove harmful content. AI tools such as natural language processing (NLP), image recognition, and machine learning models are employed to scan posts, comments, and media for violations of community guidelines. Companies like Facebook, X (former Twitter), and YouTube use these technologies to identify hate speech, graphic violence, misinformation, and other forms of harmful content quickly and efficiently.

However, as these AI-driven moderation tools have become more advanced, creators of inappropriate or harmful content have also evolved. They have developed various tactics to evade content moderation. For instance, they might use coded language, symbols, or slight alterations in spelling to bypass text-based filters. In image and video content, they may obscure harmful elements or use editing techniques to avoid detection by AI systems. This cat-and-mouse game between content moderators and those seeking to circumvent the rules has made it increasingly challenging to maintain the integrity of social media platforms.

As a result, while AI tools have significantly improved social media companies' ability to moderate content at scale, the adaptability of malicious content creators highlights the need for continuous improvement and innovation in moderation technologies. It also underscores the importance of human oversight to complement AI systems, ensuring that nuanced and context-specific judgments can be made where AI might fall short. This dynamic landscape demands a multi-faceted approach to content moderation, combining cutting-edge technology with human expertise to effectively combat harmful content online.

1.1 Literature review on social communication, censorship, and content moderation

Information systems governance of most of the existing censorship tools emphasized the need for further studies to assess and understand the relationship and impact information systems have with other stakeholders (governments, citizens, platforms, users) and how it affects the efficiency of censorship and freedom of speech (Burton, 1995; Cobbe, 2020).

As noted by Hannabuss (1994), in modern society, discussions surrounding information flow, access, and enabling technologies often reveal that information is open to more than just unbiased and objective facts and data. Instead, it is frequently laden with values, meanings, and subjectivity, especially when different parties with competing interests are involved. Information may pertain to sensitive topics or controversial issues, leading to concerns about information control at this stage.

Moreover, it is an area rife with assumptions and prejudices regarding what is obscene, offensive, and worthy of research. It is sometimes caught between the opinion that the obscene should be kept private and not made public. However, scientific belief encourages that thorough and open discussion and evaluation are vital in a free and educated society.

Some opinions state that the easy availability of knowledge has become a contemporary virtue and truism. The global community generates, disseminates, retrieves, and consolidates information, such as news, in never-before-seen ways. Activities such as computer-assisted journalism have changed media infrastructures that were already undergoing rapid change (Hannabuss, 1994).

However, since the Internet arrived at the forefront of public consciousness in 1994, general users have been concerned about the perils of freely available Internet pornography, violence, and hateful speech. Mass media stories describing pornographic websites, hate groups, and online sexual predators fuelled this concern (Hunter, 2009).

In the 2010s, the initial concept of the Internet as a digital version of the traditional open public forum or agora has undergone a transformation. It has evolved from a space primarily envisioned for the free exchanging of ideas and opinions to a concrete commercial domain. In this evolved form, it's an arena where business entities prioritize creating a space that is not only safe and dignified but also without any offensive material.

1.2 Traditional and AI moderation of problematic content

This research builds upon the existing body of literature examining the evolution of media censorship and the emergence of social media platforms, governing policies, and the implementation of AI moderation.

The digital era has introduced new forms of censorship, particularly through Internet filters. Burton (1995) critically examines the feasibility and necessity of regulating and controlling the Internet, foreshadowing many challenges that would grow with the Internet's expansion.

Cooke (2007) analyzes European content and access regulation approaches, revealing the continental philosophical and legal differences in Internet governance. Deibert's edited volume (2008) provides a thorough overview of the practice and regulations of worldwide Internet filtering, shedding light on digital censorship's social, political, and technological aspects.

Among others, Dias Oliva (2020) explores content moderation technologies and their alignment with human rights standards, a critical perspective on the governance of expression. The EU's digital services regulation efforts, including the Digital Services Act (2023), represent significant moves toward rewriting Internet governance rules, reflecting a legislative response to the complexities of digital regulation.

Roberts (2020) focuses on the concept of resilience to online censorship, discussing how users and content providers adapt and respond to censorship, which can lead to a perpetual cycle of action and reaction between censors and the censored.

These works contribute to a nuanced understanding of censorship, its justifications, mechanisms, and challenges. The literature demonstrates the dynamic interplay between societal values, legal frameworks, technological capabilities, and individual rights, underscoring the complexity of censoring and regulating media in both the traditional and digital realms.

The concept of "platforms" as political and social constructs is thoroughly analyzed by Gillespie (2010), who delves into the inherent politics of platform operations and the implications for content and social dynamics. Extending this discourse, Gillespie (2018) explores the intricate decisions involved in content moderation, framing platforms as custodians of Internet content with significant influence over digital discourse.

Moreover, Roberts (2019) takes a behind-the-scenes look at content moderation, revealing the shadowy aspects of social media that impact both the users and the platforms themselves. Together, these references frame a complex picture of social media platforms as entities that shape and are shaped by content moderation practices, reflecting a blend of technological, sociopolitical, and ethical considerations.

The challenges of content moderation in the context of misinformation, especially during critical times like a pandemic, are highlighted by Baker, Wade and Walsh (2020). Binns et al. (2017a) address the technical aspects of algorithmic moderation, specifically the biases that can be transferred from human trainers to bots. Dias Oliva (2020) discusses the potential of aligning the potential to protect liberties by aligning content moderation procedures with human rights principles.

On the other hand, Gillespie (2020) raises pertinent questions about the scalability of AI in content moderation, examining the extent to which AI can manage the vast and complex landscape of online content. Harwell (2021) discusses the promises and challenges of employing AI to tackle misinformation, indicating a trend toward technological solutions to content moderation.

Ruckenstein and Turunen (2020) suggest re-humanizing platform logic by integrating care into content moderation practices, while Ytre-Arne and Moe (2021) introduce the concept of "folk theories" to understand public perceptions of algorithms. Zeng and Kaye (2022) pivot from content to visibility moderation, using TikTok as a case study to demonstrate how algorithmic governance shapes user visibility.

These studies collectively underscore the duality of content moderation, which is neither fully human nor fully automated but rather a complex interplay between human values, biases, and the capabilities of artificial intelligence.

Content moderation has evolved as a critical aspect of social media governance. Binns et al. (2017b) investigated the inherited biases in algorithmic content moderation, revealing that automated systems often reflect the prejudices in their training data. This aligns with Gillespie's (2010) examination of the political nature of platforms, which theorizes that content moderation is not merely a technical task but one deeply entwined with socio-political values.

The consequences of platform content moderation are multifaceted. Baker, Wade, and Walsh (2020) discuss the dilemmas faced when moderating misinformation, especially during a pandemic, challenging the traditional notions of harm. Gillespie (2015) continues this dialogue by exploring the platforms' interventionist role take in content governance, shaping user discourse and social norms.

These studies collectively suggest that platform content moderation is a complex interplay of technical challenges, human judgment, legal considerations, and ethical dilemmas. They also indicate the significant impact of these moderation practices on user experience, platform trust, and the broader socio-political discourse.

Meanwhile, the European Union's initiatives on protecting minors and human dignity in audiovisual and information services reflect a legislative approach to content regulation, as detailed in the EUR-Lex summary (EUR-Lex - l24030 - EN). This legal framework is compared with perspectives from "Filters & Freedom 2.0," which debates the ethics and effectiveness of Internet content controls.

1.3 Facebook content moderation practices

Facebook's content moderation practices are a practical example of the challenges in the broader debate on Internet freedom and regulation. Biddle (2018, 2021) explores how Facebook's ad targeting system and secret blocklist can both reflect and perpetuate societal issues, such as racism and the classification of "dangerous" groups or individuals.

Eslami et al. (2015) examine user perceptions of Facebook's algorithms, revealing a disconnect between user expectations and the platform's content delivery mechanisms. The need for transparency in algorithmic content ranking is further stressed by New America and highlights the broader call for accountability in automated decision-making systems.

As reported by Goel et al. (2018), incidents like the shutdown of Facebook in Sri Lanka bring attention to the real-world consequences of online content and the social responsibility of platforms in conflict situations. This is compounded by the OHCHR's (2019) findings on the role of social media in human rights abuses in Myanmar.

Research by the Centre for Policy Alternatives (2014) and Myers West (2018) delves into the implications of hate speech on Facebook, suggesting that the platform's content moderation practices can struggle to address complex social issues adequately. The Wall Street Journal (2021) and Windwehr and York (2021) detail the shortcomings and pitfalls of Facebook's automated content moderation, underscoring the limitations of current technologies in understanding context and nuance.

In conclusion, Ytre-Arne and Moe (2021) contribute to the understanding of public sentiment toward algorithmic content moderation by introducing the concept of "folk theories of algorithms," which reflects the common misconceptions and frustrations experienced by users.

2. Methodology

In our research, we wanted to inspect the application and implications of their findings on the specific case of content moderation in Facebook, mainly focusing on the nuanced dynamics of social media content moderation and its broader societal impacts.

Our main research question was the effectiveness of implemented moderation technologies in enforcing policies for user-generated content on social media (e.g., Facebook). Our research, executed in early 2024, employed a snowball sampling technique to analyze a dataset of social media posts (N=500) characterized by terrorism content, extreme violence, explicit sexual content, and spam.

While snowball sampling is efficient and practical for qualitative research, it is acknowledged that this method introduces potential biases. These biases stem from the chain-like nature of the sampling process, which could lead to a homogeneity within the sample.

- The core limitation of this methodological approach is that it examines AI detection evasion techniques of the sample, which do not exclusively represent all the existing pages and group techniques to evade AI moderation on Facebook.
- Facebook non-disclosure of flagged or removed content repositories for research purposes.

The ethical considerations, especially given the sensitive and dangerous nature of the content, were addressed meticulously. The research adhered to strict ethical guidelines, ensuring anonymity and appropriate handling of the data throughout the research.

Subsequent selections of posts were based on their relevance and connection to the initially identified material. This iterative process continued, expanding the dataset while focusing on content relevant to AI moderation evasion techniques.

Data labeling followed strict criteria by Facebook-published guidelines and policies for Spam, Adult Sexual Exploitation, Violence and Incitement, and Dangerous Organizations and Individuals (Facebook, 2021) to maintain consistency and accuracy. The labeling process ensured that each piece of content was categorized and labeled as precisely as feasible concerning the research's parameters.

3. Results

The research utilizes the collected dataset that encapsulates the labeling of publicly accessible posts. This dataset comprises various attributes, offering insights into how users might attempt to bypass automated content moderation systems. Table 1 provides the list and explanation of AI Detection Evasion techniques identified from the dataset.

3.1 Discussion

The dataset analyzed for this research presents a range of diverse content, each representing a distinct theme frequently recurring in online content moderation. These themes encompass a spectrum of issues, from the misuse of personal images and the spread of extremist propaganda to the promotion of conspiracy theories and hate speech.

Table 1. AI Detection Evasion Techniques

AI Detection Evasion Method	Description	Purpose
Content Alteration	This involves modifying the content slightly in a way that changes its appearance or context to AI but remains recognizable or understandable to humans. It can include altering images, text, or videos.	To bypass AI detection systems that rely on specific patterns or keywords.
Blurry Image	Using images that are intentionally blurred or of low quality.	To avoid detection by AI systems that analyze visual content, as these systems often require clear, high-quality images to accurately identify content.
Contextual Reasoning	To correctly interpret the content requires an understanding of context, background knowledge, or subtle nuances.	AI systems still struggle with complex contextual understanding, allowing such content to evade detection.
External/Malicious Link Unchecked	Including links in posts that lead to external or malicious sites. This may not be immediately analyzed or flagged by AI systems.	To bypass content checks within the platform, the harmful or rule-violating content is hosted outside the direct purview of the AI.
Extra Spacing	Inserting unusual or extra spaces in the text, especially within keywords.	To fool AI detection algorithms searching for specific keywords or phrases, as the altered spacing disrupts these patterns.
Misleading Caption	Using unrelated or misleading captions relative to the actual content.	To mislead AI systems that use text analysis for content understanding, mainly when the text is used to gauge the context of images or videos.
Missed by AI Detection	Content that should have been flagged by AI systems, but was not due to various limitations in the AI's detection capabilities.	It is not an intentional evasion technique but rather an indication of areas where AI systems need improvement.
Image Alteration	Modifying images slightly, such as changing colors, adding noise, or cropping.	To evade AI detection systems that rely on image recognition algorithms by altering key features or patterns.
Language Obfuscation	Deliberate use of uncommon languages, code words, or complex linguistic structures.	To avoid detection by AI systems primarily trained on more common languages or straightforward text.
Visual Text Embedding	Embedding text within images in a way that's readable by humans but not easily detectable by AI.	To bypass AI systems that analyze text and images separately, as the AI may not recognize the embedded text as text.

Source: (authors)

Our research has identified ten distinct methods that individuals employ to evade the detection of harmful content on social media platforms. One common technique is content alteration, where users slightly modify their content to avoid keyword detection. Similarly, extra spacing between letters or words and language obfuscation using slang or alternate spellings help bypass text-based AI filters.

These tactics exploit the rigidity of AI systems, which often rely on specific patterns to identify harmful content.

In addition to text manipulation, visual content poses significant challenges. Blurry images and image alteration are frequently used to prevent visual recognition algorithms from identifying harmful elements. Furthermore, visual text embedding, where text is included within images or videos, makes it difficult for AI systems to detect harmful messages through standard text recognition techniques. Misleading captions can also deceive AI, presenting content that appears innocuous at first glance but is harmful when properly understood.

Lastly, some strategies exploit the context in which content is posted. Contextual reasoning involves creating content that seems harmless out of context but is harmful when fully comprehended. Users also share external or malicious links that AI systems might not immediately check, leading to the dissemination of harmful content. By understanding these methods, we can better appreciate the ongoing challenge of content moderation and the need for continuous advancements in AI technologies to keep up with these evolving tactics.

4. Conclusion

These findings underscore a significant and intriguing conclusion: The battle between content moderation and harmful content creators is a dynamic and ongoing struggle. As AI technologies advance, so do the methods of those seeking to circumvent them. This cat-and-mouse game highlights the importance of improving AI capabilities and also integrating human oversight and contextual understanding into moderation efforts. The future of safe and respectful online spaces relies on a multi-faceted approach, blending cutting-edge technology with human judgment to outpace those intent on spreading harmful content.

Our analysis is crucial in understanding the challenges of moderating such content, which often treads the fine line between freedom of expression and the propagation of harmful, misleading, or offensive material. For example, considering disclosure of flagged or removed content repositories for any purpose is perceived as dangerous. However, the repetitive occurrence of the harmful themes in the dataset underscores their prevalence in digital spaces and emphasizes the complexity and sensitivity required to address them effectively.

References

Baker, S. A., Wade, M., & Walsh, M. J. (2020). The challenges of responding to misinformation during a pandemic: Content moderation and the limitations of the concept of harm. Media International Australia, 177(1), 103–107. https://doi.org/10.1177/1329878X20951301

Biddle, S. (2018, November 2). Facebook Allowed Advertisers to Target Users Interested in "White Genocide"—Even in Wake of Pittsburgh Massacre. The Intercept. https://theintercept.com/2018/11/02/facebook-ads-white-supremacy-pittsburgh-shooting/

Biddle, S. (2021, October 12). Revealed: Facebook's Secret Blacklist of "Dangerous Individuals and Organizations". The Intercept. https://theintercept.com/2021/10/12/facebook-secret-blacklist-dangerous/

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017a). Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In G. L. Ciampaglia, A. Mashhadi, & T. Yasseri (Ed.), Social Informatics (s. 405–415). Springer International Publishing. https://doi.org/10.1007/978-3-319-67256-4_32

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017b). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. arXiv:1707.01477 [Cs], 10540, 405–415. https://doi.org/10.1007/978-3-319-67256-4_32

Burton, P. F. (1995). Regulation and control of the Internet: Is it feasible? Is it necessary? Journal of Information Science, 21(6), 413–428. https://doi.org/10.1177/016555159502100602

Cobbe, J. (2020). Algorithmic Censorship by Social Platforms: Power and Resistance. Philosophy & Technology. https://doi.org/10.1007/s13347-020-00429-0

Cooke, L. (2007). Controlling the net: European approaches to content and access regulation. Journal of Information Science, 33(3), 360–376. https://doi.org/10.1177/0165551506072163

Deibert, R. (Ed.). (2008). Access denied: The practice and policy of global Internet filtering. MIT Press.

Dias Oliva, T. (2020). Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression. Human Rights Law Review, 20(4), 607–640. https://doi.org/10.1093/hrlr/ngaa032

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). „I always assumed that I wasn't really that close to [her]": 33rd Annual CHI Conference on Human Factors in Computing Systems, CHI 2015. CHI 2015 - Proceedings of the 33rd Annual CHI Conference on Human Factors in Computing Systems, 153–162. https://doi.org/10.1145/2702123.2702556

EUR-Lex—L24030—EN - EUR-Lex. (b.r.). Loaded 2023, November 4, from https://eur-lex.europa.eu/EN/legal-content/summary/protection-of-minors-and-human-dignity-in-audiovisual-and-information-services-green-paper.html

Gillespie, T. (2010). The politics of 'platforms'. New Media & Society, 12(3), 347–364. https://doi.org/10.1177/1461444809342738

Gillespie, T. (2015). Platforms Intervene. Social Media + Society, 1(1), 2056305115580479. https://doi.org/10.1177/2056305115580479

Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. In Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media (p. 288). https://doi.org/10.12987/9780300235029

Gillespie, T. (2020). Content moderation, AI, and the question of scale. Big Data & Society, 7(2), 2053951720943234. https://doi.org/10.1177/2053951720943234

Goel, V., Kumar, H., & Frenkel, S. (2018, March 8). In Sri Lanka, Facebook Contends With Shutdown After Mob Violence. The New York Times. https://www.nytimes.com/2018/03/08/technology/sri-lanka-facebook-shutdown.html

Hannabuss, S. (1994). Explicit representations: Approaches to censorship. Aslib Proceedings, 46(10), 249–256. https://doi.org/10.1108/eb051372

Harwell, D. (2021, December 5). AI will solve Facebook's most vexing problems, Mark Zuckerberg says. Just don't ask when or how. Washington Post. https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/ai-will-solve-facebooks-most-vexing-problems-mark-zuckerberg-says-just-dont-ask-when-or-how/

Hunter, C. (2009). Internet Filter Effectiveness: Testing Over and Underinclusive Blocking Decisions of Four Popular Filters. Independent International Fact-Finding Mission on Myanmar. (2019, June 2). OHCHR. https://www.ohchr.org/en/hr-bodies/hrc/myanmar-ffm/index

Liking violence: A study of hate speech on Facebook in Sri Lanka. (2014, září 24). Centre for Policy Alternatives. https://www.cpalanka.org/liking-violence-a-study-of-hate-speech-on-facebook-in-sri-lanka/

Motta, M. P., & Fowler, E. F. (2016). The Content and Effect of Political Advertising in U.S. Campaigns. In M. P. Motta & E. F. Fowler, Oxford Research Encyclopedia of Politics. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228637.013.217

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. New Media & Society, 20(11), 4366–4383. https://doi.org/10.1177/1461444818773059

Roberts, M. E. (2020). Resilience to Online Censorship. Annual Review of Political Science, 23(1), 401–419. https://doi.org/10.1146/annurev-polisci-050718-032837

Roberts, S. T. (2019). Behind the Screen: Content Moderation in the Shadows of Social Media. Yale University Press. https://doi.org/10.2307/j.ctvhrcz0v

Ruckenstein, M., & Turunen, L. L. M. (2020). Re-humanizing the platform: Content moderators and the logic of care. New Media & Society, 22(6), 1026–1042. https://doi.org/10.1177/1461444819875990

The Digital Services Act package. Shaping Europe's digital future. (2023, October 23). https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

The Facebook Files. (2021, October 1). Wall Street Journal. https://www.wsj.com/articles/the-facebook-files-11631713039

Windwehr, S., & York, J. (2021, October 8). Facebook's Most Recent Transparency Report Demonstrates the Pitfalls of Automated Content Moderation. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2020/10/facebooks-most-recent-transparency-report-demonstrates-pitfalls-automated-content

Ytre-Arne, B., & Moe, H. (2021). Folk theories of algorithms: Understanding digital irritation. Media, Culture & Society, 43(5), 807–824. https://doi.org/10.1177/0163443720972314

Zeng, J., & Kaye, D. B. V. (2022). From content moderation to visibility moderation: A case study of platform governance on TikTok. Policy & Internet, 14(1), 79–95. https://doi.org/10.1002/poi3.287

DATA AND AI IN SUPPLY CHAIN MANAGEMENT

EXPLORING MULTIDIMENSIONAL BIG DATA ANALYTICS IN PROCUREMENT FOR DETECTING COLLUSION

Radoslav Delina, Gabriel Demeter

Technical University of Kosice
Faculty of Economics
Radoslav.Delina@tuke.sk, Gabriel.Demeter@tuke.sk

DOI: 10.35011/IDIMT-2024-329

Keywords

Procurement; collusion; group of suppliers, big data, transparency

Abstract

This paper explores the integration of data-driven approaches to detect and mitigate collusive behaviours in public and commercial procurement processes. As procurement represents a significant investment and is vulnerable to corruption and collusion, advanced analytical tools and machine learning algorithms are essential in promoting fair competition and integrity. This study details basic approaches how to analyse different dimensions of collusive tactics, illustrating on simple visualizations how these practices undermine market efficiency by stifling competition and inflating prices. By leveraging large datasets, this research identifies patterns and anomalies indicative of collusion and fraudulent activities among suppliers through different dimensions of analysis from analysing the history of specific supplier or his position in stabilized supplier portfolio. Through a comprehensive analysis of procurement data, including financial records and supplier performance histories, this paper open discussion on the potential of data science approaches in transforming procurement processes by ensuring competitive fairness and economic efficiency.

1. Introduction

In today's global business environment, public and commercial procurement processes play a crucial role in economic development (Relucio & Cruz, 2020). Efficient and transparent procurement practices are essential for promoting fair competition, preventing corruption, and ensuring that taxpayers' money is used effectively (Kassim & Hussin, 2010). Public procurement is a crucial process that involves the acquisition of goods and services by government agencies or public institutions. These purchases often require a significant investment and necessitate careful consideration of various factors such as price, quality, and timeliness (Kassim & Hussin, 2010). Similar problems are emerging also in commercial procurement, where collusive behaviour increases cost of the procuring organisation leading to lower efficiency and potential competitiveness. In recent years, there has been growing concern about the potential for collusion and corruption in procurement and negotiation processes, particularly in relation to bidding strategies employed in sealed bid and English auctions (Bardhi, 2014; Urminsky et al, 2023). To address these challenges, governments and organizations have begun to leverage data-driven approaches for the detection and regulation of collusive signals in procurement activities (Kiener-Manu, 2018).

2. Collusive Signals in Public and Commercial Procurement

One of the major challenges in public and commercial procurement is detecting and preventing collusion among suppliers leading to negative procurement performance. Collusion refers to an agreement between two or more suppliers to manipulate the procurement process in their Favor. Collusion can result in a lack of competition, inflated prices, and substandard products or services being procured. Collusive behaviour can take on various forms, and understanding these strategies is crucial in combating corruption and promoting fair competition in public procurement. In negotiation, collusive bidding strategies frequently include terms like cover bidding, bidding rings (as more general term), bid shading, bid withdrawal, bid-jumping, bid suppression, bid rotation or market allocations (Qi et al., 2020; OECD, 2012).

Cover bidding, also known as complementary bidding, occurs when companies agree that one will submit the lowest bid to secure the contract while the others submit artificially high or technically unacceptable bids to give the appearance of competition, but in reality, the competition is rigged. The goal is to ensure that the selected company wins without being threatened by actual competition, with other companies "covering" this operation by submitting non-competitive bids. In bid shading, colluding bidders deliberately submit bids that are lower than their actual value in order to decrease the chances of winning the contract while still maintaining the appearance of competition. This is a competitive tactic used by a single bidder independently, without any collusion with others. It involves bidding slightly less than what the bidder believes the product or service is worth to try to get a better deal while still winning the bid. One common collusive strategy in both sealed bid and English auctions is submitting bids that are very close to each other. (Kiener-Manu, 2018; Qi et al., 2020; OECD, 2012; Czibik et al, 2014). Another collusive strategy is the withdrawal of bids (Kassim & Hussin, 2010; Czibik et al, 2014). This tactic involves bidders intentionally withdrawing their bids after observing the bids of other competitors, allowing a pre-agreed bidder to secure the contract (Kiener-Manu, 2018). Another strategy employed by colluding bidders is the use of larger distance between the last bid and the bid second in the final order (Bardhi, 2014). This strategy, referred to as bid-jumping, is used to create the appearance of competition while still ensuring that a specific bidder wins the contract at a predetermined price (Kiener-Manu, 2018). Bid-jumping occurs when a bidder submits a significantly higher bid than the current leading bid to disrupt the auction and dissuade other bidders from participating further. In addition, another collusive strategy is the deliberate submission of a small number of bids. This strategy, known as bid suppression, involves colluding bidders intentionally submitting fewer bids than would be expected, thereby reducing competition and increasing their chances of winning the contract. Furthermore, there have been cases where bidders engage in bid rotation, a strategy where competitors take turns being the winning bidder to maintain the appearance of competition while ensuring that all colluding parties benefit from the contracts and managing market allocation in different forms.

There are also other type of signals like similarities in offer description between suppliers where check of plagiarism or same text strings can be a signal of agreement before the bidding process (Bardhi, 2014). These bidding strategies are often illegal and considered anti-competitive because they reduce market efficiency and harm customers or the public by increasing prices and lowering the quality of services or goods provided. It is crucial for regulatory and oversight bodies to be able to identify and act against such practices to protect a healthy competitive environment.

To tackle this issue, data-driven approaches are being employed to identify and regulate collusive signals in procurement activities. Data-driven approaches for detecting and regulating collusive signals in procurement activities involve the use of advanced analytical tools and algorithms to analyze large volumes of procurement data. By leveraging techniques such as data mining, machine learning, and predictive analytics, organizations can uncover patterns and anomalies that may indicate collusion among suppliers. Furthermore, data-driven detection and regulation of collusive signals extend beyond identifying potential collusion within procurement processes. It also encompasses the assessment of broader business and supplier risk management. By integrating data from various sources, including financial records, supplier performance history, and market intelligence, organizations can gain a comprehensive understanding of the risks associated with their procurement activities. This proactive approach enables organizations or procurers to take preventive measures to mitigate these risks.

3. Data-driven detection for preventive measures

Overall, the integration of data-driven approaches for detecting and regulating collusive signals in procurement activities represents a significant advancement in promoting fairness, integrity, and efficiency in public and commercial procurement. By delving deeper into the analysis of procurement data and embracing proactive risk management strategies, organizations can not only combat collusion but also fortify their procurement processes against a multitude of potential risks.

Collusive signals are one element of business or supplier risk management, where the impact of such a risk exposure can be significant. The supplier risk management in this context involves following measures such as:

- Analysing financial records and credit ratings, organizations can assess the financial health of their suppliers. This evaluation allows organizations to identify suppliers with low business intensity, where such a companies can server more as cover bidders as like normal business partner with potential fraudulent activity (Fakir, 2015).
- Tracking and analysing the performance history of their suppliers to identify any signs of collusion or unethical practices (P.R, 2018).
- Performing thorough background checks on potential suppliers to identify any past involvement in collusive activities or fraudulent behaviour (Zage et al., 2013).
- Establishing rigorous processes for selecting suppliers, including comprehensive due diligence and evaluation criteria, to minimize the risk of collusion and ensure that only reputable suppliers are chosen.

Additionally, organizations can leverage advanced analytics solutions to scan and identify noncompliance in transaction-intensive purchases where manual analysis may fall short.

The main approaches how to detect and regulate collusive signals in public and commercial procurement using data-driven methods including advanced analytics which can scan large amounts of procurement data and identify patterns or anomalies that may indicate collusive behaviour or machine learning approaches. Generally, a prerequisite is the existence of indicators signalizing collusive behaviour and be able to identify or visualise them (Bardhi, 2014).

From an overall efficiency perspective, it is more suitable to eliminate collusive signals with preventive measures rather than waiting for legal proof in court. Removing a signal or non-standard behaviour is much more effective and does not always depend on 100% accuracy in identifying the collusive signal but rather on increasing the likelihood of a positive impact of the measure on enhancing competition, which is a standard feature of supporting the efficiency of procurement performance.

One of the significant advantages of big data and artificial intelligence is the ability to compare potentially correct contracts with those influenced by collusive behaviour of suppliers. Simplified assertions such as a small number of suppliers, e.g., three or fewer, or a low variance coefficient, can be misleading due to the high complexity of the business environment, sector specifics such as in areas characterized by time or more permanently low margins, or low business attractiveness to companies, thus a lower number of potential suppliers. It is therefore very important not to distort the application of data science and the interpretation of results in such simplifications. Here, machine learning techniques with integrated causal algorithms come to the rescue, allowing for the analysis of comparable records in the data, thereby increasing the objectification of the results.

4. Analysis of specific behaviour

Of course, in order to expand the perspective on the validation of unfair practices informally using some indicator, it is necessary to explore various signals, their algorithmic possibilities, and the frequency of records subject to unfair practices. The prevalence of records indicating an extremely high proportion of non-standard signals such as low competition, weak savings, etc., can mean not only a cartel or collusion but also a peculiarity of a given sector or product line. However, if we have enough variability in tender results, it is possible to apply benchmarking approaches to enhance the validation of identifying non-standard, likely unfair approaches or strategies in supply or negotiation.

Applying visualisations on different signals or indicators can help to reveal potential collusive behaviour or decide on more efficient data science method to validate the signal.

Within this paper we will show several interesting examples of selected behavioural signals from public procurement environment in Slovakia, which is highly transparent and has also applied automatic notification of suppliers on published tenders registered into specific CPV code.

Regarding several recommendations from existing studies, past behavioural patterns are good starting point to analyse if something interesting for deeper understanding and analysis is behind. One of the approaches how to analyse past behaviour is to visualise different contractual dimensions. Following visualisation shows several smoothed trends on big data, where 160 thousands contracts are analysed and simple trends generalized from different dimensions, e.g. contractors' savings based on the number of their contracts. This simple example shows first interesting patterns suitable for deeper analysis.

To explain this first preliminary analysis results, we will explain dimensions, we are considering as important in the field of transactional data.

As we have already mentioned, several studies recommend analysing historical behaviour that could indicate collusive patterns. In this context, however, it is important to understand that many prerequisites for a collusive signal may lie in corrupt or other unfair and ineffective efforts, where the involvement of the buyer is necessary.

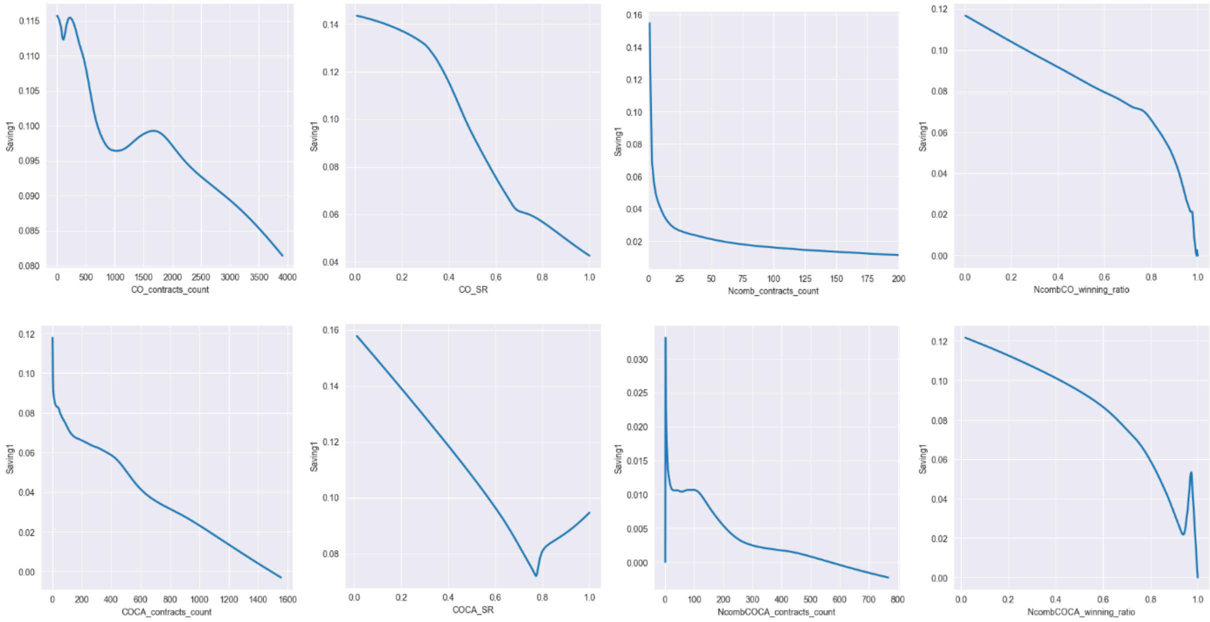


Figure 1. Aggregated and smoothed trends of savings per several dimension Source: (author)

Of course, not every signal immediately points to corrupt behaviour. It could be due to negligence, insufficient experience, errors, etc. However, statistics can help increase the likelihood that a recurring signal could be considered collusive or even corrupt. Therefore, it is appropriate to analyse the following dimensions primarily:

CO – the contractor dimension, whether a supplier or bidder, where it is important to sequentially analyse all contracts or tenders regardless of the sector or the contracting authority to whom the supplier has delivered. From a statistical viewpoint, these are not standard time series, whose methods are well-developed in econometrics. Instead, these series of contracts have a tick-like nature, where we focus more on the order of contracts rather than time. The aspect of time greatly increases the complexity of the methodology due to the immense inconsistency in the number of tenders. Announcement times, duration, and completion are unique characteristics of almost every tender. Their frequency of repetition also varies greatly among contracting authorities, where the need to issue a tender may be monthly, but for some, a similar tender may only be issued every few years. Such inconsistency complicates the understanding of the logic behind causal aspects. Therefore, in our research, we focus on the character of 'time' or rather the sequence of contracts/tenders of a given supplier. In this dimension, we recommend analysing performance or risk parameters of CO, for example, success rate (very sensitive indicator on the number of tenders due to frequent emergence of suppliers with very small amount of tenders attendance), procurement volumes, number of tenders, average values of suppliers in the tender competing with the given CO, their deviations, achieved savings, performance parameters of competitive negotiation such as bid counts, average bids per supplier, etc. Examples can include not only descriptive statistics but also benchmarks and causal rules.

COCA – the dimension of business relationships between a specific supplier (contractor) and the contracting authority. In the previous dimension, we considered the behaviour of the supplier across the entire market. In this case, it is important to distinguish and be able to compare the behaviour of the supplier in contracts or negotiations with each contracting authority separately. That is, only those tenders issued by one contracting authority and in which the supplier participated or won are analysed. This method expands the view of the possibility of benchmarking this supplier's behaviour among different contracting authorities, which may extend the analysis possibilities even in terms of non-standard tender settings or indications of corrupt behaviour.

CA – contracting authority. An interesting dimension is also the analysis from the perspective of the contracting authority's procurement, which may be specific or comparable to similar contracting authorities. However, from our perspective, it is appropriate to analyse mainly sector-specifics and therefore analyse tenders within the CA CPV (CPV for the public procurement area, although a similar principle is suitable for the B2B area based on defined categorization).

Ncomb – portfolio or cluster of suppliers attending the tender for the negotiation. In this case, this portfolio of suppliers is a significant dimension in terms of understanding possible collusive signals and agreements among suppliers in the market or according to other specific dimensions mentioned above. Analysing tenders in which this portfolio of suppliers occurs allows for understanding the stability of this portfolio, the effects of disrupting this portfolio by another, new supplier entering the negotiations, etc. Therefore, it is then appropriate to analyse multiple dimensions, such as:

NcombCO – tenders where the same supplier always wins but the portfolio of suppliers does not change. This dimension allows monitoring of risk signals given by CO in that Ncomb, comparing shares of CO in that Ncomb, and identifying patterns specific to cover offers, rotational schemes, etc.

NcombCA – tenders where the same portfolio continues to participate at a given contracting authority.

NcombCOA – tenders where at a given contracting authority, the same portfolio of suppliers participated, and the same CO always won.

Based on such dimensions, it is interesting to observe whether certain similar dimensions evoke certain trends of signals, as is the case with the graphs in Fig. 1, where evidently some signals over time, in the sequence of tenders for a given dimension, raise questions whether something unethical is occurring with the increasing number of contracts. Even though this is an aggregated, smoothed data using the Lowess method, it highlights at least some patterns of dimensions that deserve deeper attention and the application of some predictive or causal approaches.

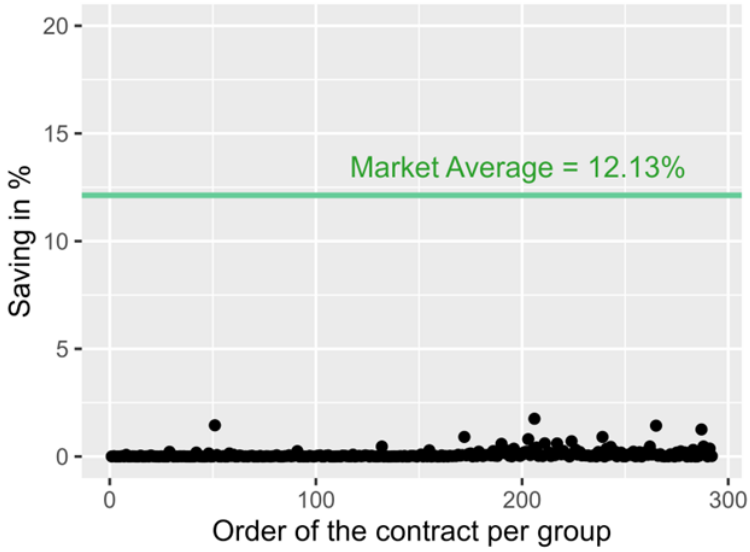


Figure 2. Comparison of savings in tenders with selected stable group of suppliers Source: (author)

An example is also shown in Fig. 2, where a consortium of suppliers was discovered, which despite a high occurrence in tenders (more than 200) shows unusually low savings achieved from negotiations, while in comparable contracts, the market average was around 12%. It is statistically improbable that in the case of adhering to proper market principles, this portfolio of suppliers (in this case, three) could maintain stability without disruption by competing suppliers. These instances merit increased attention, even in a transparent environment of digital platforms such as the EKS platform in Slovakia, where such behaviour is possible.

Conclusion

This paper underlines the importance of utilizing big data analytics to scrutinize various dimensions of procurement data and identify potential collusive behaviours that may negatively impact procurement performance. By focusing on multidimensional analysis, we propose an evolutionary approach to detect and understand the dynamics of procurement interactions. This method does not necessarily aim to legally prove collusive behaviour, instead, it focuses on identifying hypothetical collusive signals that could lead to negative outcomes. Such an approach allows for the early detection of patterns that may not be immediately evident but could suggest collusive tendencies affecting market competition.

Our findings recommend that researchers and practitioners employ big data analytics as a preventive tool to regulate and mitigate the emergence of potentially collusive signals. By analysing these signals over time, we can gain insights into the evolutionary aspects of procurement behaviours and implement timely preventive measures. This proactive approach is crucial for maintaining fair economic competition and enhancing the transparency and integrity of procurement processes.

We encourage researchers to delve into these dimensions and adopt this analytical approach to better understand the dynamics and develop effective preventive strategies. This focus on preventive measures and the identification of risk factors can help create a more competitive and economically sound procurement environment, ultimately leading to more robust and equitable market practices.

Acknowledgement

The paper is written with support of APVV within the project No APVV-20-06089 "*The research of behavioral patterns on big data within public and commercial procurement with negative impact on efficiency of procurement processes*."

References

Ávila, P., Mota, A., Pires, A R., Bastos, J., Putnik, G D., & Teixeira, J. (2012). Supplier's Selection Model based on an Empirical Study. Elsevier BV. https://doi.org/10.1016/j.protcy.2012.09.069

Bardhi, E. (2014). Integrity in Albanian Public Procurement. Principles and the Fight Against Corruption. Richtmann Publishing. https://doi.org/10.5901/ajis.2014.v3n4p185

Czibik, Á., Fazekas, M., Tóth, B. & János, T. (2014). Toolkit for detecting collusive bidding in public procurement. With examples from Hungary. January 2014. Report number: CRC-WP/2014:02Affiliation: Corruption Research Center Budapest

Fakir, J. (2015). Procurement systems should lower, not increase, organisational risk. https://www.itweb.co.za/content/WPmxVEMKBYd7QY85

Kassim, E S., & Hussin, H. (2010). Public E-procurement: A Research Synthesis. https://doi.org/10.1109/ic4e.2010.93

Kiener-Manu, K. (2018). Anti-Corruption Module 4 Key Issues: Corruption in Public Procurement. https://www.unodc.org/e4j/en/anti-corruption/module-4/key-issues/corruption-in-public-procurement.html

Nam, I. (2016). Collusion in a Telecom Market in Which the Entrant Raises the Price in Return for a Discount in Interconnection Charges by the Incumbent. RELX Group (Netherlands). https://doi.org/10.2139/ssrn.2836791

Morapitiya, P.R. (2018). The Productivity Success Factors affecting to Competitiveness of the Readymade Garment industry in Sri Lanka. https://doi.org/10.14299/ijoart.07.07.012

OECD (2012). Fighting bid rigging in public procurement. https://www.oecd.org/competition/cartels/fightingbidrigginginpublicprocurement.htm

Qi, L M., Ramly, Z M., Ariffin, H L T., Mohd, N I., & Dzahir, M A M. (2020). Construction Tender Pricing Strategies in Traditional Procurement System. IOP Publishing. https://doi.org/10.1088/1757-899x/884/1/012031

Rahmawati, D., Yaqin, M A., & Sarno, R. (2016). Fraud detection on event logs of goods and services procurement business process using Heuristics Miner algorithm. https://doi.org/10.1109/icts.2016.7910307

Relucio, F S., & Cruz, J S D. (2020). A Procurement Data Standard Tool for State Universities and Colleges. IOP Publishing. https://doi.org/10.1088/1757-899x/803/1/012005

Urminský, J., Zajarošová, M., Vávra, A. & Karady, O. (2023). Benefits of e-Auction: Real Users Evaluation. In: IDIMT 2023 – New Challenges for ICT and Management, 31st Interdisciplinary Information Management Talks (Vol. 52, pp. 321-328), Hradec Králové, Czech Republic. doi: 10.35011/IDIMT-2023-321

Zage, D J., Glass, K., & Colbaugh, R. (2013). Improving supply chain security using big data. https://doi.org/10.1109/isi.2013.6578830

SAVINGS ANALYSIS OF SINGLE-BIDDER VS. MULTI-BIDDER PUBLIC PROCUREMENTS IN SLOVAKIA

Michal Tkáč, Michal Tkáč, Michal Husnaj

University of Economics in Bratislava
Faculty of Business Economics with seat in Košice
michal.tkac1@euba.sk, michal.tkac@euba.sk, michal.husnaj@euba.sk

DOI: 10.35011/IDIMT-2024-337

Keywords

Public procurement; CPV; single bidder; savings; Slovakia;

Abstract

This study analyzes public procurement offers from the Slovak Electronic Contractual System (EKS) in order to assess the impact of bidder competition on savings in Slovak public procurement environment. Focusing on contrasts between single-bidder and multi-bidder competitions, the research employs statistical methods to evaluate the implications of competitive dynamics on fiscal efficiency. Results indicate that procurements with multiple bidders yield significantly higher savings, demonstrating the benefits of competitive bidding. The study further identifies procurement categories prone to single-bidder dominance, suggesting sectors with potential for increased regulatory oversight to enhance competition. These findings underscore the importance of competitive procurement practices in enhancing fiscal efficiency and guiding effective policy interventions.

1. Introduction

Public procurement plays a crucial role in government spending and budget management. It is a key tool for how governments, public institutions, and other publicly funded organizations obtain goods and services from external suppliers. Efficient public procurement practices can result in significant cost savings for government agencies, as well as improved transparency and accountability in the use of public funds(Public procurement - OECD, 2006). However, public procurement processes can also be susceptible to anticompetitive practices, which hinder competition and lead to inefficiencies and inflated costs (Jones et al., 2018) These anticompetitive practices, such as bid rigging, collusion among suppliers, and a lack of genuine competition in the bidding processes ultimately limit the number of bidders and reduce potential cost savings (Bhagat, 2017; Heimler, 2012). This point is reinforced by empirical research, which suggests that savings in public procurement tend to escalate proportionally with the size of the procured function and the importance of price in the contract award criteria, while being adversely affected by limited bidder participation (Chapela et al., 2017). The factor "Number of Bidders" in public procurement is a well-documented determinant of cost savings. The relationship between bidder competition and procurement outcomes has been researched in various contexts and sectors. An increased number of bidders contributes to a competitive atmosphere where each participant is incentivized to offer their best possible price. This competitive environment can prevent the prevalence of overpriced bids and limit the possibilities of collusion among bidders (Džupka et al., 2020; Onur & Taş, 2018; Chapela et al., 2017). For instance, the study titled "Optimal bidder participation in public procurement auctions" by Onur and Taş examines the optimal number of bidders needed to foster a competitive public procurement market. The research suggests that procurement costs decrease until there is participation from six to eight bidders. Having established this optimal range, policymakers have a benchmark to strive for in terms of encouraging sufficient competition to realize cost savings (Onur & Taş, 2018).

In the case of Slovakia, one major issue in public procurement is the high number of contracts awarded to only one bidder. This lack of real competition could lead to increased costs for the government. The extent and consequences of these single-bidder public procurement processes are not well-documented in academic research. In this study, we seek to compare cost savings achieved in single-bidder procurements with those in multi-bidder procurements in Slovakia. Through this analysis, we aim to understand how effective competitive bidding is at generating cost savings and ensuring value for money in public spending.

2. Methodology and description of sample

This study aims to analyse savings from public procurement competitions in Slovakia, focusing on the differences between competitions with a single bidder and those with multiple bidders. We utilized a dataset comprising over 160,000 real public procurement offers sourced from the Slovak Electronic Contractual System (EKS), a digital platform that supports electronic contracting processes.The data for this research were extracted from the EKS (www.eks.sk), which provides a comprehensive, transparent platform designed to facilitate electronic contracting. The EKS platform enables the creation, negotiation, and signing of contracts digitally, thereby enhancing efficiency and reducing paperwork. Importantly, EKS supports API approaches for big data analysis, which was crucial for handling the large volume of data in this study. The period of study spans from January 10, 2014, to December 1, 2023. This timeframe allows for a robust longitudinal analysis of trends and patterns in public procurement savings.

The Main output variable used in the study is Savings. It is defined as the difference between the estimated value of a procurement contract and the winning bid, expressed as a percentage of the estimated value. The main input variable used in this research is One Applicant. This binary variable categorizes the procurement competitions into two groups: those with only one bidder (Yes) and those with more than one bidder (No). The other variables used in the paper are:

- CPVs (Common Procurement Vocabulary): We utilized the CPV, a standardized classification system across the EU, to categorize and analyze procurement offers. This study examined four levels of CPV granularity: The first two digits identify the divisions (XX000000-Y); The first three digits identify the groups (XXX00000-Y); The first four digits identify the classes (XXXX0000-Y); The first five digits identify the categories (XXXXX000-Y).
- CO_CPV (Contractor_CPV): A unique identifier for each combination of Contractor/Applicant/Bidder and CPV, allowing for comparison across similar procurement settings.
- CO_CA_CPV (Contractor_Contracting Authority_CPV): A unique identifier for each combination of Contractor/Applicant/Bidder, Contracting Authority, and CPV, enabling detailed comparisons where the same entities are involved across different procurements.

Due to the non-normal distribution of the variables, non-parametric statistical tests were employed. The primary analytical methods included: Mann-Whitney U Test, which is used to assess the statistical significance of the differences in savings between one-bidder and multi-bidder competitions. Box-Plot Analysis, which provides a visual representation of the distribution of savings across different categories and variables. Pareto Analysis, which helps identify the CPVs in which most one bidder competitions occur. Spearman Correlation Analysis which assessed the strength and direction of association between the variables.

3. Research results

The first part of our analysis is focused on comparing the distribution of savings across the entire research sample, which represents the Slovak public procurement environment. The primary motivation for this research was to understand how the distribution of savings in procurement competitions with only one bidder differs from those with more than one bidder. This comparative approach provides insights into the impact of competitive bidding on the efficiency and effectiveness of public spending. By analyzing how the presence of multiple bidders influences the financial outcomes of procurement processes, we aim to provide empirical evidence that could guide policy decisions and strategies in public procurement. The distribution of savings based on occurrence of only one bidder in competition is presented in Figure 1.

Our analysis differentiates the savings distribution between procurement competitions with a single bidder and those with multiple bidders. For competitions with multiple bidders (One Applicant= No), the data reveals a mean savings of 19%, a standard deviation also of 19%, with the lower quartile at 3%, the median at 14%, and the upper quartile at 29%. In contrast, competitions with only one bidder (One Applicant= Yes) present a starkly different profile, with a mean savings of merely 1%, a standard deviation of 6%, and both the median and the 25th percentile standing at 0%, highlighting minimal savings in most cases.

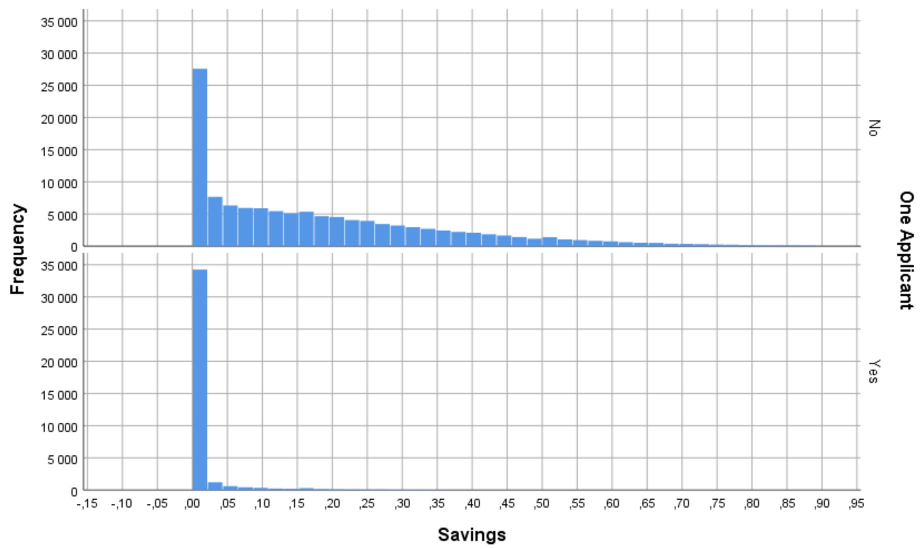


Figure 1. Distribution of public procurement savings based on presence of only one applicant

Source: (author)

To support this claim, the null hypothesis proposing that the savings distribution is identical regardless of the number of applicants (one vs. multiple) was tested. We use an independent-samples Mann-Whitney U Test. The Mann-Whitney U test provided evidence to reject the null hypothesis, with a p-value (p-value < 0.001) far below the conventional alpha level of 0.05. This significant difference in the savings distributions between single and multiple bidder competitions confirms that the presence of additional bidders lead to increasement of savings.

The second part of our analysis concentrates on the use of the Common Procurement Vocabulary (CPV) as a standard for categorizing the types of goods and services procured. Specifically, we evaluated the percentage representation of Public procurement competitions with only one bidder across different CPV categories. The main motivation behind this aspect of the study was to identify those CPV types where one bidder competitions represent at least half of all procurement activities. This analysis provides a clearer picture of which sectors may lack competitive diversity, potentially indicating monopolistic tendencies or specialized market conditions where few suppliers exist. The results of this segment of our analysis are graphically represented in Figure 2.

The analysis, as depicted in Figure 2, examines CPV codes at both the first divisional level and the second group level. At the divisional level, we identified six CPV categories where more than 50% of the procurement competitions involve only one bidder. They are: 73: Research and development services and related consultancy services, 66: Financial and insurance services, 51: Installation services (except software), 92: Recreational, cultural and sporting services, 24: Chemical products, 50: Repair and maintenance services. These categories suggest sectors where the market might be highly specialized, or where barriers to entry prevent a larger number of bidders from participating.

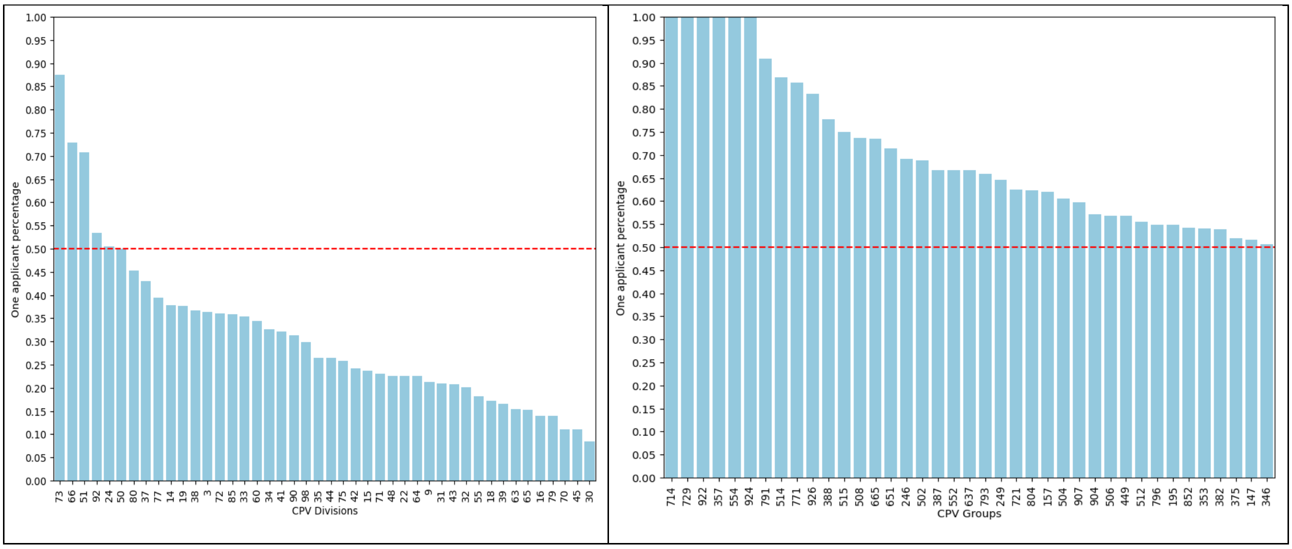


Figure 2. Overview of CPV Divisions (left) and CPV Groups (right) with majority of one-bidders competitions

Source: (author)

A more granular analysis at the second level CPV identified 47 CPV groups where competitions predominantly have only one bidder. The specific CPV codes for these groups are showed in the right graph of Figure 2, indicating a widespread presence of one bidder scenarios across various procurement categories. Additionally, we identified six CPV groups where 100% of the procurement competitions involved only one bidder. They were: 357: Military electronic systems, 554: Beverage-serving services, 714: Urban planning and landscape architectural services, 729: Computer back-up and catalogue conversion services, 922: Radio and television services, 924: News-agency services. These findings point to areas within the public procurement landscape where there is absolutely no competition, highlighting sectors potentially ripe for policy intervention to encourage more competitive bidding processes.

The third part of our study focuses on comparing the average public procurement savings per CPV with the percentage of single-bidder procurements within each CPV category. This analysis aims to test the hypothesis that a higher incidence of single-bidder procurements leads to a less competitive environment, which may be associated with a decrease in average savings per CPV. To comprehensively examine this relationship, we conducted correlation analyses across four different levels of CPV fragmentation: from Divisions (Level 1) through Groups (Level 2) and Classes (Level 3) to Categories (Level 4). These analyses help to discern how savings dynamics change as we delve into more specific categories of procurement. The graphical representation of these analyses are showed in Figure 3.

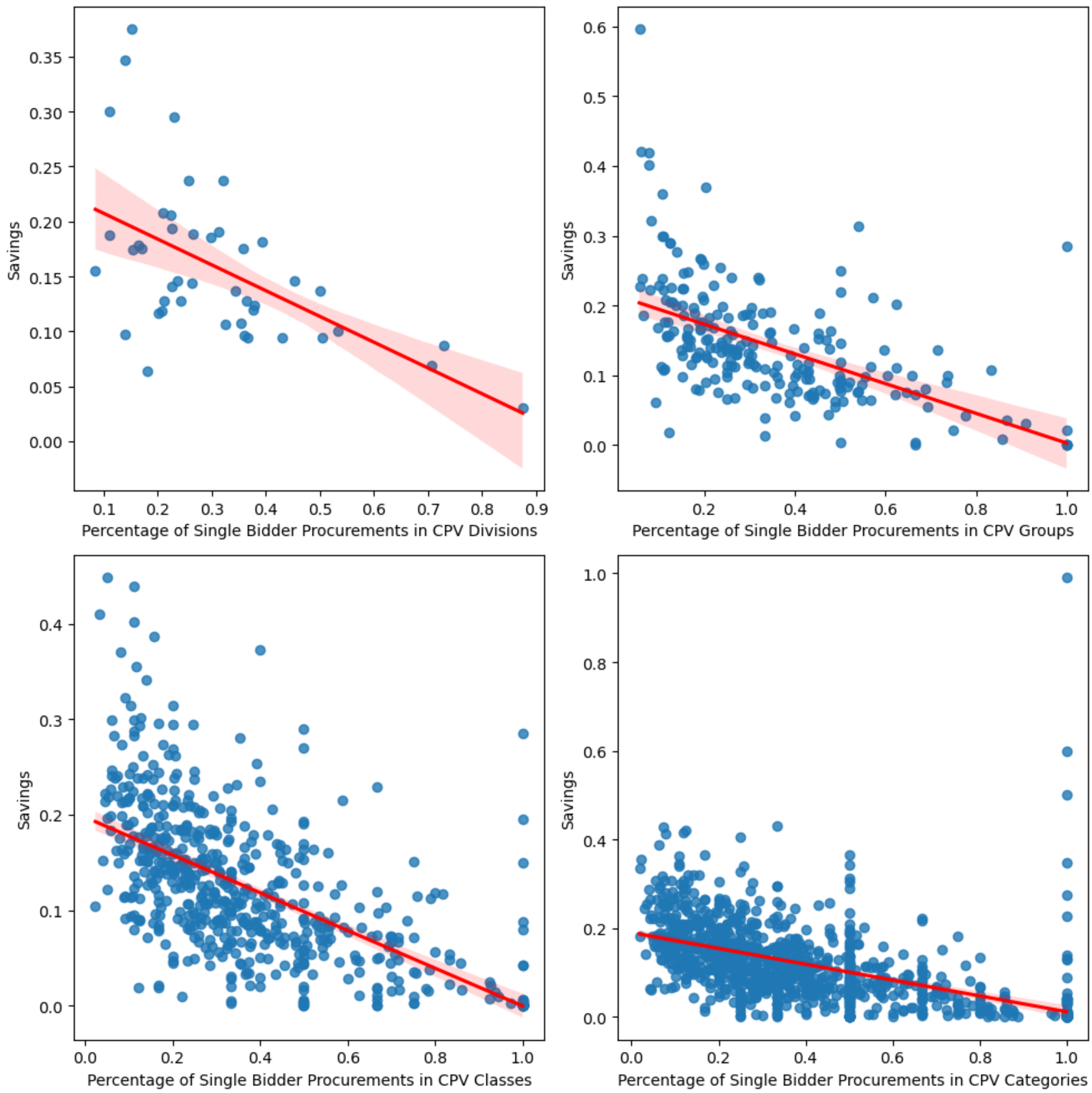


Figure 3. Overview of scatter plots comparing Savings with percentage of Single Bidders procurements in various level of CPV fragmentation

Source: (author)

The correlation analysis conducted to explore the relationship between the percentage of single-bidder procurements and average savings in public procurement at various levels of CPV classification revealed statistically significant negative correlations at all levels. For Level 1 (Divisions) a correlation coefficient of -0.525 indicates a strong negative relationship between the increase in single-bidder procurements and a decrease in average savings within CPV divisions. The statistical significance of this correlation is supported by a p-value of 0.00025. For Level 2 (Groups), the correlation coefficient deepens to -0.609 at the group level, suggesting an even stronger negative association compared to the divisional level. This result, with a p-value of less than 0.00001, underscores a robust inverse relationship between single-bidder dominance and savings at this more detailed classification level. The strength of the negative correlation increases further to -0.674, indicating that as we narrow down to specific classes within the CPV (Level 3), the impact of single-bidder procurements on reducing savings becomes more pronounced. The p-value remains below 0.00001, confirming the reliability of these findings. At the most detailed level of CPV classification (Level 4), the correlation coefficient is -0.678, the strongest among all levels tested, suggesting that the specific categories of procurement where competition is limited see the most substantial decrease in savings. Like the previous levels, the statistical significance of this correlation is extremely high, with a p-value of less than 0.00001. These results collectively indicate that there is a consistent and strengthening negative relationship between the percentage of single-bidder procurements and the average savings across increasingly specific levels of the CPV classification. The negative correlation becomes more pronounced as the CPV level becomes more detailed, illustrating that higher levels of procurement specificity are associated with more significant impacts of reduced competition on cost savings. This pattern highlights critical insights into the dynamics of competition within public procurement and underscores the importance of fostering competitive environments to enhance efficiency and savings across all levels of public contracting.

The final part of our research focused on examining how savings from various public procurement competitions within the same Common Procurement Vocabulary (CPV) category are affected by the presence of competition. To achieve this, we developed a unique identification system for each contractor (bidder) based on their Tax Identification Number. This system allowed us to assign a specific CO_CPV ID to each contractor for every procurement they won within a particular CPV. By grouping procurement competitions based on the winning contractors for each CPV, we could further subdivide these groups into two: the first group comprised competitions where the contractor was the sole bidder, and the second group included competitions where the contractor faced other bidders. This segmentation enabled us to compute and compare the average savings for both groups. The primary variable of interest, "savings difference," was calculated as the average savings from competitions with multiple bidders minus the average savings from competitions with a single bidder. The motivation behind this analysis was to explore how the distribution of savings differs when a bidder competes alone versus when competing against others within the same CPV. See first boxplot of Figure 4. A total 4,668 unique CO_CPV ID were analysed, focusing on scenarios where a contractor was alone and also faces competition within the same CPV. On average, savings increase by 15.40% when the contractor is not the sole bidder in the competition for particular CPV. This indicates a substantial improvement in savings when there is more than one bidder. On the other hand the variable savings difference has a standard deviation of 13.53%, suggesting considerable variability in how much savings increase when additional bidders are present.

Building on our earlier analysis, we introduced an additional layer of complexity by creating a unique identifier, the COCA-CPV ID. This identifier encapsulates every combination of Contracting Authority (CO) and winning contractor (CA) for all competitions within the same CPV. This step allowed us to precisely track how specific pairings between contractors and contracting authorities impact savings under different competitive conditions. For each COCA-CPV ID, we segmented the procurement data into two groups: the first group consisted of competitions where the contractor was the sole bidder, and the second group where the contractor faced competition. We then calculated the average savings for both groups and determined the "savings difference" by subtracting the average savings of the first group from the average savings of the second group. The motivation behind introducing the COCA-CPV ID and conducting this additional layer of analysis was to examine how the dynamics between specific contracting authorities and contractors influence savings outcomes within the same CPV. This approach aimed to assess whether consistent relationships between particular contractors and contracting authorities could either mitigate or exacerbate the effects of competition on procurement savings. The distribution of Savings difference for competitions within various COCA-CPV ID are presented in second box-plot of figure 4. Our dataset includes 7,450 unique Contracting authority – Winning contractor pairs which have both experience with single bidder procurement and multi bidder procurement. The average increase in savings when the contractor is not alone is 14.12%. This figure is slightly lower than in the CO_CPV analysis, suggesting that the dynamics between specific contracting authorities and contractors might not be as conducive to significant savings increases.

These findings highlight the significant impact that competition has on increasing savings in public procurement, both in generalized contractor scenarios and in specific contractor-authority relationships. The CO_CPV results generally show a higher potential for savings increase compared to COCA_CPV scenarios, suggesting that competitive dynamics might be less effective in contexts where contractors repeatedly engage with the same contracting authorities.

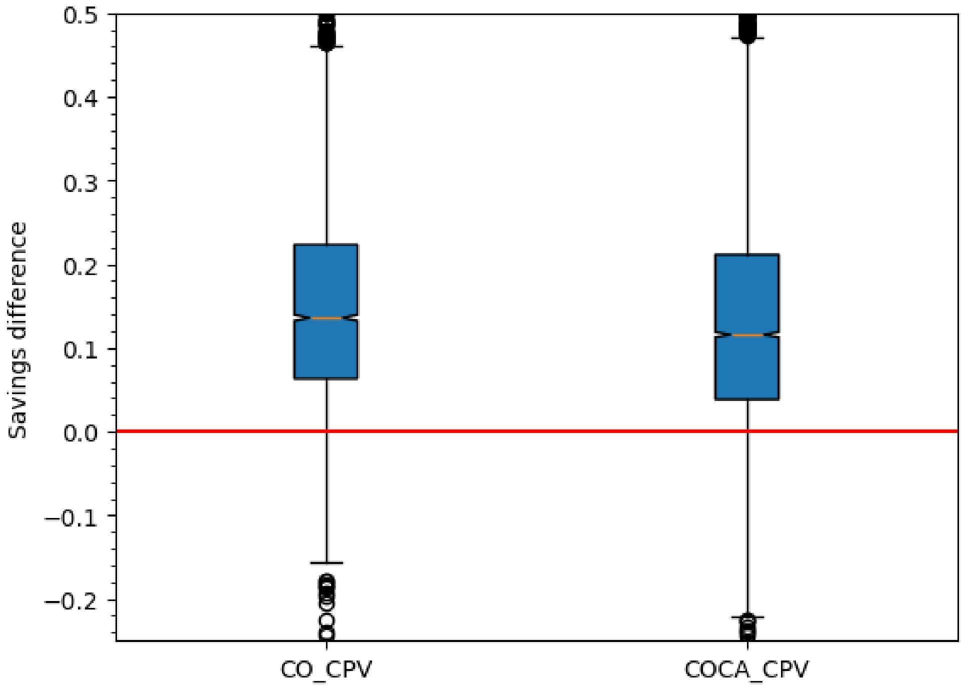


Figure 4. Boxplots of Savings Differences: Comparing Single and Competitive Bids Across CO_CPV and COCA_CPV Categories

Source: (author)

5. Conclusion

This study delves into the dynamics of savings in public procurement competitions in Slovakia, with a focus on the contrast between competitions with single bidders and those with multiple bidders. By analyzing over 160,000 real public procurement offers, the research explored the relationship between savings and competition, employing various statistical methods such as Mann-Whitney U Test, Box-Plot Analysis, and Spearman Correlation Analysis. The results reveal a stark contrast in savings 313 between competitions with single bidders and those with multiple bidders. Competitions with multiple bidders demonstrated significantly higher savings, emphasizing the positive impact of competition on fiscal efficiency in public procurement. Moreover, the analysis identified specific CPV categories where one-bidder competitions were prevalent, suggesting areas of potential market specialization or monopolistic tendencies. The findings underscore the critical role of competitive bidding in driving savings in public procurement processes. By fostering competition, governments can maximize efficiency and ensure optimal utilization of public funds. The identification of CPV categories dominated by one-bidder competitions sheds light on sectors where market conditions may hinder competitive diversity, warranting policy interventions to encourage broader participation. Moreover, the correlation analysis across different levels of CPV fragmentation reveals a consistent negative relationship between the percentage of single-bidder procurements and average savings. This implies that increased competition leads to higher savings across all levels of procurement specificity, emphasizing the importance of fostering competitive environments in public contracting. In conclusion, this research provides robust empirical evidence supporting the efficacy of competitive bidding in enhancing fiscal efficiency in public procurement. The findings underscore the need for policies and strategies aimed at promoting competition to maximize savings and ensure optimal resource allocation in public spending. By leveraging insights from this study, policymakers can develop targeted interventions to foster competitive environments, thereby driving efficiency and effectiveness in public procurement processes.

Acknowledgement

The paper is written with the support of APVV within project No APVV-20-06089 “The research of behavioral patterns on big data within public and commercial procurement with negative impact on efficiency of procurement processes”

References

Public procurement - OECD. (2006, January 1). https://www.oecd.org/gov/public-procurement/

Jones, A., Anderson, R D., & Kovacic, W E. (2018, January 1). Preventing Corruption, Supplier Collusion and the Corrosion of Civic Trust: A Procompetitive Program to Improve the Effectiveness and Legitimacy of Public Procurement. https://doi.org/10.2139/ssrn.3289170

Bhagat, G. (2017, June 1). Public Procurement: A Competition Perspective. https://doi.org/10.1177/0019556117699736

Heimler, A. (2012, October 26). CARTELS IN PUBLIC PROCUREMENT. https://doi.org/10.1093/joclec/nhs028

Džupka, P., Kubák, M., & Nemec, P. (2020, November 6). Sustainable Public Procurement in Central European Countries. Can It Also Bring Savings? Multidisciplinary Digital Publishing Institute, 12(21), 9241-9241. https://doi.org/10.3390/su12219241

Onur, İ., & Taş, B K O. (2018, September 27). Optimal bidder participation in public procurement auctions. Springer Science+Business Media, 26(3), 595-617. https://doi.org/10.1007/s10797-018-9515-2

Chapela, J G., Labeaga, J M., & Medrano, L Á. (2017, December 19). Further econometric evidence on the extent and sources of cost savings in competitively tendered contracts. Springer Science+Business Media, 56(2), 679-701. https://doi.org/10.1007/s00181-017-1365-8

EVALUATING THE BENEFITS OF E-AUCTIONS IN THE PUBLIC SECTOR AND THE IMPACT OF BREACHES OF COMMITMENTS

Markéta Zajarošová, Jaroslav Urminský, Denisa Harkabusová

VŠB-Technical University of Ostrava
Faculty of Economics
marketa.zajarosova@vsb.cz, jaroslav.urminsky@vsb.cz,
denisa.harkabusova.st@vsb.cz

DOI: 10.35011/IDIMT-2024-345

Keywords

E-auctions; benefits; survey; breach of obligations; public sector

Abstract

The growth in the use of e-auctions in the public sector is clear. The question is which benefits of e-auctions are more important for institutions, cities and municipalities and whether the evaluation of benefits is affected by possible breaches of suppliers' obligations. Data were collected by online survey, a total of 5679 e-auction users in the Czechia and Slovakia were contacted. The descriptive description of the data is complemented by a non-parametric test using statistical software. Statistical differences were confirmed between breach of commitment and ratings of transparency and simple feedback.

1. Introduction

The adoption of e-auctions in the public sector marks a significant shift towards enhancing transparency (Dráb et al., 2018), efficiency (Standaert et al., 2015), and competitiveness in the procurement process. E-auctions, or electronic auctions, leverage digital platforms to facilitate bidding on government contracts and services, allowing for real-time competition and streamlined procurement workflows. This modern approach addresses long-standing challenges in public procurement, such as opacity, inefficiency, and potential for corruption (Glas and Essig, 2018), by providing a more open and accessible bidding environment. By enabling multiple suppliers to participate simultaneously, e-auctions promote fair competition and can lead to cost savings and better value for taxpayer money. Furthermore, the digital nature of e-auctions supports comprehensive record-keeping and audit trails, which are crucial for ensuring accountability and regulatory compliance.

The objective of the paper is to analyze the main benefits of e-auctions, as perceived by their users themselves. We focus primarily on public sector that is actual user or sometimes used the e-auction in the past. The research is based on data obtained from the survey conducted between Czech and Slovak institutions, cities and municipalities. Our objective was not only to determine which benefits of e-auctions are most beneficial to the public sector, but also to determine whether there is a relationship between the evaluation of these benefits and the breach of obligations of winning suppliers.

2. Literature review

All over the world, public entities are part of the running of the economic system. They are active in various markets in an effort to secure their needs and fulfil their essential mission. Governments and public (state-owned) organizations realize their purchases through public procurement. The estimate is that about 15% of global GDP flows through public procurement systems (Fazekas, Blum, 2021). Public entities have to respect the public interest. The pressure of the society generates the stimulus for improving the efficiency of purchasing behaviour as a strict economic condition and requires of public entities for fiscal responsibility. The successive digitization of processes and activities offers new ways and tools for purchasing in this sense. The electronic reverse auctions (eRA) represent one of them. This tool can be considered one of the most important market mechanisms for negotiation between organizations worldwide.

One of the main benefits lies in the possibility of achieving significant savings in various types of products and services (Wang, Zhang, Liu, 2020). Savings are achieved based on the mutual interaction of competitors within the auction process, which may push the product price down. But a specific kind of motivation may be presented that differs from the standard private subjects' purchasing. Public entities want to achieve economic, environmental, or social goals or their mutual combination. Strictly economic goals do not necessarily have to play a primary role.

There are several types of auctions, differing in the mechanisms behind them (Liu, 2021). One represents reverse auctions, and their electronic form electronic reverse auction (eRA). Electronic Reverse Auctions (eRA), also known as online reverse auctions, are a procurement method and e-commerce tool used in the purchasing process. In an eRA, a buying organization, often referred to as the buyer or the procuring entity, invites potential suppliers to participate in a competitive bidding process conducted online through a specialized eAuction platform. The primary goal of an eRA is to obtain goods or services at the best possible price while promoting transparency and efficiency in the procurement process. There is still a limited understanding of the determinants of auction savings that exist in this process, especially factors that can change information asymmetry during auctions (Drab et al., 2022).

In public management and government procurement, electronic reverse auctions (eRAs) are often used to improve transparency, cost-effectiveness, and efficiency (Pearcy and Giunipero, 2008) in the acquisition of goods and services. Public management eRAs are typically characterized by their focus on transparency, compliance with regulations, and the consideration of factors, such as quality, environmental sustainability, and socio-economic objectives (Bosio et al., 2022).

3. Research design

This part included methodical notes of the research. Basic conditions of the research in the sense of data collection are described. Next, we introduce brief characteristics of the applied methods.

3.1. Data collection

Data collection was carried out through a survey. The questionnaire itself consisted of 14 questions. Closed or semi-closed types of questions were represented. A battery of questions containing selected factors of the benefits of electronic auctions was also used, and evaluation was done through a five-point Likert scale (see Boone and Boone 2012; Joshi et al. 2015). The Likert contained 5 degrees (from 1-very beneficial to 5-no benefit). The lowest/highest value expressed a strict agreement/disagreement with the question. The neutral position was represented by degree 3. Only selected issues will be presented due to the predefined maximum range of the paper.

The data collection itself took place from 16.3. until 21.3.2021. They were collected for the purpose of the final thesis (see Ondřejková, 2021). The survey was conducted using an electronic questionnaire, which was distributed by the company PROEBIZ s.r.o. Its core activity lies in the provision of services in the field of e-auctions. The company used a database of institutions, cities and municipalities in the Czechia and Slovakia who were subsequently contacted.

The sample size were 5679 respondents. Of these, 3722 respondents were contacted in Slovakia and 1957 in Czechia. Only 53 respondents out of all those approached responded to the request. The total return is therefore slightly below 1 %. The very low return may have been due to the time of data collection. Data collection was carried out during the pandemic period. The impacts of the pandemic were widespread and fundamental. It often meant an extreme workload for the employees. In sum, employees may not have had the time disposition, or desire to answer the questions. The small sample represents one of the main limitations of the presented article.

The original sample of 53 respondents had to be further reduced due to non-compliance with the pre-condition of non-response. After reduction, the final sample of respondents are *N=24* and *34*.

3.2 Applied methods

The applied methods are based on the nature of the research, the defined objective, and the character of the dataset. Due to the limitations of the paper, selected results are presented. Standard descriptive statistics were used, followed by verification of the data distribution. In the last step, we tested the relationships between the selected variables using non-parametric tests. All relevant hypotheses were tested at the α=0.05. Finally, three basic steps were realized: a) Descriptive statistics, b) Normality test, c) Jonckheere-Terpstra tests.

4. Results

The results are divided into two parts. Part one reflected basic descriptive characteristics. Part two presented results in sense of the relationship between a) the winners of the e-auctions have not followed the terms of the tender and evaluation benefits of e-auctions, b) the winners of the e-auctions refused to sign the contract and evaluation benefits of e-auctions.

4.1 Descriptive statistics

The selected descriptive statistics were used for the summarization of our data set. Based on the nature of the data, we show all three basic categories of measures: measures of central tendency, measures of variability, and frequency distribution, see Figures 1 and 2.

The winners of the e-auctions have not followed the terms of the tender (% of cases)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 % - 5 %	18	34,0	75,0	75,0
	6 % - 10 %	1	1,9	4,2	79,2
	11 % - 20 %	3	5,7	12,5	91,7
	21 % - 50 %	2	3,8	8,3	100,0
	Total	24	45,3	100,0	
Missing	System	29	54,7		
Total		53	100,0		

The winners of the e-auctions refused to sign the contract (% of cases)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 % - 5 %	28	52,8	82,4	82,4
	6 % - 10 %	2	3,8	5,9	88,2
	11 % - 20 %	3	5,7	8,8	97,1
	21 % - 50 %	1	1,9	2,9	100,0
	Total	34	64,2	100,0	
Missing	System	19	35,8		
Total		53	100,0		

Figure 1. Frequency distribution Source: (own elaboration, IBM SPSS 2024)

The tables above show the basic structure of the sample according to their experience with e-auctions. This data indicates that the majority of auction winners adhere to the terms of the tender. The data shows the frequency of cases where winners of e-auctions refused to sign the contract, categorized by the percentage of such refusals. The majority of winners (28 out of 34, or 82.4%) fall within the 0% to 5% refusal range. A smaller portion of cases (6 out of 34, or 17.6%) have higher refusal rates. This indicates that contract refusal is mostly infrequent among e-auction winners.

The overall evaluation of the selected e-auction advantages for institutions, cities and municipalities in the Czechia and Slovakia are reflected in Figure 2. The ranking according to the average values shows that the respondents most value speeding up communication within the tender process, increasing transparency and archiving in one place.

Statistics

	increased transparency	speeding up communication within the tender procedure	reducing the costs associated with the tendering procedure	financial savings	obtaining an objective offer on the market	data archiving in one place	it is not only necessary to compete on price, but also on other conditions	simple feedback
N Valid	34	34	34	34	34	34	34	34
Mean	2,03	1,97	2,62	2,15	2,35	2,06	2,79	2,12
Median	1,50	2,00	3,00	2,00	2,00	2,00	3,00	2,00
Std. Deviation	1,337	1,167	1,303	1,184	,950	1,347	1,250	1,274
Variance	1,787	1,363	1,698	1,402	,902	1,815	1,562	1,622
Skewness	1,156	1,274	,337	,979	,789	1,231	,316	1,076
Std. Error of Skewness	,403	,403	,403	,403	,403	,403	,403	,403
Kurtosis	,129	1,016	-,739	,244	,710	,424	-,857	,258
Std. Error of Kurtosis	,788	,788	,788	,788	,788	,788	,788	,788
Minimum	1	1	1	1	1	1	1	1
Maximum	5	5	5	5	5	5	5	5

Figure 2. Evaluation of e-auction benefits – descriptive statistics Source: (own elaboration, IBM SPSS 2024)

4.2 Relationship between e-auction benefits evaluation and breach of obligations

First, we tested the assumption of normality data distribution. A normality test helps determine whether our data are normally distributed or not. The Shapiro-Wilk tests were used because of the character of our data. It tests the hypothesis that the distribution of the data deviates (or not) from a normal distribution. Finally, the observed p-values $p \leq 0.05$ indicated that the distribution of our sample is significantly different from a normal distribution. Based on the results the nonparametric method was applied. The Jonckheere-Terpstra test was chosen to identify the potential differences among classes (Field, 2009). The test was used to identify potential differences in the evaluation of e-auction benefits in relation to the situations in which winners of e-auctions did not follow the terms and conditions (% of cases), see Figure 3. The test was used than to identify potential differences in the evaluation of e-auction benefits in relation to the situations in which winners of e-auctions refuse to sign the contract (% of cases), see Figure 4.

Jonckheere-Terpstra Test[a]

	increased transparency	speeding up communication within the tender procedure	reducing the costs associated with the tendering procedure	financial savings	obtaining an objective offer on the market	data archiving in one place	it is not only necessary to compete on price, but also on other conditions	simple feedback
Number of Levels in Winners of e-auctions did not follow the terms and conditions	4	4	4	4	4	4	4	4
N	24	24	24	24	24	24	24	24
Observed J-T Statistic	87,500	74,000	67,500	83,500	68,000	67,000	82,000	98,500
Mean J-T Statistic	59,500	59,500	59,500	59,500	59,500	59,500	59,500	59,500
Std. Deviation of J-T Statistic	13,620	13,999	14,594	14,195	13,522	14,127	14,612	14,173
Std. J-T Statistic	2,056	1,036	,548	1,691	,629	,531	1,540	2,752
Asymp. Sig. (2-tailed)	,040	,300	,584	,091	,530	,595	,124	,006

a. Grouping Variable: Winners of e-auctions did not follow the terms and conditions (% of cases).

Figure 3. The situations in which winners of e-auctions did not follow the terms and conditions (% of cases) and evaluation of e-auction benefits Source: (own elaboration, IBM SPSS 2024)

Jonckheere-Terpstra Test[a]

	increased transparency	speeding up communication within the tender procedure	reducing the costs associated with the tendering procedure	financial savings	obtaining an objective offer on the market	data archiving in one place	it is not only necessary to compete on price, but also on other conditions	simple feedback
Number of Levels in Winners of e-auctions refuse to sign the contract	4	4	4	4	4	4	4	4
N	34	34	34	34	34	34	34	34
Observed J-T Statistic	118,500	97,500	99,500	100,000	107,500	103,500	101,000	141,000
Mean J-T Statistic	89,500	89,500	89,500	89,500	89,500	89,500	89,500	89,500
Std. Deviation of J-T Statistic	20,649	20,868	21,472	21,219	20,627	20,817	21,624	21,134
Std. J-T Statistic	1,404	,383	,466	,495	,873	,673	,532	2,437
Asymp. Sig. (2-tailed)	,160	,701	,641	,621	,383	,501	,595	,015

a. Grouping Variable: Winners of e-auctions refuse to sign the contract (% of cases).

Figure 4. The situations in which winners of e-auctions refuse to sign the contract and evaluation of e-auction benefits Source: (own elaboration, IBM SPSS 2024)

An overall view of the test results points to an important fact. The respondent's experience with e-auctions does not have a significant influence on the evaluation of the selected benefits in most cases.

The only statistically significant differences were observed in the assessment of *increased transparency* and *simple feedback*. We can say that users who have more frequent experiences of non-compliance with terms and conditions rate transparency and simple feedback less positively. For public institutions, cities and municipalities, experiences of non-compliance with tendered conditions are real and then have an impact on how users of these systems view transparency and feedback. It is important for every e-auction user that the whole procedure goes smoothly to the end. The question is whether this type of auction is the case or whether the supplier would not have followed the conditions in another type of auction.

In the case of the second investigated relationship, i.e., winners of e-auctions refuse to sign the contract, statistically significant differences appear only in the factor *simple feedback*. The failure of the supplier to sign the contract becomes a much more serious problem for e-auction users. From the descriptive results we see that this happens rarely, with most cases up to 5%. But, if this is happening, the more experience users have with it, the less positive they are about simple feedback. It is logical that the feedback cannot be simple if the selected supplier subsequently refuses to sign the contract under the conditions, he himself set in the e-auction.

For other e-auction benefits, no relationship was found between these observed variables.

5. Conclusion

Many studies can be found on the benefits of e-auctions. The benefit of this analysis is that actual users of e-auctions responded. The sample consisted exclusively of public institutions, cities and municipalities from the Czechia and Slovakia. This practical contribution is always advantageous as it provides answers from the real market. Respondents rated the benefits of e-auctions on a Likert scale of 1 to 5. Respondents most value speeding up communication within the tender process, increasing transparency and archiving in one place. The relationship between situations where e-auction winners did not comply with the terms and conditions and *increased transparency* and *simple feedback* has been demonstrated. Similarly, the relationship between situations where e-auction winners refused to sign a contract and *simple feedback* has been demonstrated. A limitation of this work is the small sample of respondents, although thousands of institutions, cities and municipalities were surveyed. We suggest repeating the survey to obtain more responses.

Acknowledgement

This paper is financed by Student Grant Competition of the Faculty of Economics, VŠB-Technical University of Ostrava; project's registration number is SP2024/092. All support is greatly acknowledged and appreciated.

References

Bosio, E., Djankov, S., Glaeser, S. & Shleifer, A. (2022). Public Procurement in Law and Practice. American Economic Review, 112(4), 1091-1117. https://doi.org/10.1257/aer.20200738

Boone, H. N., & Boone, D. A. (2012). Analyzing Likert Data. TheJournal of Extension 50(2), 1-5. Retrieved April 08, 2023 from https://joe.org/joe/2012april/tt2.php

Dráb, R., Delina, R. & Štofa, T. (2018). Electronic Auctions: Role of Visibility Settings in Transparency Analysis. Quality Innovation Prosperity, 22(2), 100-111. DOI 10.12776/QIP.V22I2.1059

Dráb, R., Štofa, T. & Delina, R. (2022). Analysis of the efficiency of electronic reverse auction settings: big data evidence. Electronic Commerce Research, 22(2), 427-450. DOI 10.1007/s10660-020-09433-0

Fazekas, M., & Blum, J. R. (2021). Improving Public Procurement Outcomes Review of Tools and the State of the Evidence Base. Policy Research Working Papers 9690. World Bank Group, Governance Global Practice. June 2021. https://doi.org/10.1596/1813-9450-9690

Field, A. (2009). Discovering Statistics Using SPSS. 3rd Edition, Sage, London, ISBN 978-1-84787-906-6

Glas. A. H. & Essig, M. (2018). Factors that Influence the Success of Small and Medium-sized Suppliers in Public Procurement: Evidence from a Centralized agency in Germany. Supply Chain Management, 23(1), 65-78. DOI 10.1108/SCM-09-2016-0334

Joshi, A., Kale, S., Chandel, S. & Pal, D. K. (2015). Likert Scale: Explored and Explained. British Journal of Applied Science & Technology, 7(4), 396-403. DOI 10.9734/BJAST/2015/14975

Liu, P. (2021). Balancing Cost Effectiveness and Incentive Properties in Conservation Auctions: Experimental Evidence from Three Multi-award Reverse Auction Mechanisms. Environmental and Resource Economics, 78, 417-451. https://doi.org/10.1007/s10640-021-00538-0

Ondřejková, S. (2021). Analýza užívání e-aukcí v českých a slovenských městech. Diplomová práce. Ostrava: VSB-TUO, Faculty of Economics, Marketing and Business Department

Pearcy, D. H., & Giunipero, L. C. (2008). Using E-procurement Applications to Achieve Integration: What Role does Firm Size Play? Supply Chain Management: An International Journal, 13(1), 26–34. https://doi.org/10.1108/13598540810850292

Standaert, W., Muylle, S. & Amelinckx, I. (2015). An Empirical Study of Electronic Reverse Auction Project Outcomes. Electronic Commerce Research and Applications, 14(2), 81-94. DOI 10.1016/j.elerap.2014.12.001

Wang, Q., Zhang, R., & Liu, J. (2020). Price/time/intellectual efficiency of procurement: Uncovering the related factors in Chinese public authorities. Journal of Purchasing and Supply Management, 26(3). https://doi.org/10.1016/j.pursup.2020.100622

ADVANCED DATA ANALYTICS ORGANIZATION - AS TIME WENT

Milos Maryska, Felix Espinoza, Petr Doucek, Marie Kovářová

Prague University of Economics and Business, Faculty of Informatics and Statistics

milos.maryska@vse.cz, felix.espinoza@vse.cz, doucek@vse.cz, xbalm21@vse.cz

DOI: 10.35011/IDIMT-2024-353

Keywords

Advanced Data Analytics; Advanced Data Analytics Organization; Advanced Data Analytics Organizational

Abstract

A temporal view of the evolution of the advanced data analytics organization presents a very interesting perspective on an area that is hardly addressed at all in the literature. The aim of the paper is to determine what attention is paid to the issue of the organization of advanced data analytics in terms of the number of published papers and what the authors' conclusions are, if publications on the topic exist. The paper establishes the methodology of the analysis, sets out the procedure for conducting the analysis of selected publication databases and identifies the keywords that will be used for the analysis in the selected publication databases in which the corresponding publications will be identified.

Five keywords were identified based on the analysis carried out. A very interesting result is the finding that the scientific community hardly pays attention to the issue of advanced data analytics in organizations, and on the contrary, it is more focused on consulting companies such as Gartner and others, which publish their analyses on the topic.

The paper identifies an area that is currently not adequately addressed and should be further analysed and scientifically investigated.

1. Introduction

Advanced data analytics has become a crucial tool in various fields such as clinical research, business, and healthcare. Organizations are increasingly recognizing the importance of advanced data analytics capabilities in driving business value and improving operational performance (Korsten et al., 2022). The use of advanced analytics, including mathematical optimization techniques, has been acknowledged as a valuable resource for enhancing manufacturing planning, control, and overall operational performance (Chae et al., 2014).

In the healthcare sector, advanced data analytics has been instrumental in clinical research and healthcare delivery. It has enabled the selection of patient groups of interest, data integration from various sources like pharmacy and radiology, and improved patient care through monitoring and analysis (*Advanced Data Analytics for Clinical Research Part I: What Are the Tools? - Nicolas Zhou,* *Erin M. Corsini, Shida Jin, Gregory R. Barbosa, Trey Kell, Michael H. Antonoff, Mara B. Antonoff, 2020*, n.d.). Additionally, the integration of advanced analytics with electronic health records has shown promise in providing operational decision support in healthcare settings.

Moreover, advanced analytics is applied not only in clinical research but also in areas like procurement fraud prevention, where big data analytics is employed to apply advanced analytic techniques to large and diverse datasets (Phillips & Lanclos Iii, 2014). Various industries benefit from the use of advanced analytics, including mathematical optimization techniques, to enhance operational performance (Chae et al., 2014).

The use of data and advanced analytics is gaining interest across various business domains, offering opportunities to uncover complex patterns and relationships that support strategic planning and operational performance (Bose, 2009; Chae et al., 2014). Companies worldwide are increasingly utilizing advanced analytics to analyze structured and unstructured data, enabling them to make informed decisions based on past circumstances, present events, and projected future actions (*Big Data Analytics Capability and Organizational Performance Measures: The Mediating Role of Business Intelligence Infrastructure - Aboobucker Ilmudeen, 2021*, n.d.).

In terms of firm agility, data analytics plays a crucial role in increasing organizational agility through fit, emphasizing the significance of aligning data analytics practices with organizational objectives. (Ghasemaghaei et al., 2017) Additionally, integrating data analytics with business processes contributes to business value by improving data quality and enhancing analytical capabilities within organizations (*How Does Business Analytics Contribute to Business Value? - Seddon - 2017 - Information Systems Journal - Wiley Online Library*, n.d.).

The evolution of advanced data analytics has led to the development of sophisticated tools and methodologies that enable organizations to extract valuable insights from large and complex datasets. From clinical research to business operations, the application of advanced analytics is transforming decision-making processes and driving performance improvements across various sectors.

The advancement of business intelligence and data analytics has significantly enhanced management systems by providing valuable insights for strategic planning and decision-making. (Mansell & Ruhode, 2019) By incorporating advanced data analysis techniques, organizations can improve their operational and marketing performance, ultimately leading to better firm performance. (Chaudhuri et al., 2021) The adoption of robust business analytics tools, integrated with artificial intelligence, has further enhanced organizational data-driven cultures, fostered innovation, and improved competitive advantage.

The goal of this paper is the literature review about development of the advanced data analytics and its organization in companies based on the searching for papers/articles based on defined key words.

2. Methodology

This review is performed to provide an overview and summary of up-to-date studies related to Advanced Data Analytics and its organization.

During Q1/2024, we analysed two key literature databases Web of Science, Scopus. We did not intentionally analysed Google Scholar. Authors were searching for papers published from 1990.

Authors used following combination of key words for the searching in selected databases:

- Advanced Data Analytics

- Advanced Data Analytics Organization
- Advanced Data Analytics Organization Company

The process of articles/papers selection can be described by the following steps:

- key words definition
- selection of databases
- document search based on defined key words
- analysis of selected documents based on the content of the abstract
- selection of the most relevant documents based on the abstract
- detailed analysis of the whole documents
- selection of the most relevant documents based on the detailed analysis

3. Results

In the first part of the analysis, we focus on the total number of articles registered in Web of Science and Scopus according to the most specific set of keywords, which is "Advanced Data Analytics Organization Company"

In the second part of the chapter, we focus on the results of the analysis of the most relevant papers. Relevance is assessed based on a detailed analysis of both the abstract and the content of the full papers. Given the size of the paper, we focus on only the 5 most relevant papers or conference papers in each database, which were selected based on the process described above.

3.1 Number of publications by keywords

An initial analysis was performed over both databases and we analyzed each of the keyword groups separately. As expected, the highest occurrence of articles was recorded for the most general keyword combination, i.e. "Advanced Data Analytics". An interesting finding is the fact that the first mentions of advanced data analytics are as early as 1990, but the number of articles between 1990 and 2010 is very low. In total, the total number of articles published between 1990 and 2010 is less than 200 articles/papers in each of the databases analyzed. These numbers include redundant articles as some of the articles or papers may be registered in each of the databases. Deduplication of articles is only performed in the last four steps of the above process.

After 2010, there was a gradual increase in the number of published articles, with the most significant increase starting in 2016, peaking in 2020 for Web of Science and 2021 for Scopus.

A very interesting development is especially in the case of the Scopus database, where between 2021 and 2022 there was a decrease in the number of published articles containing the keywords "Advanced Data Analytics", whereas in the case of the Web of Science database there was an increase in the number of published articles between 2021 and 2022. Conversely, there was a decrease in the number of articles published in the Web of Science database between 2022 and 2023, whereas there was a slight increase in the number of articles published in the Scopus database over the same period.

If we focus on more specific searches, i.e. "Advaned Data Analytics Organization" and "Advanced Data Analytics Organization Model", the resulting numbers of published articles are significantly higher and also.

Table 1. Number of Articles Available in Scopus/WOS Based on the Year

Year	Web of Science			Scopus		
	Advanced Data Analytics	Advanced Data Analytics Organization	Advanced Data Anlaytics Organization company	Advanced Data Analytics	Advanced Data Analytics Organization	Advanced Data Anlaytics Organization company
1990-2000	2	-	-	7	-	-
2001-2005	8	-	-	12	-	-
2006-2010	157	-	3	105	19	1
2011-2015	1.137	77	11	829	97	21
2016	677	67	10	316	44	7
2017	814	57	9	520	53	9
2018	1.110	95	17	718	81	13
2019	1.142	106	25	1.434	151	21
2020	1.305	115	29	2.056	190	32
2021	1.237	95	14	2.159	169	30
2022	1.272	115	15	1.044	154	19
2023	973	94	20	1.095	135	20

Source: (authors)

If we perform a similar analysis in terms of the countries of origin of the authors, or the countries under which they publish their articles - typically based on the country of the university - we find that the largest number of publications are from authors from the USA and China, followed by India and, somewhat surprisingly, England and Germany.

Table 2. Number of Articles Available in Scopus/WOS Based on the Country

	Web of Science			Scopus		
Countries	Advanced Data Analytics	Advanced Data Analytics Organization	Advanced Data Anlaytics Organization company	Advanced Data Analytics	Advanced Data Analytics Organization	Advanced Data Anlaytics Organization company
USA	3.771	287	50	2.383	252	25
China	1.188	61	12	2.262	74	
England	771	80	14	627	78	16
India	758	93	12	1.444	220	33
Germany	681	52	14	469	50	14
Italy	535	55	14	380	35	
Australia	509	59	12	375	55	11
Canada	493	49		367	38	
Spain	349	38				
France	327	31		255	32	
South Korea	314					
Japan	240					
Netherlands	210	42				
Malaysia					37	
Saudi Arabia				237		
Greece				209		

Source: (authors)

3.1 Key publications

For the detailed analysis below, we have selected only articles that contain a combination of the following keywords: Adavnced Data Analytics Organization Company. We did not analyze more general keyword combinations further.

Articles registered in the Web of Science database

Unfortunately, the most cited article with over 1.000 citations called "The role of environmental disclosures as tools of legitimacy: A research note" from Cho, Ch published in 2007 (Cho & Patten, 2007) is not related to the topic.

On the contrary the second highest cited paper with 350 citations titled "How 'Big Data' Is Different" by (Davenport et al., 2012). The text underscores the importance of organizational models in adapting big data and advanced analytics and suggesting that companies adept at leveraging it are gaining deeper business insights, fostering innovation, and responding swiftly to change, contrasting sharply with traditional data analysis approaches, thus highlighting the pivotal role of organizational structure in navigating the evolving landscape of data-driven decision-making and facilitating the emergence of dynamic information ecosystems.

Article "Making Advanced Analytics Work For You" (Barton & Court, 2012) published in Harvard Business Review has 297 citations. The text stresses the critical role of organizational models in effectively harnessing big data, emphasizing the necessity for companies to identify relevant data sources, build focused analytics models aligned with business objectives, and most importantly, cultivate a culture and capabilities conducive to implementing analytical insights throughout the organization, highlighting that embracing big data requires a shift in mindset alongside technical adaptation, with potential benefits exceeding initial expectations.

A very interesting article is "Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems" by Ben Shneiderman (Ben, 2020)which underscores the significance of organizational models in ensuring the ethical governance and proposing 15 recommendations across team, organizational, and industry levels to enhance the reliability, safety, and trustworthiness of AI systems, ultimately aiming to mitigate risks and maximize benefits for individuals, organizations, and society.

The article "How Can SMEs Benefit from Big Data? Challenges and a Path Forward" (Coleman et al., 2016) highlights the crucial role of organizational models in facilitating the adoption of big data analytics by small and medium enterprises (SMEs), stressing the urgency for SMEs to embrace this technology to avoid falling behind, and proposes a big data maturity model as a framework to overcome barriers and enhance SMEs' capability, underlining the importance of addressing challenges through collaboration among stakeholders including policymakers, IT, business management, and data science communities to sustain European industrial and business success.

Article "Big Data Analytics in Building the Competitive Intelligence of Organizations" by (Ranjan & Foropon, 2021) has 211 citations underscores the critical role of organizational models in effectively utilizing big data for Competitive Intelligence (CI), emphasizing the need for frameworks and process models to integrate big data analytics into CI strategies, highlighting challenges faced by organizations and indicating a preference for decentralized informal processes over formal structures in CI, thus accentuating the importance of adaptable organizational approaches to leverage big data effectively for informed decision-making. This article is also cited in the Scopus database, where it has 134 citations.

The article "Defining analytics maturity indicators: A survey approach" highlights the pivotal role of organizational models in the adoption and advancement of analytics within companies, emphasizing the evolution of analytics organization across various aspects defined by the DELTA model and indicating the correlation between organizational maturity and the complexity of analytics techniques employed, underscoring the importance of adaptable organizational structures in fostering analytics innovation and maturity.

Articles registered in the Scopus database

In the case of the Scopus database article "Tutorial: Big data analytics: concepts, technologies, and applications" by H.J. Watson (Watson, 2014) p 264 citation underscores the critical role of organizational models in effectively harnessing the potential of big data analytics, highlighting the necessity for clear business alignment, committed sponsorship, and skilled personnel, while also emphasizing the emergence of big data as a new generation of decision support data management, and raising significant privacy concerns necessitating careful implementation within organizations.

The article "Managing a big data project: The case of Ramco cements limited" (Dutta & Bose, 2015) underscores the vital role of organizational models in the successful implementation of Big Data projects, emphasizing the need for a comprehensive framework that addresses the complexities associated with such projects, highlighting key factors including clear understanding of business problems, detailed project planning, cross-functional teams, management involvement, and a culture of data-driven decision-making to ensure project success, thus emphasizing the importance of adaptable organizational structures in navigating the challenges of Big Data initiatives.

Article "Business analytics adoption process: an innovation diffusion perspective" (Nam et al., 2019) a highlights the significance of organizational models in driving the adoption of business analytics (BA) within organizations, underscoring the importance of considering technological, organizational, and environmental factors in each stage of BA adoption, thus emphasizing the crucial role of adaptable organizational structures in facilitating the successful integration of BA processes.

The latest article listed here is "Big data analytics in turbulent contexts: towards organizational change for enhanced agility" by Y. Barlette and P. Baillette (Barlette & Baillette, 2022) leveraging the capabilities of big data analytics (BDA) within the context of Industry 4.0, highlighting the necessity for organizational adaptation to fully exploit BDA's potential for enhancing organizational agility and achieving enhanced performance in turbulent environments, thus underscoring the importance of adaptable organizational structures and top management involvement in driving organizational changes to optimize BDA utilization and responsiveness.

4. Conclusion

Although these articles analyze the issue of data analytics organization from different perspectives, none of them discuss in detail specific recommendations and concrete models of how data analytics should be organized. The articles address whether or not it should be part of the IT department, but they no longer address the questions of how to organize data analytics internally, how to organize the company as a whole, whether, for example, to address analytics as a hub and spoke model, as described by, for example, Gartner (Create a Hybrid Centralized and Decentralized Data and Analytics Organizational Model, n.d.), complete centralisation or, on the contrary, specific decentralisation,

In view of this fact, it is necessary to state that this area is not sufficiently scientifically covered and it is appropriate to address it not only from the perspective of the organization of data analytics in the organization, but it is also possible to take into account, for example, the size of the organization, the industry or the data maturity of the organization.

The article has identified an area that is currently not adequately addressed and should be further analysed and scientifically investigated.

Acknowledgement

Paper was processed with support from institutional-support fund for long-term conceptual development of science and research at the Faculty of Informatics and Statistics of the Prague University of Economics and Business (IP400040) and the project IG 409033.

References

Advanced Data Analytics for Clinical Research Part I: What are the Tools? - Nicolas Zhou, Erin M. Corsini, Shida Jin, Gregory R. Barbosa, Trey Kell, Michael H. Antonoff, Mara B. Antonoff, 2020. (n.d.). Retrieved March 30, 2024, from https://journals.sagepub.com/doi/10.1177/1556984520902783

Barlette, Y., & Baillette, P. (2022). Big data analytics in turbulent contexts: Towards organizational change for enhanced agility. Production Planning and Control, 33(2-3), 105-122. Scopus. https://doi.org/10.1080/09537287.2020.1810755

Barton, D., & Court, D. (2012). Making Advanced Analytics Work For You. HARVARD BUSINESS REVIEW, 90(10), 78-+.

Ben, S. (2020). Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. ACM TRANSACTIONS ON INTERACTIVE INTELLIGENT SYSTEMS, 10(4), 26. https://doi.org/10.1145/3419764

Big data analytics capability and organizational performance measures: The mediating role of business intelligence infrastructure-Aboobucker Ilmudeen, 2021. (n.d.). Retrieved March 30, 2024, from https://journals.sagepub.com/doi/10.1177/02663821211055321

Bose, R. (2009). Advanced analytics: Opportunities and challenges. Industrial Management & Data Systems, 109(2), 155-172. https://doi.org/10.1108/02635570910930073

Chae, B. (Kevin), Yang, C., Olson, D., & Sheu, C. (2014). The impact of advanced analytics and data accuracy on operational performance: A contingent resource-based theory (RBT) perspective. Decision Support Systems, 59, 119-126. https://doi.org/10.1016/j.dss.2013.10.012

Chaudhuri, R., Chatterjee, S., Vrontis, D., & Thrassou, A. (2021). Adoption of robust business analytics for product innovation and organizational performance: The mediating role of organizational data-driven culture. Annals of Operations Research. https://doi.org/10.1007/s10479-021-04407-3

Cho, C. H., & Patten, D. M. (2007). The role of environmental disclosures as tools of legitimacy: A research note. ACCOUNTING ORGANIZATIONS AND SOCIETY, 32(7-8), 639-647. https://doi.org/10.1016/j.aos.2006.09.009

Coleman, S., Goeb, R., Manco, G., Pievatolo, A., Tort-Martorell, X., & Reis, M. S. (2016). How Can SMEs Benefit from Big Data? Challenges and a Path Forward. QUALITY AND RELIABILITY ENGINEERING INTERNATIONAL, 32(6), 2151-2164. https://doi.org/10.1002/qre.2008

Create a Hybrid Centralized and Decentralized Data and Analytics Organizational Model.(n.d.). Gartner. Retrieved February 27, 2024, from https://www.gartner.com/en/documents/3980295

Davenport, T. H., Barth, P., & Bean, R. (2012). How "Big Data" Is Different. MIT SLOAN MANAGEMENT REVIEW, 54(1), 43-+.

Dutta, D., & Bose, I. (2015). Managing a big data project: The case of Ramco cements limited. International Journal of Production Economics, 165, 293-306. Scopus. https://doi.org/10.1016/j.ijpe.2014.12.032

Ghasemaghaei, M., Hassanein, K., & Turel, O. (2017). Increasing firm agility through the use of data analytics: The role of fit. Decision Support Systems, 101, 95-105. https://doi.org/10.1016/j.dss.2017.06.004

How does business analytics contribute to business value? - Seddon-2017-Information Systems Journal-Wiley Online Library. (n.d.). Retrieved March 30, 2024, from https://onlinelibrary.wiley.com/doi/10.1111/isj.12101

Korsten, G., Aysolmaz, B., Turetken, O., Edel, D., & Ozkan, B. (2022). ADA-CMM: A Capability Maturity Model for Advanced Data Analytics. http://hdl.handle.net/10125/79362

Mansell, I. J., & Ruhode, E. (2019). Inhibitors of business intelligence use by managers in public institutions in a developing country: the case of a South African municipality. South African Journal of Information Management, 21(1), Article 1. https://doi.org/10.4102/sajim.v21i1.1004

Nam, D., Lee, J., & Lee, H. (2019). Business analytics adoption process: An innovation diffusion perspective. International Journal of Information Management, 49, 411-423. Scopus. https://doi.org/10.1016/j.ijinfomgt.2019.07.017

Phillips, T. B., & Lanclos Iii, R. J. (2014). Data Analytics in Procurement Fraud Prevention: Defense Technical Information Center. https://doi.org/10.21236/ADA626749

Ranjan, J., & Foropon, C. (2021). Big Data Analytics in Building the Competitive Intelligence of Organizations. INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT, 56, 102231. https://doi.org/10.1016/j.ijinfomgt.2020.102231.

Watson, H. J. (2014). Tutorial: Big data analytics: Concepts, technologies, and applications. Communications of the Association for Information Systems, 34(1), 1247-1268. Scopus. https://doi.org/10.17705/1cais.03465

ONLINE REPUTATION OF AI AND ML SUPPLY CHAIN FIRMS & SOLUTIONS: AN EMPIRICAL STUDY

Michal Konečný, Yaroslava Kostiuk
Institute of Technology and Business in České Budějovice
Faculty of Corporate Strategy
27826@mail.vstecb.cz, 26567@mail.vstecb.cz

Kristián Kalamen, Zuzana Dzilská
University of Economics in Bratislava
Faculty of Business Management
kristian.kalamen@euba.sk, zuzana.dzilska@euba.sk

DOI: 10.35011/IDIMT-2024-361

Keywords

Corporate reputation, AI, ML, supply chain management, sentimental analysis

Abstract

This paper explores the challenges associated with sustainably building corporate reputations and supply chain solutions in Artificial Intelligence (AI) and Machine Learning (ML). Special emphasis is placed on online reputation, which is a key factor in forming a responsible and sustainable image and is considered a valuable but also vulnerable intangible asset. A sample of the top 10 companies and supply chain solutions in AI and ML was selected for the research. This sample was selected based on the ranking of top 10 companies and supply chain solutions in AI and ML published on supplychaindigital.com on May 24, 2023. The analysis was conducted using sentiment analysis method. The findings identified in this study provide insights into the challenges associated with sustainable reputation building for companies. These findings provide a better understanding of how to achieve sustainable reputation development for companies and supply chain solutions in AI and ML but can also be applied to other relevant business areas.

1. Introduction

Digital transformation allows supply chains to access vast, accurate data, which should be combined with analytics for better management (Straková et al., 2022). AI is an effective technique due to its ability to handle uncertain and dynamic information (Rodriguez et al., 2021). Modern supply chains differ greatly from the past and require technological solutions to manage increasing complexity. While some AI applications in supply chains exist, studies on AI in these processes are limited. Machine learning, natural language processing, and robotics are potential transformation factors (Riahi et al., 2021). As global supply chain complexity increases, businesses seek effective business analytics (BA), business intelligence (BI), and AI tools to manage supply chain risks. Existing tools often rely on subjective expert assessments, which could be partially or fully replaced by AI for greater objectivity and efficiency. Transitioning from BA to AI in supply chain risk management (SCRM) is complex (Zigiene et al., 2022). Despite rapid AI research growth in supply chains and production management, empirical knowledge on AI adoption in small and medium enterprises (SMEs) and its impact on sustainable practices and supply chain resilience is lacking (Dey et al., 2023). AI is expected to become a significant tool in supply chain management (Hao & Demir, 2023), but broader AI adoption faces challenges due to insufficient explanation or limited approaches (Kosasih et al., 2023). Recent disruptions like pandemics (Straková et al., 2021) highlighted the need for resilient supply chains (SCRes) capable of handling interruptions. Advanced information processing techniques, such as AI, contribute to building resilient supply chains and improving performance (SCP) (Belhadi et al., 2021). AI can analyze complex data in dynamic situations, such as supply chain disruptions (Gupta et al., 2021). During COVID-19, AI emerged as a tool to enhance supply chain resilience through business continuity capabilities (Modgil, Singh & Hannibal, 2022). The pandemic emphasized the need for increased supply chain resilience (SCR), sparking research interest in overcoming risks and disruptions for successful project management. Every supply chain requires innovative projects for competitive advantage (Naz et al., 2022). AI and big data analysis (BDA) significantly enhance supply chain resilience and resource management efficiency (Zamani et al., 2023). Supply chain management (SCM) evolves in response to Industry 4.0 elements like AI, machine learning (ML), the Internet of Things (IoT), and big data (BD). These tools address SCM challenges at all levels, improving operations, demand volatility management, cost fluctuations, and data-driven decision-making. Industry 4.0 tools are catalysts for SCM development, with increased usage recently observed (Younis & Wuni, 2023). Attention to agri-food supply chain efficiency grows due to logistics costs and Industry 4.0 technologies like AI, driving research in operations management (Bhilat, El Jaouhari & Hamidi, 2024). AI is increasingly seen as a competitive advantage in operations and SCM (OSCM), but implementation challenges and limited empirical studies persist (Cannas et al., 2023). AI in food supply chains (FSC) can address food safety, quality, and waste through improved transparency and traceability. However, FSC technology adoption literature is in early stages, with limited success factor information (Dora et al., 2022). AI applications in customer service accelerate market processes, but the impact on supply chains in developing countries is still uncertain (Khalifa, Abd Elghany & Abd Elghany, 2021). Demand forecasting is crucial in SCM 4.0 for balancing supply and demand, enhancing decision-making efficiency. AI can analyze data throughout the supply chain, improving performance and understanding customer needs (Terrada, El Khaili & Ouajji, 2022). The rise of generative AI (Gen-AI) like ChatGPT has significant implications for business operations and SCM (O & SCM), but empirical evidence on its impact is limited (Wamba et al., 2023). Consumers' higher expectations push companies to optimize processes and fast product delivery, emphasizing online reputation management (Zrakova, Demjanovičová & Kubina, 2019). Enhancing business process quality improves economic results and corporate social responsibility reputation (Strakova & Kostiuk, 2023). Online corporate reputation (OCR) management is vital for economic performance, with online comments considered intangible assets impacting tangible value (Ramos & Casado-Molina, 2021). Corporate reputation is crucial for success, requiring constant monitoring for threat detection (Pollak, Dorcak & Markovic, 2021). Pandemic-driven consumer behavior changes highlight maximizing benefits (Markovic et al., 2022). Internet benefits for business focus on online activities and promotional tools (Dorcak, Strach & Pollak, 2015). Individuals leave digital footprints in online communities, making global reputation data essential (Allahbakhsh et al., 2022). AI-driven digital avatars, or "virtual influencers," are now key marketing partners for major brands, avoiding human influencer risks (Allal-Chérif, Puertas & Carracedo, 2024). The rise of social media amplifies online reputation, but effective management solutions are still lacking (Samara et al., 2021).

2. Objectives and methods

The main purpose of the study is to provide an analytical perspective on the online reputation of the top 10 companies and supply chain solutions in the AI and ML space. The research sample consists of the top 10 companies and solutions in AI and ML ranked by supplychaindigital.com (2023). Sentiment analysis will be the main methodological tool to quantify the sentiment or polarity of the top ten Google search results for a specific entity. This method, known as opinion mining, automates the quantification of subjective content to identify the author's stance on a given topic. The process starts with identifying relevant industry representatives and competitors. The first ten search results will be considered in the sentiment analysis. To minimize bias, a proxy anonymizer will be used. Using a well-known name of the study subject or object as the search phrase is essential for accurate analysis results. The sentiment of the results will be evaluated based on positive, neutral, and negative ratings, providing a picture of the online reputation of the research subject. The sentiment scores will be summed to give a final value, used as a baseline to judge a company's success in a segment. Search results for each of the ten positions will be quantified according to Table 1:

Table 1. Sentiment of results / position of results

Sentiment/Position of the result	1	2	3	4	5	6	7	8	9	10
+										
Positive sentiment	20	19	18	17	16	15	14	13	12	11
x										
Company owned website	10	9	8	7	6	5	4	3	2	1
±										
Neutral sentiment	2	2	2	2	2	2	2	2	2	2
-										
Negative sentiment	-20	-19	-18	-17	-16	-15	-14	-13	-12	-11

Source: (Liu, 2012 In: Dorčák, Pollák & Szabo, 2014)

Incognito search mode will be used to minimize personalization of results, quantifying only organic results and excluding ads. Multiple links to the subject's website within the results will be assigned a neutral sentiment. Polarity will be determined based on the link's title and description. To determine overall sentiment polarity, the polarity of individual words in the text will be analyzed. Adjectives and adverbs are suitable for determining text orientation. Superlative adjectives and positive or negative meanings of adjectives and verbs signal the message's mood. It is important to consider the object to which the words apply and assign numerical values to determine overall sentiment. Each entity analyzed can score a maximum of 155 points, representing 100%, with one percentage point equaling 0.645 points. The ranking, indicating the relative position of the test subjects (SA score), will be determined based on their overall online reputation percentage.

3. Results and discussion

We subjected selected Artificial Intelligence (AI) and Machine Learning (ML) companies and supply chain solutions to basic sentiment analysis. Table 2 below shows the values of each monitoring indicator for each of these analysed companies and solutions:

Table 2. Basic reputation score

Brand/Position results	1	2	3	4	5	6	7	8	9	10	Score SA (%)
AWS SageMaker	10	2	2	2	2	2	2	2	2	2	18,06
Google Cloud Vertex AI	10	2	2	2	16	15	14	2	2	2	43,23
Microsoft Azure ML	10	2	2	2	2	2	2	2	2	2	18,06
IBM Watson	10	2	2	2	-16	2	2	2	12	2	12,90
SAP	10	2	2	2	16	2	2	2	2	2	27,10
OCI AI Services	10	2	2	2	2	2	14	2	2	2	25,81
Siemens	10	19	2	2	2	2	2	13	2	2	36,13
Dataiku	10	2	2	2	16	2	2	2	2	2	27,10
SAS Machine Learning	10	2	2	2	2	2	2	2	2	2	18,06
Databricks	10	2	2	2	16	2	2	2	2	2	27,10

Source: (authors)

After identifying the sentiment of each of the ten search results for each company and supply chain solution in AI and ML, they were assigned a score based on the sentiment rating scale from Table 1. Table 2 shows the resulting scores. Based on the sum of these scores, the ranking of the selected AI and ML companies and supply chain solutions can be determined: 1. Google Cloud Vertex AI, 2. Siemens, 3. SAP, Dataiku, and Databricks with equal scores, 4. OCI AI Services, 5. AWS SageMaker, Microsoft Azure ML, and SAS Machine Learning with equal scores, and 6. IBM Watson. Sentiment analysis shows that the companies' websites ranked in the top positions in search results across all companies and supply chain solutions surveyed in the AI and ML space. This phenomenon is likely the result of SEO optimization that seeks to increase the visibility of a company's website and ensure it ranks in the top positions in search results. In the case of AWS SageMaker, the company's website ranked 1st, 4th and 9th in Google search results. For Google Cloud Vertex AI, the company's website ranked 1st, 2nd, 3rd and 10th in Google search results. For Microsoft Azure ML, the company's website ranked 1st, 2nd, 4th, 5th, 6th, 7th, 8th and 10th in Google search results. For IBM Watson, the company's website ranked 1st, 3rd, and 10th in Google search results. In the case of SAP, the company's website ranked 1st, 3rd and 10th in Google search results. For OCI AI Services, the company's website was ranked 1st, 2nd, and 8th within Google search results. For Siemens, the company's website was ranked 1st, 3rd, and 5th in Google search results. For Dataiku, the company's website was ranked 1st in Google search results. For SAS Machine Learning, the company's website ranked 1st, 2nd and 7th in Google search results. For Databricks, the company's website was ranked 1st in Google search results. The second finding that emerged from the sentiment analysis is that some posts, such as those on wikipedia.org, were considered neutral because they did not express any particular sentiment. In the case of AWS SageMaker, wikipedia.org was in the 3rd position within Google search results. In the case of IBM Watson, wikipedia.org was in the 2nd position within

Google search results. For SAP, wikipedia.org was ranked 2nd, 7th and 8th within Google search results. For Siemens, wikipedia.org was ranked 4th and 10th in Google search results. For Dataiku, wikipedia.org was ranked 2nd in Google search results. For Databricks, wikipedia.org was ranked 4th in Google search results. The positivity of the result in the 5th position within Google search results for Google Cloud Vertex AI can be explained by the fact that towardsdatascience.com (2021) describes the new Vertex AI platform as a development in the field of machine learning on the Google Cloud platform, indicating a positive attitude towards this new development and its benefits. The positivity of the result in the sixth position within Google search results for Google Cloud Vertex AI can be explained by Datadog Docs (2024) describing Google Cloud Vertex AI as a tool that allows developers to train their own machine learning models with minimal knowledge, indicating a positive attitude towards its ability to provide effective tools for developers. The positivity of the result in the seventh position within Google search results for Google Cloud Vertex AI can be explained by the positive review of the book *Learning Google Cloud Vertex AI: Build, deploy, and manage machine learning models with Vertex AI* by Hemanth Kumar K on amazon.com (2023), where this product received a rating of 5 stars out of 5 possible. The negativity of the post in the fifth position in Google search results for IBM's Watson can be explained by reference to an article entitled *What Ever Happened to IBM's Watson?* published in The New York Times (2021) regarding the fact that IBM's artificial intelligence was set to transform industries and generate wealth for society. Neither of the above came to fruition, however, and IBM ultimately opted for a more modest vision in the case of Watson. The positivity of the ninth position post in Google search results for IBM Watson can be explained by the positive review on live-agent.cz (2023), where the product received a rating of 4.1 stars out of a possible 5. The positivity of the result in the fifth position within the Google search results for SAP can be explained by a blog post on itica.cz (2024), where it is positively reviewed as the largest and most expensive enterprise system in the world used by the largest companies. The positivity of the result in the seventh position within Google search results for OCI AI Services can be explained by the fact that thesoftwarereport.com (2024) describes Oracle's new OCI Generative AI service as a pioneer in the field, streamlining the process of adapting AI models and providing benefits to businesses. The positivity of the result in the second position within the Google search results for Siemens can be explained by the fact that BSH domácí spotřebiče s.r.o. (2024) describes Siemens home appliances on its website as cutting-edge technology and welcomes readers to the future of modern living, promoting smart technology and distinctive home appliance design. The positivity of the result in the eighth position within the Google search results for Siemens can be explained by the fact that Siemens Czech (2024), in a post on its Facebook profile, announces that it is launching a new variant of Siemens' ultra-high-performance charging station, which has a maximum output of 400 kW and enables dynamic charging. The positivity of the result in the fifth position within Google search results for Dataiku can be explained by the fact that the article on xflow.cz (2024) describes Dataiku as a means that enables businesses to create value from their data in a human-centric way, while at the same time breaking down silos and encouraging collaboration. The positivity of the result in the fifth position within Google search results for Databricks can be explained by the fact that the portal cz.billigence.com (2024) describes the Databricks platform as a tool that simplifies data science and machine learning processes, indicating a positive perception of the benefits and advantages of this technology.

4. Conclusion

An effective online reputation management model focuses on key determinants of reputation, such as Google search results and actively managed social media profiles. Sentiment analysis based on Google search results has several weaknesses that can affect the reliability and validity of the results.

Google's search algorithms may favor certain sources or types of information, which can lead to biased results. Search results may be tailored based on the user's search history and geographical location, causing variability in displayed results among different users. Websites with higher traffic and popularity may be prioritized over sites with higher quality content, which can skew sentiment analysis. Search algorithms may sometimes display unrelated or irrelevant results that can confuse sentiment analysis. Words with multiple meanings can lead to incorrect sentiment interpretation if not properly contextualized. Google search results may be limited to a certain time frame or specific sites, which can affect the representativeness of the data. Results in different languages may be unevenly distributed, affecting global sentiment analysis. Recommendations to overcome the aforementioned weaknesses of sentiment analysis include using multiple search engines to obtain a broader range of data, including data from professional databases and social media to increase representativeness, manually checking data samples to improve quality and eliminate unrelated results, implementing algorithms to filter out low-quality or irrelevant results, using advanced algorithms capable of better understanding context, developing techniques for better recognition of irony and sarcasm in texts, using multilingual models for sentiment analysis to ensure even coverage of different language groups, and obtaining data from various geographical areas for better global representativeness. According to Sasko (2014), online reputation management tools that are actively used can greatly facilitate the communication of AI and ML firms and supply chain solutions with the public, helping them to obtain and evaluate feedback and respond quickly with crisis marketing communication actions. If an Internet user searches for information about a company or an AI and ML supply chain solution and does not come across positive indicators of its or their reputation, this may affect their overall assessment of that company or solution. This is particularly relevant for potential customers who have not yet had direct experience with the company or the AI and ML solution and are forming their opinion based on information from the Internet. Selected AI and ML companies and supply chain solutions are doing well in this respect, but it is important that they continue to improve their reputation and minimize negative publicity. This can be achieved through active online communication such as distributing positive information about the company through reputable online media or social networks. The most effective way to remove negative or neutral publicity from the first page of search results is to implement targeted marketing communications in an online environment.

Acknowledgement

This article is one of the partial outputs of the currently solved research grant VEGA no. 1/0110/24. This article is one of the partial outputs of the currently solved research project IVSUPS2305.

References

Allahbakhsh, M., Amintoosi, H., Dustdar, S., & Motahari-Nezhad, H. R. (2022). Sharing Reputation Data Across Online Communities. IEEE Internet Computing, 26(4), 60-67

Allal-Cherif, O., Puertas, R., & Carracedo, P. (2024). Intelligent influencer marketing: how AI-powered virtual influencers outperform human influencers. Technological Forecasting and Social Change, 200

Belhadi, A., Mani, V., Kamble, S. S., Khan, S. A. R., & Verma, S. (2021). Artificial intelligence-driven innovation for enhancing supply chain resilience and performance under the effect of supply chain dynamism: an empirical investigation, Annals of Operations Research

Bhilat, E. E., El Jaouhari, A., & Hamidi, L. S. (2024). Assessing the influence of artificial intelligence on agri-food supply chain performance: the mediating effect of distribution network efficiency. Technological Forecasting and Social Change, 200

Cannas, V. G., Ciano, M. P., Saltalamacchia, M., & Secchi, R. (2023). Artificial intelligence in supply chain and operations management: a multiple case study research. International Journal of Production Research

Databricks. Billigence Pty Ltd. Retrieved 12 April 2024, https://cz.billigence.com/technologie/databricks/

Dataiku. Xflow. Retrieved 12 April 2024, https://www.xflow.cz/firmy/detail/659-dataiku

Dey, P. K., Chowdhury, S., Abadie, A., Yaroson, E. V. & Sarkar, S. (2023). Artificial intelligence-driven supply chain resilience in Vietnamese manufacturing small- and medium-sized enterprises. International Journal of Production Research

Dora, M., Kumar, A., Mangla, S. K., Pant, A., & Kamal, M. M. (2022). Critical success factors influencing artificial intelligence adoption in food supply chains. International Journal of Production Research, 60(14), 4621-4640

Dorčák, P., Pollák, F., & Szabo, S. (2014). Analysis of the Possibilities of Improving an Online Reputation of Public Institutions. IDIMT-2014: Networking Societies - Cooperation and Conflict: 22nd Interdisciplinary Information Management Talks. Poděbrady, Czech Republic, 2014. pp. 275–282

Dorcak, P., Strach, P., & Pollak, F. (2015). Analytical View of the Perception of Selected Innovative Approaches in Marketing Communications. Quality Innovation Prosperity, 19(1), 74-84

Durisinova, M. (February 9, 2023). Recenze IBM Watson Assistant. LiveAgent. Retrieved 9 February 2023, https://www.live-agent.cz/reviews/recenze-chatbota/ibm-watson-chatbot/

Google Cloud Vertex AI. Datadog Docs. Retrieved 12 April 2024, https://docs.datadoghq.com/integrations/google_cloud_vertex_ai/

Gupta, S., Modgil, S., Meissonier, R., & Dwivedi, Y. K. (2021). Artificial Intelligence and Information System Resilience to Cope With Supply Chain Disruption. IEEE Transactions on Engineering Management.

Hao, X. & Demir, E. (2023). Artificial intelligence in supply chain decision-making: an environmental, social, and governance triggering and technological inhibiting protocol. Journal of Modelling in Management

Inteligentní řešení pro váš domov. BSH domácí spotřebiče s.r.o. Retrieved 12 April 2024, https://www.siemenshome.bsh-group.com/cz/

Khalifa, N., Abd Elghany, M., & Abd Elghany, M. (2021). Exploratory research on digitalization transformation practices within supply chain management context in developing countries specifically Egypt in the MENA region. Cogent Business & Management, 8(1)

Kosasih, E. E., Papadakis, E., Baryannis, G., & Brintrup, A. (2023). A review of explainable artificial intelligence in supply chain management using neurosymbolic approaches. International Journal of Production Research

Lakshmanan, L. (2021, May 19). Giving Vertex AI, the New Unified ML Platform on Google Cloud, a Spin. Towards Data Science. https://towardsdatascience.com/giving-vertex-ai-the-new-unified-ml-platform-on-google-cloud-aspin-35e0f3852f25

Learning Google Cloud Vertex AI: Build, deploy, and manage machine learning models with Vertex AI (English Edition). Amazon. Retrieved 12 April 2024, https://www.amazon.com/Learning-Google-Cloud-VertexAI/dp/9355515359

Liu, B. (2012). Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies, Chicago, IL, USA: Springer Cham, 167, ISBN 978-3-031-02145-9

Lohr, S. (2021, July 16). What Ever Happened to IBM's Watson? The New York Times. https://www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html

Markovic, P., Pollak, F., Vavrek, R., & Kostiuk, Y. (2022). Impact of Coronavirus Pandemic on Changes in e-Consumer Behaviour: Empirical Analysis of Slovak e-Commerce Market. Ekonomický časopis, 70(4), 368-389

Modgil, S., Singh, R. K., & Hannibal, C. (2022). Artificial intelligence for supply chain resilience: learning from Covid-19. International Journal of Logistics Management, 33(4), 1246-1268

Naz, F., Kumar, A., Majumdar, A., & Agrawal, R. (2022). Is artificial intelligence an enabler of supply chain resiliency post COVID-19? An exploratory state-of-the-art review for future research. Operations Management Research, 15(1-2), 378-398

Oracle Unveils OCI Generative AI Service, Pioneering Enterprise Adoption. The Software Report. Retrieved 28 February 2024, https://www.thesoftwarereport.com/oracle-unveils-oci-generative-ai-service-pioneeringenterprise-adoption/

Pollák, F., Dorčák, P., & Markovič, P. (2021). Corporate Reputation of Family-Owned Businesses: Parent Companies vs. Their Brands. Information, 12(2), 89

Ramos, C. M. Q., & Casado-Molina, A. M. (2021). Online corporate reputation: A panel data approach and a reputation index proposal applied to the banking sector. Journal of Business Research, 122, 121-130

Riahi, Y., Saikouk, T., Gunasekaran, A., & Badraoui, I. (2021). Artificial intelligence applications in supply chain: A descriptive bibliometric analysis and future research directions. Expert Systems with Applications, 173, 114702

Rodriguez, M. A., Alemany, M. M. E., Boza, A., Cuenca, L., & Ortiz, A. (2021). Artificial Intelligence in Supply Chain Operations Planning: Collaboration and Digital Perspectives. Boosting Collaborative Networks 4.0, 598, 365-378

Samara, F., Ondieki, S., Hossain, A. M., & Mekni, M. (2021). Online Social Network Interactions (OSNI): A Novel Online Reputation Management Solution. 2021 7th International Conference on Engineering and Emerging Technologies (ICEET 2021), 415-420

SAP program co to je. Itica. Retrieved 12 April 2024, https://itica.cz/blog/sap-blog/sap-program-co-to-je/

Sasko, J. (2014, June 9). Dbáte na hodnotu svojej značky? Online reputačný manažment. Podnikajte.sk. https://www.podnikajte.sk/marketing/online-reputacny-manazment

Siemens Czech (2024, March 26). Společnost Siemens uvádí na trh novou variantu ultra výkonné dobíjecí stanice #SichargeD s maximálním výkonem 400 kW. Umožňuje dynamické dobíjení až čtyř vozidel najednou. Facebook. https://www.facebook.com/photo/?fbid=810148444488909&set=a.643913847779037&locale=cs_CZ

Strakova, J., & Kostiuk, Y. (2023). Importance of Business Process Quality for Creating Added Value and Raising Reputation of Companies in Low-Carbon Economy. Energies, 16(17), 6388

Straková, J., Koraus, A., Váchal, J., Pollák F., Cernak, F., Talíř, M., & Kollmann, J. (2021). Sustainable Development Economics of Enterprises in the Services Sector Based on Effective Management of Value Streams. Sustainability, 13(16)

Straková, J., Talíř, M., & Váchal, J. (2022). Opportunities and threats of digital transformation of business models in SMEs. Economics & Sociology, 15, 159-171

Terrada, L., El Khaili, M. & Ouajji, H. (2022). Demand Forecasting Model using Deep Learning Methods for Supply Chain Management 4.0. International Journal of Advanced Computer Science and Applications, 13(5), 704-711

Top 10 AI and ML supply chain firms & solutions. (2023, May 24). Supply Chain Digital. https://supplychaindigital.com/digital-supply-chain/top-10-ai-and-ml-supply-chain-solutions

Wamba, S. F., Queiroz, M. M., Jabbour, C. J. C., & Shi, C. V. (2023). Are both generative AI and ChatGPT game changers for 21st-Century operations and supply chain excellence? International Journal of Production Economics, 265.

Younis, H., & Wuni, I. Y. (2023). Application of industry 4.0 enablers in supply chain management: Scientometric analysis and critical review. Heliyon, 9(11)

Zamani, E. D., Smyth, C., Gupta, S., & Dennehy, D. (2023). Artificial intelligence and big data analytics for supply chain resilience: a systematic literature review. Annals of Operations Research, 327(2), 605-632

Zigiene, G., Rybakovas, E., Vaitkiene, R., & Gaidelys, V. (2022). Setting the Grounds for the Transition from Business Analytics to Artificial Intelligence in Solving Supply Chain Risk. Sustainability, 14(19), 11827

Zrakova, D., Demjanovicova, M., & Kubina, M. (2019). Online reputation in the transport and logistics field. 13th International Scientific Conference on Sustainable, Modern and Safe Transport, 40, 1231-1237

ONLINE REPUTATION OF CHATBOTS THROUGH THE LENS OF POST-PANDEMIC ACCELERATED DIGITIZATION IN E-COMMERCE

František Pollák

University of Economics in Bratislava
Faculty of Business Management
frantisek.pollak@euba.sk

Institute of Technology and Business in České Budějovice
Faculty of Corporate Strategy
frantisek.pollak@vste.cz

Peter Markovič

University of Economics in Bratislava
Faculty of Business Management
peter.markovic@euba.sk

DOI: 10.35011/IDIMT-2024-369

Keywords

E-commerce; accelerated digitization; artificial intelligence; supply chain; COVID-19 pandemic; online reputation vanagement

Abstract

This study examines the impact of information technology on business competitiveness, driven by a surge in digitization due to the global pandemic. E-commerce is projected to quadruple by the end of the decade, illustrating this transformation. Additionally, the U.S. market's rapid AI growth, expected to be substantial by 2030, heralds a new era of automation and innovation across sectors. Europe shows a slight delay in AI market growth but is poised for significant catch-up. The research focuses on evaluating the online reputation of the 7 Best Chatbots of 2024, as selected by Forbes, in both U.S. and European markets. Using sentiment analysis of Google search results, the study finds varying levels of online presence and reputation among these AI solutions. U.S.-based search results show a slight advantage, but localized findings suggest minimal regional differences, indicating uniformity in content across markets. The study highlights a predominantly generic online presence, except for one notable entity, underscoring the developmental nature of these chatbots' online visibility. While achieving a basic resolution, the research calls for deeper analysis to explore complex relationships further, laying the foundation for future studies on AI adoption, market dynamics, and digital visibility.

1. Introduction

Information technology has been helping businesses increase their competitiveness for decades (Maryška, Doucek & Novotný, 2012; Delina, & Tkáč, 2010). The gradual digitization of the last decade, especially when it comes to increasing the presence of business companies in the social media environment (Sagapova, Dušek & Pártlová, 2022), took on a significantly accelerated character in the context of the global pandemic. From day to day, trading companies were forced to saturate the demand of their customers almost exclusively in the form of e-commerce. The Statista platform (2024a) reports that the volume of sales in an international context increased by almost 2.5 trillion US dollars between the years, from the original 3.351 trillion in 2019 to 5.784 trillion in 2023. This volume is expected to rise to more than 8 trillion by the year 2027. Over the course of one decade, this would mean an almost fourfold increase in sales, which can only confirm the statement about accelerated digitization. Another interesting phenomenon is the rise of artificial intelligence in terms of the market in the United States, Statista (2024b) reports that between 2020 and 2030 the AI market will grow from 45 billion in 2020 to 273 billion in 2030. Europe lags this assumption, but the catch-up is not significant, Statista (2024c) reports that the market will grow from 23 billion in 2020 to almost 210 billion in 2030 from the initial inter-decadal comparison. We can thus monitor the emergence of a new market in real time. In the context of artificial intelligence, automation should take control over the implementation of a significant number of supporting and main processes in business (Adam, Wessel & Benlian, 2021; Miklosik, Evans & Qureshi, 2021). Currently, artificial intelligence tools are most widely applied in the processes of consumer and technical services, process automation, or cyber security. By 2025, he foresees the greatest growth within the processes of supporting the internal functions of business companies, or processes in the field of human resources management (Thormundsson 2024a). While the current state of involvement of AI tools in business processes is described by representatives of business companies as either the same state as their competitors, or the state at the level of industry leaders (Thormundsson 2024b). In layman's terms, that doesn't sound very convincing. While the state values of the industry are relatively low, at the same time the expectations are relatively high, we consider it necessary to examine the Real Visibility of AI tools in the online environment, with the aim of defining a baseline for further investigation. The research problem is defined as the need to investigate the reputation level of the best AI Chatbot applications. At the same time, compare the level of reputation on the market in the country of origin (United States) with the level of reputation on the selected European market characterized by the nature of the catch-up market. The research questions are formulated as follows:

RQ1: To what extent are the selected platforms established in the online space from the point of view of their presence in Google search results?

RQ2: To what extent is it possible to track the regional specifics of applications in the online space from the point of view of their basic level of online reputation?

The study as such is divided into four parts, where the introduction is followed by a chapter describing the methodological apparatus used. Subsequently, the findings of the empirical analysis are presented and discussed.

2. Objectives and methods

The research problem is defined as the need to investigate the reputation level of the best AI Chatbot applications. At the same time, compare the level of reputation on the mother market of the United States with the level of reputation on the selected European market characterized by the nature of the catch-up market. The object of the research is the market segment of AI solutions represented by Chatbots; computer programs that mimic human conversation and make it easy for people to interact with online services using natural language. They help businesses automate tasks such as customer support, marketing and even sales. The research set thus represents a relatively difficult to quantify portfolio of AI solutions, while the sample was identified based on Forbes (2024) advisors, who selected the 7 best AI solutions for 2024. To find the best chatbots for small businesses Forbes advisors analyzed the leading providers in the space across several metrics. From pricing and plans to ease of use to the key features that matter most to businesses looking to use a chatbot to streamline customer service and supercharge sales, such as prebuilt conversations templates, AI-powered chatbots, natural language processing and more. Advisors also considered user reviews and customer support to get a better understanding of real customer experience. These 7 best AI solutions for 2024 are thus representatives of the developing industry which were subjected to an analysis of their online reputation in the Google search engine environment. Through the basic sentiment analysis (SA), we reviewed in a semi-automated way through an online tool MonkeyLearn (2024) the polarity of the first 10 search results for each of the analyzed subjects. While the first ten results by default represent the first page of search results for a specific subject. The basic values were positive, neutral, and negative sentiment. Based on the chosen methodology (Liu, 2012; Pollák, Dorčák & Markovič, 2021), these values were extended by the value "subject's own page". Based on the position in the search, the measured values were assigned points according to the following Table 1.

Table 1. Sentiment Individual Results/Position of Results

Sentiment / Position	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Positive sentiment (+)	20	19	18	17	16	15	14	13	12	11
Own website (x)	10	9	8	7	6	5	4	3	2	1
Neutral sentiment (±)	2	2	2	2	2	2	2	2	2	2
Negative sentiment (–)	–20	–19	–18	–17	–16	–15	–14	–13	–12	–11

Source: (Pollák, Dorčák & Markovič, 2021)

We have excluded sponsored links from the first ten search results. If we find our own page on several links in the form of sub-domains, we classify it as an occurrence of a neutral nature. To analyze the polarity of the search result, we used the MonkeyLearn online service. While we used the Title and Per extensum/ perex of the link as text. Regarding the proper name of the subject, we used the official names of AI solutions as defined in the overview table in the Forbes (2024) ranking. Solutions that did not have a mention of technology or tools in their name were not further modified or supplemented. As such, the collection took place in two geographical and linguistic locations with the aim of creating an empirical base for answering the stated research questions. The data collection and analysis itself took place in the month of April 2024. From the point of view of data processing, it is a semi-automated collection without the intervention of the human factor in the process of evaluating the sentiments of the results. The summary score for each of the analyzed subjects (SA) was converted into percentages according to the chosen methodology (Pollák, Dorčák & Markovič, 2021). The data itself was subsequently evaluated using the MS Excel spreadsheet. The findings are presented through a summary table and a histogram.

3. Results and discussion

Individual AI solutions were subjected to a basic sentiment analysis (SA) in the language and geographic localization environment of Slovakia and likewise in the language and geographic localization environment of the USA. The results of the measurements are presented in the following Table 2.

Table 2. Findings across measures

Rank	Subject/ Result sentiment	SA Score SR (%)	SA Score US (%)
1.	LivePerson	83	83
2.	WP-Chatbot	67	67
3.	HubSpot Chatbot Builder	51	50
4.	Salesforce Einstein	40	43
5.	Genesys DX	30	47
6.	Intercom	17	34
7.	Drift	15	18

Source: (Own processing)

As for the qualitative description of the individual findings, they can be summarized for the performed analyzes as follows.

3.1. Qualitative evaluation of findings

The results are divided within the sub-chapter based on language and geographical location.

3.1.1. Slovak language and geographical localization

HubSpot Chatbot Builder- First it finds its own page, followed by its two sub-domains. Subsequently, there are three links to websites promoting AI applications within their own solutions or third-party solutions. The seventh and eighth place belongs to the link to your own subdomain. The ninth reference compares the features of the AI with a competitive solution, we classify it as a reference of a neutral nature. The last section is closed by a positive link discussing the application possibilities of the given AI.

Intercom- In the first two places in the organic search results, we find irrelevant links, probably due to the choice of a name referring to a specific communication technology. In third place is a link to the organization's own page. The links in the remaining positions are completely irrelevant to the evaluated subject due to the inappropriately chosen name. An interesting finding is the fact that when it comes to paid links, they link exactly to a specific AI solution.

Drift- The first four places are occupied by links of an irrelevant nature. The company's website is only in fifth place in the search results. The links in the sixth to tenth place are also irrelevant from the point of view of content.

Salesforce Einstein- In the first place we find the subject's own page. The following three links are linked to the sub-domain of the subject. In the fifth position we find a link of positive sentiment, this is a page presenting the use of technology. Next in order are two links linking to the company's sub-domain. The eighth position refers to the occurrence of positive sentiment. order closes the link with a neutral reference to the sub-domain and essentially a generic reference to the technology, which is classified as positive.

WP-Chatbot- In the first place we find a link to the organization's own page, it is followed by a link to the sub-domain. The third place belongs to the link explaining technology, it has a positive sentiment polarity. It is followed by a pair of similar messages of positive sentiment. The sixth is a link to the organization's sub-domain. The following is a page with a tutorial, the title and perex of which we evaluate as the occurrence of positive sentiment in the seventh position. This is followed by a sub-domain link and a pair of essentially generic positive sentiment links.

LivePerson- In the first position we find the organization's own page. In the second position we find a link to LinkedIn of positive sentiment. Followed by a link to the sub-domain and a positive Wikipedia link. The following is a trio of links to major social media outlets, all showing the nature of the positive sentiment. The ranking is rounded off by three links of positive sentiment that point to significantly well-managed optimization for search engines, which is the first in the examined sample not to show a significant artificial nature of the results.

Genesys DX- Custom page is followed by four links to variable sub-domains. The link in the fifth position has a positive sentiment. The following is a link to LinkedIn, which again has the character of positive sentiment. This is followed by a custom sub-domain and a negative link of a generic nature. It is followed by a link of positive sentiment, a tutorial on the YouTube platform, to which we assign the character of positive sentiment based on the description. The order closes the link to its own sub-domain.

3.1.2. English language with USA geographic localization

HubSpot Chatbot Builder- The first four links are completely identical to the search localized to Slovakia. In fourth place, we find a positive sentiment page describing the benefits of a specific AI solution. The sixth link is a link to the sub-domain of your own page. The seventh link refers to a page discussing the high price of an AI solution. The analyzer classifies it as neutral sentiment. The following are links with neutral and twice positive sentiment. The evaluation is thus closed with an almost identical score regardless of language localization.

Intercom- In the first place we find the organization's own page. This is followed by an irrelevant link to Wikipedia. In the third place we find a link to platform X and the profile of the given AI solution, the sentiment of the link shows signs of a positive reference. The fourth to ninth positions are occupied by references of an irrelevant nature. On the last rung we find a message of positive sentiment.

Drift- In the first place we find the company's own page. The following five links are content-irrelevant. The seventh place belongs to the sub-domain of the organization. Eighth to tenth place is dominantly irrelevant in nature.

Salesforce Einstein- In the first two places we find a link to our own page and then its sub-domain. The third place refers to positive sentiment. The fourth is a link to a sub-domain and subsequently to a page reporting on technology, we evaluate the sentiment of this link as positive. The same situation is repeated on the next two bars. While the link of the sixth position links to a renowned medium. The seventh to ninth rungs belong to links to sub-domains. The tenth reference is basically a generic mention of technology, based on the methodology we classify it as a positive sentiment.

WP-Chatbot- The first nine links are identical to the links recorded in the analysis localized to the Slovak market. The tenth link is almost completely identical in character to the tenth link within the local search, but in this case it is a different domain publishing an essentially identical text.

LivePerson- The first seven cities are almost identical to the geographically regionally allocated measurement. The links on the eighth to tenth rungs have a significantly positive sentiment.

Genesys DX- The first five links are identical to the regionally localized measurement. The following three links are of a positive sentiment nature. The ranking is closed by the neutral sentiment of the own sub-domain and the link of positive sentiment on YouTube.

3.2. Discussion of specific findings

When it comes to search results in the first ten positions, reputable entities generally benefit from maximum optimization (Pollák, Markovič & Majdúchová, 2023). It is the domain of developing markets, respectively of established products, that we find results of a neutral nature in the first ten search places during the analysis. The combination of developing markets and established products creates a situation that could be described as a reputational vacuum. We find this vacuum in the analysis of the selected sample. All links, regardless of order, seem very generic and even artificial. Thematically, they are limited only to the direct applicability of solutions to the tool portfolio of companies. The texts have the form of tutorials with the absence of non-generic content. As for specific AI solutions with a clear and specific designation, the links have almost the same content and source, regardless of geographic location. If we consider the linguistic and geographical localization of the Slovak market, the order, content, and sources of the results are taken from the linguistic and geographical localization of the United States due to the lack of local search results. Regarding solutions with general designations, or designations that have a wide range of synonyms, within our analysis we find a significantly lower number of relevant links and thus a lower level of overall reputation. However, in this case, the geographical localization of the mother country turns out to be a significant advantage in the comparison of the target and reference market within the analysis carried out by us. In general, it is possible to notice the phenomenon of reputation catching up, we noted this phenomenon in several research carried out in the past. This is a phenomenon where a significantly positive reputation of a parent entity increases the reputation of a subsidiary, in this case an application developed as part of a popular website building solution. From the point of view of the overall ranking, based on a simple sentiment analysis, it is possible to confirm the significantly developing nature of the market, regardless of regional specificity. Of the analyzed entities, only one, namely LivePerson, shows signs of an optimal online reputation. Second in line WP-Chatbot benefits from a relatively high level of reputation of the parent entity. The following Figure 1. presents a graphic illustration of the findings.

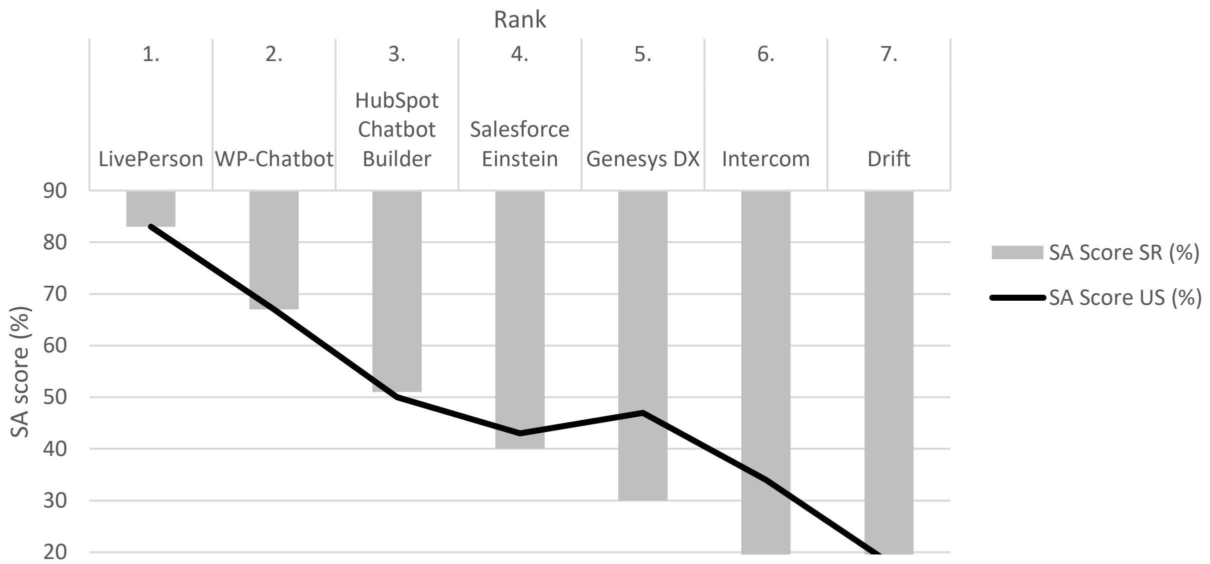


Figure 1. Reputation of entities Source: (author)

4. Conclusion

In general, therefore, the global market represented by localization in the United States means an advantage for both AI solutions with conservative and creative names, while the conservative ones reach a level roughly half higher than those with creative names without mentioning the epithet Chatbot or AI. However, the advantage over regional localization is not significant. For the best subjects in the sample, regional localization does not play a role. The results are either identical, or we find basically the same sentiment levels in the different places. This phenomenon is quite interesting and would require a more detailed investigation. When it comes to answering the first of the research questions, it can be stated that the platforms, except for two, are relatively weakly established in the online environment of Google search results. The online reputation level of most entities ranges from 15-50%. Considering that these are the most important representatives of the industry, this value is low in comparison with measurements on equally important global samples (Pollák et al, 2022; Pollák, Markovič & Majdúchová, 2023). Regarding the regional specificity and thus the second of the research questions, it is possible to observe a relatively small specificity in favor of the mother market. However, the content is largely the same regardless of the analyzed region. Regionally specific content for the benefit of the reference catch-up market is essentially non-existent. With this finding, it is possible to declare the research problem solved. This solution has the nature of a basic description of the facts. For a thorough review, it will be necessary to analyze the identified connections in depth, which we see as researchers as the biggest challenge in further researching the topic. However, with the development of the market, we have defined the initial reference base for further investigation. From the point of view of limitations, it is possible to clearly point to a relatively small sample, which was defined from a relatively indefinite set of the expanding offer of AI solutions. As we mentioned earlier in the text, the issue as such is of a considerably developing nature. Future research on the topic is essential for monitoring the development of the industry, identifying trends, and describing the phenomenon of the emergence of artificial intelligence as such. The development of issues for the needs of transfer of knowledge into practice is quite difficult due to the continuous growth of the industry. We also note a relatively rare paradox, where the application level of the topic often exceeds the academic level from the point of view of accumulation of knowledge and description of phenomena. Examining the possibilities and perspectives is thus not only prospective, but also necessary to a certain extent.

Acknowledgement

This article is one of the partial outputs of the currently solved research grant VEGA no. 1/0110/24.

References

Adam, M., Wessel, M., & Benlian, A. (2021). AI-based chatbots in customer service and their effects on user compliance. Electronic Markets, 31(2), 427-445

Delina, R., & Tkáč, M., (2010). Trust Building Mechanisms for Electronic Business Networks and Their Relation to eSkills, World Academy of Science. Engineering and Technology, 6 (71), 80-390

Liu, B. (2012). Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, Chicago, IL, USA, 2012.

Maryška, M., Doucek, P., & Novotný, O. (2012). Requirements of companies on the knowledge ICT specialists for the ICT administrator role. 4th World Conference on Educational Sciences. Barcelona, 2012. Procedia - Social and Behavioral Sciences, Vol. 46 (2012), 4389 – 4393

Pollák, F., Dorčák, P., & Markovič, P. (2021). Corporate Reputation of Family-Owned Businesses: Parent Companies vs. Their Brands. Information, 12(2), 89

Pollák, F., Markovič, P., Vavrek, R., & Konečný, M. (2022). Return to the New Normal: Empirical Analysis of Changes in E-Consumer Behavior during the COVID-19 Pandemic. Behavioral Sciences, 12(3), 85

Pollák, F., Markovič, P., & Majdúchová, H. (2023). Reputation of Electric Vehicles in the Environment of Carbon Reduction and Accelerated Digitization. Energies, 16(9), 3836

Sagapova, N., Dušek, R., & Pártlová, P. (2022). Marketing Communication and Reputation Building of Leading European Oil and Gas Companies on Instagram. Energies, 15(22), 8683

Forbes. (2024). 7 Best Chatbots Of 2024. Available from internet: https://www.forbes.com/advisor/business/software/best-chatbots/

MonkeyLearn. (2024). Free Online Sentiment Analysis Tool. Available from internet: https://monkeylearn.com/sentiment-analysis-online/

Miklosik, A., Evans, N., & Qureshi, A.M.A. (2021). The use of chatbots in digital business transformation: A systematic literature review. IEEE Access, 9, 106530-106539

Statista.com. (2024a). Retail e-commerce sales worldwide from 2014 to 2027. Available from internet: https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/

Statista.com. (2024b). Artificial intelligence (AI) market size in the United States from 2020 to 2030. Available from internet: https://www.statista.com/forecasts/1451309/market-size-of-ai-us

Statista.com. (2024c). Artificial Intelligence - Europe. Available from internet: https://www.statista.com/outlook/tmo/artificial-intelligence/europe

Thormundsson, B. (2024a). Activities in which American and European companies are using artificial intelligence (AI) in 2023 and use forecast for 2025. Available from internet: https://www.statista.com/statistics/1455572/ai-use-united-states-europe/

Thormundsson, B. (2024b). Has advanced is your company in terms of data, analytics, and artificial intelligence (AI) function in 2023?. Available from internet: https://www.statista.com/statistics/1455666/ai-function-analytics-advancement-united-states-europe/

ACADEMIC BUSINESS CO-OPERATION

ACADEMIC BUSINESS CO-OPERATION – CASES OF CZECHIA AND TAIWAN

Jan Ministr

Technical University of Ostrava
Faculty of Economics
jan.ministr@vse.cz

Tomáš Pitner

Masaryk University
Faculty of Informatics
pitner@muni.cz

DOI: 10.35011/IDIMT-2024-379

Keywords

Cooperation between academic environment and business; contractual research; new trends in research in business; government role

Abstract

This paper explores the cooperation between academic environments and businesses in Czechia and Taiwan, focusing on their economic structures and the role of universities. Both countries have export-oriented economies, with Czechia specializing in machinery and automotive products and Taiwan in integrated circuits and technology. They are deeply integrated into global supply chains and prioritize developing a skilled workforce. SMEs are vital to both economies, contributing significantly to employment and GDP. The paper highlights the importance of government facilitation and university-industry collaborations in driving innovation and economic growth, drawing lessons from Taiwan's experience to inspire similar advancements in Czechia. Key strategies include the development of science parks, fostering an entrepreneurial culture, and the establishment of germination centers to support early-stage innovation as well as fostering world-class universities, knowledge transfer, licensing and contractual research.

1. Introduction and Context

1.1 Goal and Structure of the Paper

Global significance of Taiwan semiconductor industry together with growing direct presence of the semiconductor production facilities (fabs) in the heart of Europe (namely East Germany) with expected related value-chain to span across the entire (at least) Central Europe attracts attention of Europe to investigate the roots of Taiwanese success in core technology needed for virtually every

PC and electronic gadgets but also each car. Historically, the Taiwanese progress from agriculture-based to high-tech economy started in early 70s with decisions to develop dedicated science parks that would eventually concentrate a) qualified human power, b) proximity of universities, c) natural resources (e.g. water), energy and d) know-how that time mostly imported from the U.S.

Similarities between Taiwanese and other countries in terms of the structure of economies and academic sectors would mean more chance to successfully transfer Taiwanese experience to other context – from mostly semiconductor to other industrial (high-tech) sectors and apply models of academic-industrial cooperation. On the other hand, significant differences in important factors would limit transferability of related instruments to Czechia.

Nevertheless, to be able to transfer as much inspiration from Taiwan could significantly help the Czech academic sector to reflect the needs of Czech economy and society better than today. It holds true particularly in current economic perspectives setup after the announcement of the huge-scale semiconductor industry in Europe, specifically in Czechia and Germany, in the latter case directly from Taiwan's TSMC.

2. Structure of Czech and Taiwanese Economies

The Czech and Taiwanese economies share several similarities despite being geographically distant and culturally distinct. For the explorative part, we used ChatGPT 4o (paid version) for outlining the main areas of comparison between the two countries. Then, for each identified factor, we used regular authoritative statistics, such as OECD or specialized ones.

Here are the main factor representing similarities and differences between the two countries:

Export-Oriented Economies. Both the Czech Republic and Taiwan have economies that are heavily reliant on exports. The Czech Republic's economy benefits significantly from exporting machinery and machinery products, such as cars ($25.5B), broadcasting equipment ($14.1B), motor vehicles parts and accessories ($14B), computers ($10.9B), and office machine parts ($5.4B), largely to European markets, mainly neighbouring countries. Total export volume in 2022 was $237 billion. (OEC-CZE, 2023).

Taiwan, similarly, is a major exporter of integrated circuits ($223B), office machine parts ($26.4B), computers ($16B), refined petroleum ($14.9B), and broadcasting equipment ($13.2B), with a significant portion going to global markets but mostly located in East Asia and Pacific (China, US, Japan, Singapore, Hong-Kong). Total export volume in 2022 was $542 billion. (OEC-TWN, 2023).

Integration into Global Supply Chains. Both economies are deeply integrated into global supply chains. The Czech Republic's integration is mainly within the European Union, benefiting from its central location in Europe. Taiwan is a key player in the global technology supply chain, providing essential components for a variety of electronic products worldwide. For both countries, the volume of international trade per capita is very high with larger surplus for Taiwan:

- Czechia: export = $22.5 K, import = $21.3 K
- Taiwan: export = $23.3 K, import = $17.8 K

Sources: (OEC-CZE, 2023; OEC-TWN, 2023).

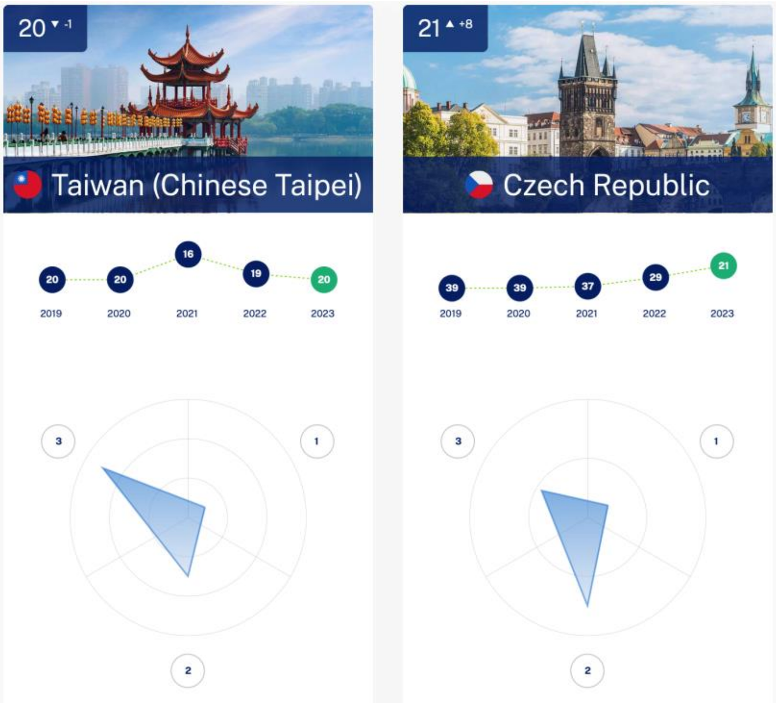


Figure 1. World Talent Ranking TWN-CZE Source: (IMD, 2024)

Skilled Workforce. Both countries have invested in developing a skilled workforce to support their advanced manufacturing sectors. The Czech Republic has a strong tradition in engineering and technical education, while Taiwan has focused heavily on education and training in science and technology fields. Positions in global rankings are very similar – total World Talent Ranking by IMD (2024) shows total scores of 69-70 for both, though. See Fig. 1. Source: IMD (2024); data until 2023

7. Total public expenditure on education: TWN: **3.52%**; CZE: **5.05%**
8. Total public exp. on education per student: TWN: $ **6,540.30**; CZE: **5,895.32**
9. Pupil-teacher ratio (primary education): TWN: **12.10**; CZE: **18.04**

Investment in Research and Development. Taiwan is a leader in technological innovation, especially in the semiconductor industry. The volume of investment is huge for Taiwan, it is on place 3rd on the OECD rank (Wikipedia-OECD, 2021) according to GDP proportion (3.77%) behind Israel and South Korea and before the U.S. Czechia showed on rank 20th with 2.00% which is significantly less. Note that the rank does not distinguish the origin of the spending: both private and government is summed up. Taiwan reaches even a better place for per capita spendings: 2nd rank with $2187 vs $736 for Czechia. Note that some institutions do not include Taiwan into their statistics, such as the World Bank (World Bank, 2024).

Biggest difference is when it comes to the source of R&D budgets (OECD, 2021):

- Czechia's largest sources are from business sector 36%, government 32%, and rest-of-world (mostly EU-funds) makes 30%
- Taiwan's largest sources are from business sector 86% and the rest from the government. Namely, according to the NSTC, 82.5 percent of R&D spending by enterprises came from large-sized companies with a workforce of more than 500.

Small and Medium-Sized Enterprises. SMEs play a vital role in both economies. In the Czech Republic, SMEs are a backbone of the economy, contributing to employment and GDP. Taiwan's economy is also characterized by many SMEs, especially in the technology sector, which are highly dynamic and innovative. According to Country Profiles by OECD (2018), in 2018, "*there were roughly 1.155 million active enterprises in the Czech Republic. 99.83% of these firms were SMEs with less than 250 employees each. Together, they employed almost 1.88 million people, or 57.68% of the Czech Republic's workforce. Micro-firms dominated the business landscape, comprising 96.4% of all SMEs in 2018 (roughly stable from 2017)"*. While in Taiwan, "*as of 2021, Taiwan boasted more than 1.59 million small and medium enterprises (SMEs), according to the White Paper on Small and Medium Enterprises in Taiwan, 2022. This accounts for more than 98 percent of all enterprises, an all-time high. Further, SMEs employed 9.2 million people, representing more than 80 percent of the total workforce. In the same period, Taiwanese SMEs exceeded NT$26 trillion in total revenue of which nearly 90 percent was generated in the domestic market"* as officially reported by the government of Taiwan in (MEA, 2022).

Foreign Direct Investment (FDI). Both countries attract significant foreign direct investment due to their strategic locations, skilled workforces, and favourable business environments. The Czech Republic attracts FDI primarily from Europe but the very recent announcement from ON Semiconductor changes the landscape: onsemi *selects the Czech Republic to establish end-to-end silicon carbide production for advanced power semiconductors* (onsemi, 2024) by a direct investment of $2 B.

Stability, Trade Relationships and International Partnerships. Both countries actively engage in international trade relationships and seek to build partnerships to enhance their economic standing. The Czech Republic benefits from its membership in the European Union, while Taiwan, despite its unique international status, maintains strong trade relations with major economies like the United States and Japan. These similarities highlight the shared characteristics that have contributed to the economic success of both the Czech Republic and Taiwan.

Summary. The Czech and Taiwanese economies share notable similarities despite their geographic and cultural differences. Both are highly export-oriented, with the Czech Republic focusing on machinery and automotive products, and Taiwan on integrated circuits and technology-related goods. Their integration into global supply chains is a critical economic driver, with the Czech Republic benefiting from its central location in Europe and Taiwan from its pivotal role in the global technology sector. Additionally, both countries boast high international trade volumes per capita, reflecting their strong global economic linkages. They also prioritize the development of a skilled workforce, as evidenced by similar rankings in global talent indices and substantial investments in education and technical training.

In terms of research and development, Taiwan leads with significant investments, particularly in the semiconductor industry, ranking third globally in GDP proportion spent on R&D. The Czech Republic, while investing less, still maintains a strong R&D presence. Both economies are heavily supported by small- and medium-sized enterprises (SMEs), which play a crucial role in employment and GDP contribution. Foreign direct investment is significant in both countries, drawn by strategic locations and skilled workforces, with recent major investments such as ON Semiconductor in the Czech Republic highlighting this trend. Finally, both countries actively engage in international trade and partnerships, with the Czech Republic leveraging its EU membership and Taiwan maintaining strong trade ties with major economies despite its unique international status.

3. University and Research Systems

3.1 Financing

Both countries place a strong emphasis on higher education to drive economic growth and innovation. In the Czech Republic, higher education is predominantly funded by the state, and tuition fees for public universities are relatively low for both domestic and EU students. In Taiwan, while public universities also receive significant government funding, there is a higher reliance on tuition fees from students. Czechia and Taiwan invest significantly in their university sectors to develop skilled workforces.

However, the differences are clear. According to (OECD, 2022): *All OECD countries devote a substantial share of national output to educational institutions. In 2019, OECD countries spent on average 4.9% of their gross domestic product (GDP) on primary to tertiary educational institutions. In the Czech Republic, the corresponding share was 4.3%. Between 2008 and 2019, funding for educational institutions from all sources grew by 47% in the Czech Republic. Over the same period, the increase in GDP was lower with 22%. Consequently, expenditure on educational institutions as a share of GDP grew by 0.8 percentage points over the same time.* Also, relative to GDP, public spending on primary to tertiary education (3.9%) is lower than the OECD average (4.4%). Taiwan devotes more, i.e. 5% of GDP to education. About a third of the education budget goes to higher education (NCCE, 2022).

3.2 Research and Development

Universities in both countries are key players in research and development (R&D). They collaborate with industry and government to promote innovation, particularly in science and technology fields. Taiwanese universities are particularly strong in technology, engineering, and applied sciences, reflecting Taiwan’s economic focus on high-tech industries. Czech universities have a broader focus, with strengths in engineering, natural sciences, and humanities, reflecting its diverse industrial base but also tradition.

According to OCAC (2023), In terms of science parks supervised by the NSTC (National Science and Technology Council = Ministry of Science and Technology, Taiwan), *firms in the Hsinchu Science Park, the Central Taiwan Science Park and the Southern Taiwan Science Park, accounted for 47.6 percent of all R&D spending by enterprises in Taiwan in 2022. These science parks house many semiconductor manufacturers, computer and peripherals suppliers, communications gadget developers, optoelectronics producers, precision machinery makers and biotech technology developers, with semiconductor firms spending the most on R&D.*

3.3 Internationalization

Both countries are increasingly focusing on internationalization, encouraging student and faculty exchanges, international collaborations, and attracting foreign students to enhance the global standing of their universities. Czech universities primarily use Czech as the language of instruction, though many programs, especially at the postgraduate level, are available in English. In Taiwan, while Mandarin Chinese is the primary language of instruction, there is also a significant number of programs offered in English to attract international students. The educational philosophies and administrative practices are influenced by European traditions in the Czech Republic and by Japanese and American models in Taiwan. Taiwanese universities like National Taiwan University (NTU) often rank higher in international university rankings compared to Czech universities. Taiwan has numerous partnerships with universities in the United States, Japan, and other Asian countries, reflecting its geopolitical focus. The Czech Republic, being part of the European Union, is deeply integrated into the Erasmus+ program, promoting student exchanges primarily within Europe.

There are traditionally ties between universities from both countries, such as numerous memoranda of understanding but also deeper collaborations in form of double-degrees. Recently, there are also government-supported internationalization activities between Taiwan and Czechia facilitating inter-university collaboration, e.g. the recent *program of cooperation between UAAT (University Academic Alliance in Taiwan) and ICU (CZE) university networks.*

3.4 Governance

Both the Czech Republic and Taiwan have a mix of public and private universities. Public universities tend in both cases to be more prominent and receive more funding from the government. There is no big difference is this respect. However, university governance structures differ, with Taiwanese universities typically having more centralized administrative frameworks influenced by American and Japanese systems.

Czech universities often have more autonomy and are influenced by European models of governance, with significant faculty participation in decision-making. Apart of financing by national grants (both education and research), the government cannot impose direct governance onto universities as their autonomy is guaranteed by law. The only body composed from representatives from other than academic environment is the Board of Trustees deciding (but merely passively): "The Board of Trustees is the supervisory body of a public university, which grants consent for the acquisition or transfer of higher-value real estate or movable property, the establishment of an easement or pre-emptive right, and the founding of a legal entity and the contribution of assets to it. … Board members, who cannot be employees of the university, are representatives of public life, local government, and state administration." So, no direct involvement from industry in university's governing bodies.

3.5 The Academic-Industrial Ecosystem

Mei-Chih Hu et al (2016) identifies determinants for the performance of the university–industry research collaboration (UIRC) on the example of NTHU (National Tsing Hua University) which is a leading university in Hsin-chu science park being called the Silicon Valley of Taiwan and hosting companies such as Taiwan Semiconductor Manufacturing Company (TSMC). NTHU coordinates the new UAAT-ICU cooperation program with Czechia since 2024 in *Engineering and Technology* which makes the collaboration a potential source of inspiration for Czech universities.

Fig 2 shows a dominating position of electrical engineering together with other engineering + natural sciences over other disciplines in licensing. However, income from licensing is significantly smaller than from direct research contracts, as visible from Fig 3.

Though the numbers of around NT$ 150 M ($ 4.6 M as of current exchange rate)/year may look large, "*most of the research funding comes from the public sector, with the private sector only contributing to roughly 4% of the NTHU's total R&D funding. This proportion is normal for Taiwan's national universities, as all their expenses are funded by the government to meet the national goals of industrial development and strategy chosen and targeted by the state.*" (Mei-Chih Hu et al, 2016).

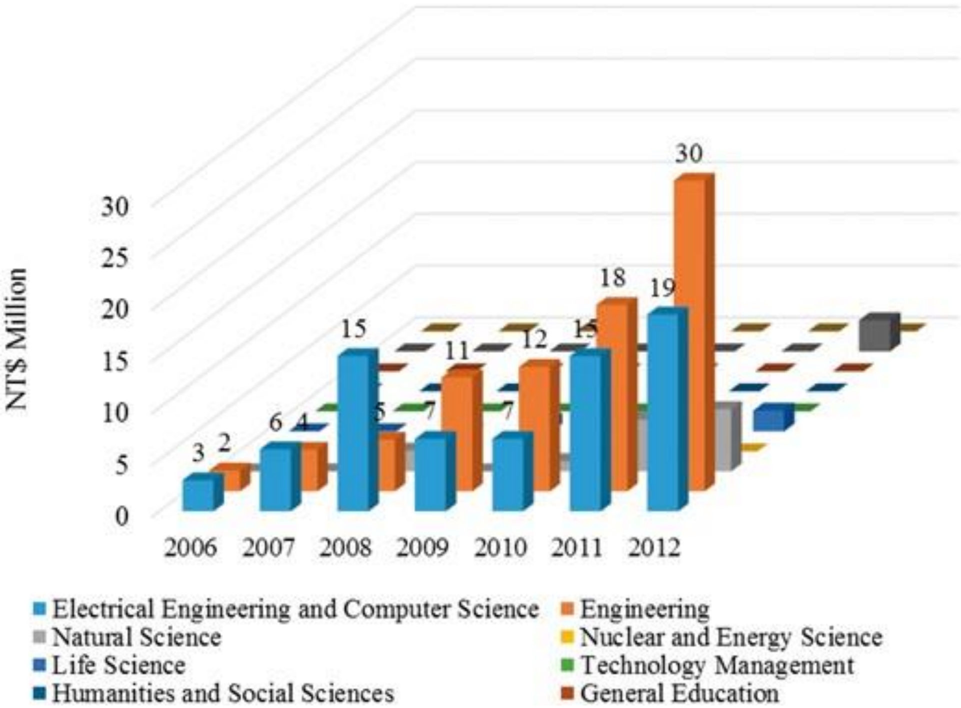


Figure 2. Value of technology licensing activity at the NTHU, by college, 1990–2012.

Source: (Mei-Chih Hu et al., 2016)

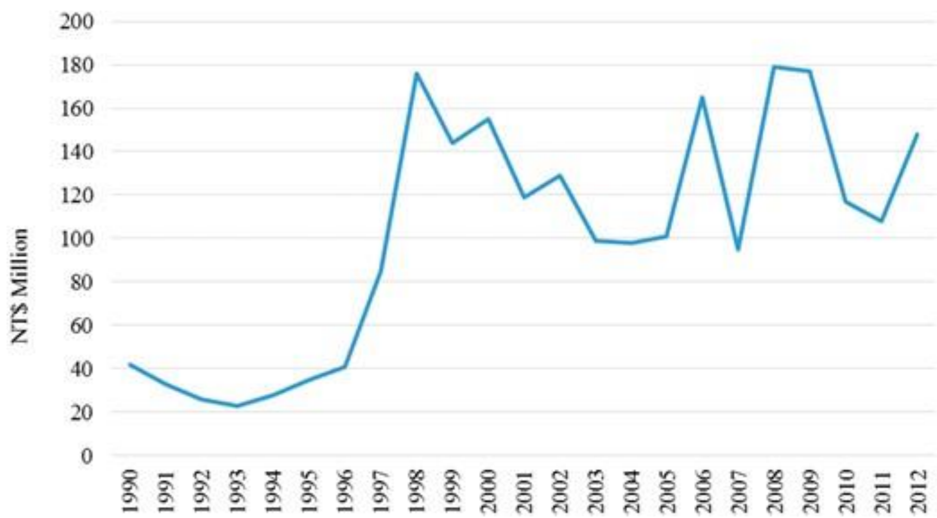


Figure 3. Contract-based R&D collaboration conducted by the NTHU, 1990–2012.

Source: (Mei-Chih Hu et al., 2016)

4. Conclusion

Mei-Chih Hu et al (2016) claims that countries that are newer to technological development ("latecomers") which is the case of both Taiwan and Czechia with Taiwan starting earlier, rely more on the spread of technology, inflow of knowledge, and participation in innovative activities within key industrial sectors to boost their innovation systems and speed up their progress.

In “latecomer” countries, the government plays a crucial role as a *facilitator by developing institutional mechanisms and infrastructure, directing limited resources into selected strategic industries*. (Hu and Mathews, 2005). *Universities in these countries are pivotal, aligning with government policies to enhance national competitiveness* …. (Wong, Ho, and Singh, 2007).

Some measures taken during the last two-three decades in Taiwan might serve as possible inspiration for Czechia:

- Development Plan for World Class Universities and Research Centers of Excellence (2006) defining top evaluation criteria to be *patents, industrial collaboration, licensing,* and so on. It provides extra funding to get into the world class.
- The program on *germination centers* helping technological advancements and enhance global competitiveness in high-tech by early stages *of innovation and entrepreneurship*.
- Fostering an *entrepreneurial culture* by providing strategized industrial services at crucial junctures of a *start-up development* (including *technology valuation, R&D value-added, financial, marketing, and business strategy partnership*).

References

Hu, M. C., & Mathews, J. A. (2005). National innovative capacity in East Asia. Research policy, 34(9), 1322-1349.

IMD - International Institute for Management Development (2024). World Talent Ranking. https://www.imd.org/centers/wcc/world-competitiveness-center/rankings/world-talent-ranking/

Mei-Chih Hu, Shih-Chang Hung, Hsien-Chen Lo, Yung-Ching Tseng. (2016). Determinants of university–industry research collaborations in Taiwan: The case of the National Tsing Hua University, *Research Evaluation,* Volume 25, Issue 2, April 2016, Pages 121–135, https://doi.org/10.1093/reseval/rvw005

Ministry of Economic Affairs of Taiwan (2022). White Paper on Small and Medium Enterprises in Taiwan, 2022, available at: https://www.moeasmea.gov.tw/list-en-2572

National Center on Education and the Economy (NCEE) (2022). Top Performing Countries – Taiwan. https://ncee.org/country/taiwan/

OECD (2022). Education at a Glance 2022, DOI: https://doi.org/10.1787/3197152b-en

Observatory of Economic Complexity (OEC CZE) (2023). Czech Republic Profile. https://oec.world/en/profile/country/cze

Observatory of Economic Complexity (OEC TWN) (2023). Chinese Taipei Profile. https://oec.world/en/profile/country/cze

onsemi (2024). onsemi Selects the Czech Republic to Establish End-to-End Silicon Carbide Production for Advanced Power Semiconductors https://www.onsemi.com/company/news-media/press-announcements/en/onsemi-selects-the-czech-republic-to-establish-end-to-end-silicon-carbide-production-for-advanced-power-semiconductors

Wong, P. K., Ho, Y. P., & Singh, A. (2007). Towards an “entrepreneurial university” model to support knowledge-based economic development: the case of the National University of Singapore. World Development, 35(6), 941-958.

World Bank (2024). World Bank Open Data. https://data.worldbank.org/indicator/GB.XPD.RSDV.GD.ZS

Wikipedia-OECD (2021). List of sovereign states by research and development spending (by OECD 2021 data) https://en.wikipedia.org/wiki/List_of_sovereign_states_by_research_and_development_spending

Overseas Community Affairs Council (OCAC) (2023). R&D spending accounted for almost 4% of Taiwan’s GDP in 2022. https://www.ocac.gov.tw/OCAC/Eng/Pages/Detail.aspx?nodeid=329&pid=58794598

CONTRACT RESEARCH IN ECONOMICS STATISTICS: EXPERIENCE, PITFALLS, PERSPECTIVES

Jakub Fischer, Kristýna Vltavská

Prague University of Economics and Business
Faculty of Informatics and Statistics
fischerj@vse.cz, kristyna.vltavska@vse.cz

DOI: 10.35011/IDIMT-2024-387

Keywords

Contract research, higher education institutions, knowledge transfer

Abstract

The paper aims to describe the role of contract research at selected Czech higher education institutions (HEIs) and to analyse their revenues from contract research as the share on the total revenues in the last five years. There are substantial differences between HEIs, the share varies from 0.14% to 11.66%, according to yearly reports of HEIs. There is no clear trend in the share of contract research in the last five years, with the exception of two HEIs where the share of contract research is generally low. Contract research accounts for the majority of total knowledge transfer income, more than 90% in some HEIs. Secondly, the paper brings the authors' own experience from 30 projects of contract research and comparison to the projects of basic research (topic choice, speed and deadlines, feedback from sponsor, use of the results, popularization, additional costs, cost reporting, awards). Following the experience, the authors mention some perspectives and pitfalls regarding to contract research projects.

1. Introduction

According to Remr (2014, p. 19), contract research is part of the complementary activities of research organisations; contract research was not yet defined by Act No. 130/2002 Coll., on support for research and development from public funds in 2014, but its definition could already be found in the text of the Community Support Framework. Remr (2014), in his work published 10 years ago, refers to the then 2006 version of the Framework (EC, 2006); an updated 2022 version is now available (EC 2022).

Law 130/2002, as currently amended, already defines contract research as research carried out on behalf of an undertaking under a directly applicable European Union regulation, for which a research organisation or research infrastructure is used, whereby the undertaking owns the results of the research activities, bears the risk of failure and provides the research organisation or research infrastructure with a reasonable remuneration for the services received.

According to Remr (2014), in practice, the implementer of contract research (i.e. the research organisation) often designs the research design, constructs the relevant instruments, analyses the data obtained and prepares the final outputs, but each of the steps is always consulted with the sponsor, who approves major decisions regarding the implementation of contract research. We will return to this thesis in our paper.

Contract research has come to the forefront of the debate in universities in the context of the methodologies for evaluating research organisations (RVVI, 2018), which include all public universities and some private ones.

Contract research is part of knowledge transfer. This should be newly explicitly anchored by the Higher Education Act as part of the mission of universities (HEIs, 2024).

The aim of this paper is to quantify the role of contract research at Czech universities of different specializations, to analyze the income from contract research in the last five years, to discuss the experience gained from own contract research in the field of economic statistics in the last 10 years and to describe the opportunities and pitfalls that await researchers in contract research in the field of economics.

First, we quantify the role of contract research at different universities (Chapter 2). Then, in Chapter 3, we provide selected examples of successful contract research projects in the field of economic statistics and summarize the lessons learned, be it the specific role of the contract research sponsor, the collaboration with students, or the possibility of transferring the results of contract research into teaching.

2. Role of the Contract Research at the Czech HEIs

2.1 Data and Methodology

Data on contract research are a mandatory part of the tables in the universities' annual management reports. These tables, in the uniform structure prescribed by the Ministry of Education, Youth and Sports of the Czech Republic (MEYS), also contain data on knowledge transfer. In addition to income from contract research, income from knowledge transfer includes income from licence agreements, income from paid training courses for staff and income from consultancy and advisory activities.

In our analysis, we follow 11 public HEIs of different specializations. The sample includes large universities (Charles University, Masaryk University), technically oriented universities (Czech Technical University in Prague, University of Chemistry and Technology Prague), regional universities with a higher representation of technical fields (VSB-Technical University of Ostrava, University of West Bohemia), regional universities with lower representation of technical fields (Silesian University in Opava, University of South Bohemia, University of Ostrava), non-university HEIs (College of Polytechnics Jihlava) and monothematically focused economic HEI (Prague University of Economics and Business).

For each of these schools, we calculate the share of contract research income in total income and the share of contract research income in knowledge transfer income. This way of comparison does not need to take into account the size of the HEI measured by the number of employees.

The data are obtained from publicly available annual reports of the universities. Data are only available for HEIs as a whole, not for individual faculties. This prevents a more detailed view of the importance of contract research by discipline group.

2.2 Results

First, let's look at the share of contract research revenue in total public university revenue (Table 1). The University of Chemistry and Technology Prague has the highest share, ranging from 7.83% in 2021 to 11.66% in 2019. The University of Chemistry and Technology Prague is a specific university. In addition to its relatively narrow disciplinary focus, it is also characterised by a low number of students compared to similar universities; it is a university with a lower proportion of educational activities compared to creative activities and a social role, partly represented by contract research. VSB-Technical University of Ostrava and Czech Technical University in Prague have the second highest share, around 4% in 2018-2022. Both schools have a strong representation of technical disciplines. Of the more broadly disciplinary universities operating in the regions, contract research is only significantly represented at the University of West Bohemia, most likely by technically oriented faculties (data disaggregated by faculty are not available in the universities' annual management reports). At all other universities, whether regional, non-university, economically oriented, contract research accounts for less than one per cent of total revenues. The same is true for the largest universities (Charles University around a quarter of a percent, Masaryk University around half a percent). In terms of development over time, it is not possible to trace a clear trend at individual universities, except perhaps at the Prague University of Economics and Business, where the share has been increasing very slightly over time, and Silesian University in Opava, which started to implement these activities only in 2021. In absolute terms, it is not without interest that the two largest schools in terms of student numbers, Charles University and Masaryk University, are only 5th and 6th in the ranking of universities in terms of absolute revenues from contract research, with their revenues being about one-fifth that of the much smaller University of Chemistry and Technology Prague.

Table 1. Share of contract research revenue in total public university revenue (%)

	2018	**2019**	**2020**	**2021**	**2022**
University of Chemistry and Technology Prague	8.79	11.66	11.02	7.83	9.87
VSB – Technical University of Ostrava	5.76	3.98	4.11	4.14	4.21
Czech Technical University in Prague	4.30	4.25	3.51	3.90	3.88
University of West Bohemia	2.87	2.50	1.21	2.79	2.99
Masaryk University	0.63	0.68	0.65	0.53	0.43
Prague University of Economics and Business	0.27	0.35	0.32	0.39	0.39
Charles University	0.22	0.24	0.22	0.27	0.31
University of South Bohemia	0.48	0.41	0.50	0.48	0.25
College of Polytechnics Jihlava	0.33	0.55	0.34	0.36	0.23
University of Ostrava	0.15	0.12	0.13	0.27	0.21
Silesian University in Opava	0.02	0.00	0.00	0.06	0.14

Source: author

Table 2. Share of contract research in knowledge transfer (%)

	2018	2019	2020	2021	2022
University of West Bohemia	96.47	97.05	97.05	98.41	96.32
University of South Bohemia	89.65	94.18	94.42	95.41	94.95
VSB – Technical University of Ostrava	93.07	90.72	92.48	89.88	90.33
University of Chemistry and Technology Prague	93.03	88.66	93.52	92.72	87.26
Charles University	75.14	83.51	85.12	79.36	84.17
Czech Technical University in Prague	86.71	88.43	78.73	79.70	80.01
Masaryk University	96.17	95.88	87.62	84.84	79.43
University of Ostrava	61.56	36.53	37.72	69.56	77.02
College of Polytechnics Jihlava	81.97	95.37	100.00	93.75	64.74
Silesian University in Opava	100.00	.	.	36.51	60.47
Prague University of Economics and Business	45.28	46.50	50.86	53.39	51.53

Source: author

The share of contract research in knowledge transfer (Table 2) accounts for more than half of knowledge transfer revenues at most universities, with contract research accounting for more than 75% of total contract research revenues at most of the sampled universities. The exception is the Prague University of Economics and Business, where it is only around 50%, with the other half being roughly equally represented by income from paid training courses for research staff and from consultancy and advisory activities.

3. Own Contract Research: lessons learned (discussion)

At the Department of Economic Statistics, Faculty of Informatics and Statistics, Prague University of Economics and Business, we have been doing contract research for the last 10 years. Before that, we were more intensively involved in basic research (Czech Science Foundation projects), in recent years we have been working on a CSF project from time to time, but the focus has been more on contract research; all the time we have been working on various projects of the Technology Agency of the Czech Republic, in the vast majority as co-recipients. Based on this experience, we would like to describe the key differences between basic research projects (CSF) on the one hand and contract research projects on the other, and to highlight some of the advantages and pitfalls of contract research. Technology Agency projects are usually somewhere in between, with much depending on the specific programme announced and sometimes the specific call.

1. Choice of the topic, determining the expected outcome. While in basic research projects the topic and expected results are specified by the researcher, in contract research this is specified by the sponsor. This can be both an advantage and a disadvantage in terms of contract research: on the one hand a clear brief gives the researcher a clear direction, on the other hand it can be binding and outside the main direction of the researcher's scientific activity.

2. Speed and deadline. The turnaround time for basic research projects is usually calculated in years, for contract research projects it is usually in months. There is also a difference in flexibility, which is considerably higher for basic research projects (in interim reports, any deviations from the timetable are often easier to justify).

3. Feedback from the sponsor. Feedback in the case of basic research projects takes place primarily in the form of comments on the interim and final reports. In contrast, feedback tends to be significant for contract research projects, although in our experience it also varies considerably from one sponsor to another. In some cases, the sponsors are more concerned with the general fulfilment of the assignment and leave the methodology entirely to the project investigator, in other cases the feedback is continuous, very strong and the sponsor continuously consults very specifically on the individual sub-steps of the project.

4. Use of the result. Contract research projects have a much stronger exploitation of the result, the impact on practice is much more obvious than written scientific publications. Knowing that we have helped someone to solve a problem (Fischer & Mazouch, 2022; Fischer et al., 2017; Mazouch et al., 2023) is often more motivating than a "comma" (record), albeit valuable, in the list of publications.

5. PR and media. For basic research projects, any media coverage must be carried out by the researcher himself, with possible support from the PR department of his department. Grant agencies only publicise the results of the best basic research projects. On the other hand, in many cases, the sponsor of contract research promotes the result itself or presents it together with its outputs (Fischer & Mazouch, 2022), or significantly assists the popularisation (Fischer et al., 2019). However, there are also cases where the sponsor considers the results to be strictly confidential and would prefer to keep the collaboration as such secret (here, for obvious reasons, we do not provide a citation).

6. Additional time and financial costs. In connection with point 5, the additional time or financial costs for the presentation of the results, whether it is desired by the sponsor (e.g. in the form of a presentation at a council meeting, a presentation to sponsor’s partners) or arises spontaneously from the demand of various media, should be calculated and included in the costing.

7. Cost reporting. Contract research projects are much looser, the sponsor usually does not need to report the cost structure in any way, with the exception of cases of funding from so-called Innovation Vouchers (Fischer et al., 2017).

8. Awards. The best publications, often resulting from basic research projects, are usually eligible for various external or internal awards (Rector's Award, Dean's Award). For contract research, this changes over time; also in the context of the aforementioned M17+, some schools or faculties are starting to award prizes for societal impact. An important social award for contract research projects is the knowledge that the results are of such value to the sponsor that they are willing to pay for them from their own resources.

9. Linking to teaching. Introducing the results of research activities into teaching, or conversely involving students in the research project, is generally possible in all types of research projects. In our experience, students tend to be more interested in practice projects, seeing that contract research project contributes to something specific and helps someone specific.

5. Conclusion

This paper provides an insight into the role of contract research at HEIs in the Czech Republic. In the first part we analysed the role of contract research at Czech HEIs in terms of the share of contract research income in the total income of the HEI. This varies significantly between HEIs, being negligible in the tens of percentages at some HEIs, and around 10% at others. The role of contract research varies by type of HEI, with large universities with strong research not having a very high share of contract research income. Contract research is the most significant item of knowledge transfer income, exceeding 90% at many HEIs.

In the second part of the paper, we compare the different nature of basic research and contract research projects from a researcher's perspective, outline the advantages and disadvantages of both types of projects, and provide practical advice for engaging in contract research.

Overall, we perceive an untapped potential of contract research in Czech HEIs. The challenge for universities, their management and researchers is to look for ways to increase the share of funds from contract research and thus strengthen the third role of universities. We see it as an opportunity for further research to make international comparisons and seek inspiration from foreign best practice.

Acknowledgement

This paper is prepared under the support of the Institutional Support for the Long-Time Conceptual Development of the Research and Science at the Faculty of Informatics and Statistics, Prague University of Economics and Business.

References

European Commision. (2006). Community Framework for State Aid for Research and Development and Innovation (2006/C 323/01). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52006XC1230(01)

European Commision. (2022). Framework for State aid for research and development and innovation (2022/C 414/01). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022XC1028(03)

Fischer, J., Doseděl, T., & Vltavská, K. (2019). Sociální stratifikace v České republice a polarizace české společnosti [Social stratification in the Czech Republic and polarization of Czech society]. Praha. Asociace samostatných odborů – ASO.

Fischer, J., & Mazouch, P. (2022). Společenství obcí: potenciální úspory z rozsahu, dopad na snížení administrativní zátěže ředitelů a na zvýšení kvality správy škol [Community of municipalities: potential economies of scale, impact on reducing the administrative burden of principals and improving the quality of school administration] [online]. Praha: Vysoká škola ekonomická. https://kest.vse.cz/wp-content/uploads/page/242/studie_FINAL_2-11.pdf.

Fischer, J., Mazouch, P., & Vltavská, K. (2017). Analýza vývoje produktivity a doporučení na zvýšení konkurenceschopnosti společnosti INPEKO spol. s r. o [Analysis of the Productivity Development and Recommendations for Competitiveness Improvement of the INPEKO Company]. Praha: Vysoká škola ekonomická.

Mazouch, P., Hulík, V., & Fischer, J. (2023). Poptávka po vzdělání v mateřských a základních školách v městě Humpolec do roku 2036 [Demand for education in nursery and primary schools in the city of Humpolec until 2036]. Praha: Vysoká škola ekonomická.

Remr, J. (2014). Vybrané právní aspekty postavení a činnosti výzkumných organizací (Selected legal aspects of the position and activities of research organisations). [Dissertation thesis. PF MU]. https://is.muni.cz/th/kphcu/LLM_DIS_06.pdf

RVVI. (2018). Methodology for Evaluating Research Organisations and Research, Development and Innovation Purpose-tied Aid Programmes. https://vyzkum.gov.cz/FrontClanek.aspx?idsekce=799796&ad=1&attid=998025

ÚV ČR. (2024). Návrh zákona, kterým se mění zákon č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů, a další související zákony (Bill amending Act No. 111/1998 Coll., on Higher Education Institutions and on Amendments and Additions to Other Acts (Act on Higher Education Institutions), as amended, and other related acts.)). https://odok.cz/portal/veklep/material/KORNCNUTLVWM/

STAKEHOLDER ENGAGEMENT IN DISASTER MANAGEMENT: PARTICIPATORY DESIGN AND CO-CREATION WITHIN THE PANTHEON PROJECT

Simon Bittner, Ilona Grabmaier, Sarah Kainz
Johanniter Österreich Ausbildung und Forschung gemeinnützige GmbH
simon.bittner@johanniter.at

Kiril Shtefchyk
Institute for Public Security of Catalonia
kiril.shtefchyk@gencat.cat

Cristina Barrado
Computer Architecture Department, Technical University of Catalonia
cristina.barrado@upc.edu

Daoíz Zamora
Mobility Ion Technologies SL (MION)
daoiz.zamora@miontechnologies.com

Satenik Bakunts
Scientific Research Center, Crisis Management State Academy
satenbakunc@yahoo.com

Ioanna Triantafyllou
National and Kapodistrian University of Athens, Department of Geology and Geoenvironment
ioannatriantafyllou@yahoo.gr

John Tsaloukidis
Center for Security Studies (KE.ME.A.)
j.tsaloukidis@kemea-research.gr

Nick Petropoulos
Hellenic Police
npetropoulos@astynomia.gr

PANTHEON

DOI: 10.35011/IDIMT-2024-393

Keywords

DM; participatory design; stakeholder engagement; vulnerable groups; co-creation; PANTHEON

Abstract

This paper deals with participatory design and stakeholder engagement in Disaster Management (DM) and was conducted as part of the EU-funded project PANTHEON. The PANTHEON project aspires to build improved community disaster resilience by developing a community-based, digital platform for DM using Smart City Digital Twin (SCDT) technology. In this paper, we describe the participatory design process that was carried out as part of the project in order to develop design criteria for PANTHEON together with stakeholders (civil protection, military, firefighters, police, emergency medical services, etc.) and potential end users. We also discuss strengths and limitations of such an approach and ultimately argue that participatory design is crucial for bridging the gap between the needs of potential end-users and technology developers. For a participatory design process to succeed, it is indispensable to gather information not only about potential users and stakeholders but also about the specific context in which they are operating, in order to better understand their needs. The combination of a solid methodology, an appropriate and advanced technology, and a well-implemented community participatory process are thus key to an effective DM that ultimately helps to enhance the resilience of communities.

1. Introduction

Our society has become increasingly dependent on technology, and first responders have not been immune to this phenomenon. Indeed, there are many technological innovations that enhance the capabilities of disaster management professionals, enabling them to better manage specific crises. However, there is a significant gap between the needs of potential end-users (civil protection, military, firefighters, police, emergency medical services, etc.) and the products offered by technology providers, often due to a lack of communication or exchange between the two parties. One way to solve this issue is by using *participatory design*, an approach where potential end-users and stakeholders are directly involved in the design process, in order to ensure that the end product is targeted towards their needs and expectations (Interaction Design Foundation, 2024).

This paper describes a participatory design process in the field of Disaster Management (DM), as part of the project PANTHEON. The process addresses the actual needs of users and aims at bringing the two ends of the market closer together. Such a process helps to prevent technology suppliers from developing products that do not add significant value and end up not being used, wasting time and money; it also helps to ensure that the concerns and needs of end users are heard and taken seriously. However, participatory design processes have their limitations and require a thorough analysis of the context in which a system is to be implemented. It is therefore important not only to gather information from stakeholders and potential end-users, but also to study the context in which they operate, in order to understand their needs, possibilities and limitations.

The EU-funded project PANTHEON aims to enhance current disaster resilience approaches and models by developing evidence-based tools for policy and disaster plan analysis and evaluation by using Smart City Digital Twin (SCDT) technology. Digital Twins are virtual representations of physical systems that are updated through a constant information exchange between physical and virtual systems (VanDerHorn & Mahadevan, 2021). The aim of PANTHEON is to create Digital

Twins of the focus areas Athens/Greece and Vienna/Austria including information relevant for DM processes, with the possibility of one day adding data on other regions. The purpose of the study described here was to identify design requirements for the PANTHEON technology using a participatory approach. The aim was to elicit possible applications for a system that could assist relevant experts in disaster response planning, training and implementation using a Digital Twin environment, and to derive and develop concrete design criteria to inform the development of a meaningful product for DM end-users.

This article describes the groundwork that was laid for this participatory design process in previous studies within the PANTHEON project (context analysis), the methodology used, and some recommendations for the PANTHEON technology. We then discuss the selected approach and compare it with other available techniques and tools within the realm of participatory design. The focus of this paper lies on the selected methodology. Detailed results and recommendations for the PANTHEON system can be found in Bittner et al. (2023), which is currently under review by the European Commission.

2. Context Analysis

For a participatory design process to succeed, it is indispensable to gather information not only about potential users and stakeholders, but also about the specific context in which they are operating. In the context of this study, we therefore collected relevant context information such as regulatory DM (Disaster Management) frameworks in the focus areas of the PANTHEON project (Athens and Vienna), potential risks and DM mechanisms, information on vulnerable groups, and recommendations for a participatory governance model. All these aspects were analysed with respect to their implications for specific design recommendations for PANTHEON. The complete outcomes can be found in Bittner et al. (2023) as well as the other project reports referenced in the text.

2.1 Regulatory Framework

Regulations, emergency plans and responsibilities among actors involved shape the composition of DM during all phases of a disaster. In order to understand the links and dependencies within which stakeholders and end-users operate as well as to ensure that the developed system operates within the legality of all pilot regions, the regulatory dimension needs to be closely examined. To facilitate this required understanding among researchers involved and to derive regulation-based design recommendations, a comprehensive literature analysis was done on regulatory frameworks in the realm of DM in Greece/Athens and Austria/Vienna (Tsaloukidis et al., 2023). Since the focus areas are limited to the urban centres Vienna and Athens, cross-border aspects were not taken into account, although this would need to be considered if the system were to be used on a larger scale.

The analysis of regulatory frameworks in the pilot areas informed different development pathways of the project and were used to derive contextual design recommendations for PANTHEON. The main takeaways for these design recommendations were: i) to identify and integrate all relevant emergency management actors, both professionals and local communities; ii) to provide a common and unified channel of communication; and iii) to enable the adaptability of system components to local and regional emergency plans and legislation.

2.2 Risks in the Pilot Areas

For all work in the field of DM, it is essential to have a clear overview and accurate judgement regarding risks affecting a community and their potential effects. In the context of this project it was thus analysed and assessed what implications the identified risks for both pilot areas have in regards to concrete design recommendations for PANTHEON. Triantafyllou et al. (2024) used a risk assessment matrix, based on the aspects “likelihood of occurrence” and “impact from hazard”, to evaluate the risks emanating from different hazards for the focus regions. The assessment methodology included literature research as well as expert interviews and an expert survey as a participatory approach. The most relevant hazards for the region of Vienna were determined to be heat waves, technological accidents, terrorist attacks, and cyber-attacks, with heat waves receiving a particularly high exposure score. Floods were not considered as high risk, because due to its effective flood protection system, Vienna has been safe from river flooding for the last decades (Stadt Wien, MA 45 - Wiener Gewässer, 2017). For Athens, wildfires, heat waves, and earthquakes received the highest scores, followed by cyber-attacks. These hazards (heat waves and cyber-attacks for Vienna, and wildfires and earthquakes for Athens, to cover a diverse set of scenarios) were chosen as pilot cases for the PANTHEON system (see Karamousadakis et al., 2024).

2.3 The Inclusion of Vulnerable Groups in PANTHEON

In Kainz et al. (2024), a vulnerability and capacity assessment was carried out for the focus regions Austria/Vienna and Greece/Athens, using participative methods. This consisted of literature research, interviews, and an online survey where experts on disaster management as well as community representatives and representatives of vulnerable groups were asked for their assessment. This culminated in a list of factors deemed to make people and communities in these regions particularly vulnerable in disaster situations. The list included factors such as age (children and elderly persons), financial resources, language and communication barriers, and cognitive and physical disabilities as well as mobility. These and other factors were considered by the experts to put people at higher risk of being harmed or put at a disadvantage in a disaster situation. Considering and empowering vulnerable groups is an important element in building resilience and reducing vulnerability is one of the main opportunities for disaster risk reduction. For the PANTHEON system, it was decided that vulnerabilities would be considered by including aggregated data on vulnerable groups, such as demographic data and localities such as kindergartens and hospitals, in the Digital Twin, giving a rough overview of the concentration of social vulnerabilities within an area (Bittner et al., 2023). Additionally, it was recommended to use the Digital Twin to inform members of vulnerable groups—children, for example—through workshops, as part of the curriculum or field excursions.

2.4 Participatory Governance Model

To come up with a methodology for a participatory governance model, Bittner et al. (2023) examined various approaches to inform, involve, and mobilize communities in their Decentralized and Participatory Governance report. Their analysis was grounded in a comprehensive review of the literature, two workshops, and an online survey involving stakeholders, community representatives, and academic experts in the field of community engagement. The findings produced were used to contextualize the role of community engagement and to define takeaways for an inclusive system design which also takes especially vulnerable and marginalized people into account. Essential recommendations derived from this part of the context analysis include: to involve affected local communities in the development of disaster response plans when applying simulations via the PANTHEON system, thereby incorporating local knowledge; to identify, include and display local community institutions (sport clubs, cultural associations, etc.) within the SCDT; and to create points of contact to local community institutions for disaster planning and response operations.

3. Methodology of the Participatory Design Process

Based on the groundwork laid in the aforementioned studies, a methodology was developed for a participatory design process for the PANTHEON system by Bittner et al. (2023). The first step was to identify potential stakeholders and end-users which could then be further integrated into the participative process. Both DM stakeholder (Civil Protection Authorities, First Responders and Emergency Services, utilities and infrastructure providers, private companies, media, donors, governmental and policy making authorities) and community and citizen stakeholders (local communities and citizens, NGOs/associations, charities and informal groups) were involved in the process. They can be grouped according to their responsibilities in a disaster situation into administrations (high level and political decisions), first responders (operative and on-site responsibilities), and community representatives (with no responsibility a priori).

As a way of operationalizing the idea and context of the PANTHEON project, a common vocabulary was defined (to minimize misunderstandings) and a concise yet open and undetermined description of PANTHEON’s possibilities was developed (to support purposeful onboarding during participatory design). This way the project’s aims and objectives were kept in focus, while enabling creative and exploratory work interacting with external experts of diverse fields.

To get an overview of existing participatory design methods, a literature review was conducted in which numerous texts dealing with this topic were researched, culminating in a selection of carefully chosen methods used within the context of three workshops (held via Microsoft Teams and Zoom) with stakeholders and end-users that took approximately 2 hours each. Two workshops were held with participants from Austria and one with participants from Greece. The workshops with the Austrian participants involved the attendance of ten experts in total, while the workshop with the Greek experts involved a total of 18 attendees.

The workshops were designed as a group discussion in which experts were asked to talk about a variety of topics the workshop leaders provided, using the following questions as a guide: 1) How can communities become more resilient to disasters? 2) How can the effectiveness of disaster management be increased? 3) How should PANTHEON be designed to enhance community resilience to disasters and support effective disaster management?

In the two Austrian workshops, participants could simultaneously write and comment on an online whiteboard using the interactive browser-based collaboration tool "Conceptboard" (Conceptboard, 2024) which helped to facilitate the group discussion. Due to technical restrictions, participants in the Greek workshop were instead given access to an online form with the guiding questions on it. The following techniques were used in the workshops:

Brainstorming/brainwriting: Using the digital whiteboard (or, in the case of the Greek workshop, the online form), participants individually wrote down ideas and thoughts. Each person then presented those, and the group as a whole discussed them.

Timelines: Together with the participants, a timeline detailing the distinct disaster scenarios of "terror" and "large-scale forest fire" was developed. This timeline was used to show how a disaster develops and to talk about when using PANTHEON would make sense within the disaster management cycle, as well as requirements for the PANTHEON system.

Visioning, following the *Walt Disney method* (adapted and sequential): Participants were asked to imagine themselves as "dreamers," then as “critics” and “realists”.

The workshops were recorded and the resulting data was transcribed and translated into English. All personal information was anonymized, collected data was treated in accordance with the GDPR and participants were informed about and consented to this data processing before the conduction of each workshop. As part of the qualitative content analysis applied, the transcript was structured according to categories identified through open-coding and its scope was reduced by paraphrasing (Mayring, 2002). Finally, the findings were interpreted and translated into a catalogue of design criteria. Although data was collected and risks were assessed separately for the pilot regions, a unified catalogue of design recommendations for the whole PANTHEON system was devised, to ensure that the system could be used for other regions in the future as well.

4. Results

While the actual results of PANTHEON’s participatory design are not the focus of this paper, this section gives a brief insight into the content identified during the process. One of the topics discussed during the workshops was frequently faced issues and problems in disaster management. The participants talked about several different important aspects, and for some of them, recommendations regarding the PANTHEON system (i.e. the technical solution developed in the PANTHEON project) were given. Among others, the attendees mentioned that clear definitions, responsibilities and contact persons are crucial for efficient communication and coordination during disaster management operations, which is trained via cross-organisational exercises and trainings. The participants suggested using the PANTHEON system in the context of these trainings and exercises. Attendees also talked about the “chaos phase” (this expression was used by all Austrian stakeholders), i.e. the first phase of a disaster or an emergency, wherein the situational picture is often still provided by a single individual, which creates a very subjective image of the situation. They suggested using the PANTHEON system, which incorporates drones and sensor data, for creating a realistic real-time situational picture in the chaos phase. It was also proposed to use the system for simulating and modelling CBRNe (Chemical, Biological, Radiological, Nuclear, explosive) events, which could help to define preparedness plans for this hazard.

The participants were asked to specifically identify possible application areas for the PANTHEON system. They identified possible applications for the phases before (prevention/preparedness), during (response), and after (recovery) a disaster, which were later grouped into categories: Before a disaster, the system could be used for *(A) Planning and early warning according to simulations*, and *(B) Training and exercises*. During a disaster, it could be utilized for *(C) Assessing the situational picture* or *(D) Cross-organisational communication*, and after a disaster it’s use could be in *(E) Documentation and Evaluation*. These five application areas were then used as a basis for a catalogue of specific design recommendations for the PANTHEON system, incorporating results from the workshops as well as from earlier results within the project (Bittner et al., 2024; Kainz et al., 2024; Triantafyllou et al., 2024).

The catalogue included general design recommendations for the system, such as ensuring user friendliness and including data on buildings as well as locations of community organisations in the Digital Twin. It also included design recommendations for each of the specific applications listed above. For *Application (A) Planning and early warning* according to simulations, recommendations included that the system should be able to detect on-site spaces according to their size, that the system should suggest a prioritisation of actions to be taken according to calculations, and that local communities should be included in emergency plan development. For *Application (B) Training and exercises,* aspects related to content (e.g. that behavioural or safety instructions should be embedded in specific areas) as well as to the user interface (e.g. that it should be possible to zoom in or out) were mentioned. Recommendations for *Application (C) Situational picture* included for instance that the system should acquire information from various but reliable resources and disseminate accordingly to stakeholders while simultaneously protecting sensitive data. Recommendations concerning the data sources (e.g. satellite data and google maps) and user interface suggestions were also listed. For *Application (D) Cross-organisational communication*, recommendations were similar as for *Application (C),* with a focus on reliable and accurate data transmission and data sources. The recommendations for *Application (E) Documentation and evaluation* comprised similar recommendations as the other application areas, such as using satellite data and remote sensing, allowing the users to zoom in and out, and adding time stamps.

In addition to devising the design catalogue, potential risks, hurdles and issues for the implementation of specific applications of PANTHEON that were discussed during participatory research activities were collected in a table. For instance, one potential risk is that the use of IT-systems may be a liability during response operations, if it fails due to infrastructure damages or network overload. This risk is only relevant for *Applications (C)* and *(D)*.

Finally, a list of relevant data which could be incorporated in the Digital Twin mentioned by the participants during the workshops was drawn up. This list included data such as existing flood risk maps, information on neighbourhood vulnerability indicators such as population density, unemployment rates and social cohesion, and the locations of fire stations (important for instance for wildfire scenarios) and other critical infrastructure.

5. Discussion and Conclusion

The aim of this paper was to describe the participatory design process used within the PANTHEON project. The goal of this participatory approach was to devise concrete design criteria to be used by the technicians within the PANTHEON consortium. These could then be used for the development of a tool with the goal of enhancing community disaster resilience by improving existing disaster management mechanisms. The foundation was laid by a context analysis for identifying relevant hazards, vulnerabilities, and regulatory frameworks within the focus regions of the project, before participatory workshops with DM experts and community and citizen stakeholders were conducted. It culminated in a list of possible application areas for the PANTHEON system, a catalogue with specific design recommendations for each of these possible applications, a list of possible risks, hurdles and problems, and a list of relevant data which should be incorporated into the PANTHEON system's Digital Twin. This process ensured that the system is designed with the relevant issues for the specific regions and the stakeholders and end-users in mind and can be seen in Figure 1.

The specific methods for the participatory design process were chosen after conducting literature research and carefully considering practical limitations such as time constraints, and included brainstorming/brainwriting, timelines, the Walt-Disney Method, and the use of the interactive online whiteboard 'Conceptboard', embedded within three workshops. According to Ryan et al. (2020), workshops are among the most widely used and successful community engagement methods within the realm of disaster preparedness. They offer a personal and collaborative environment and enable the participants and the workshop-leaders to establish a closer relationship and exchange ideas and knowledge. Although some may prefer the personal atmosphere of an in-person workshop, practical considerations may make online workshops more feasible. The field of participatory design offers a myriad of different methods and tools, and researchers should always inform themselves thoroughly of the possible options before choosing those that fit their requirements best.

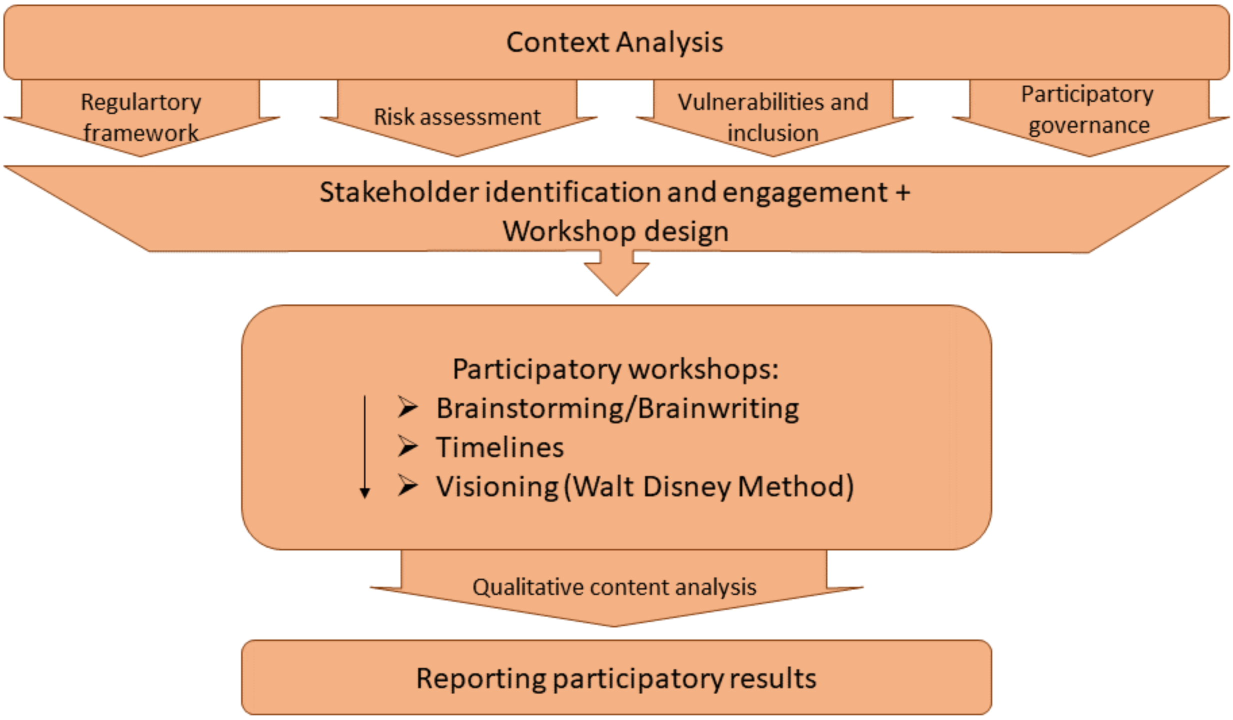


Figure 1. A flow chart showing the methodology used for the PANTHEON participatory design process

In the study presented here, *brainwriting/brainstorming* served as a suitable way to engage stakeholders during the workshop. Participants were given the opportunity to freely express their ideas on what the presented technology could do for their specific area of expertise within DM. Each individual participant was given space to formulate their vision as part of an individual level journaling exercise, before sharing and discussing the collected ideas. However, the lack of goal-orientation of this process limits its implementation to the purpose of exploring possibilities and may face limitations when developing design specifications.

The next method, using *timelines* to identify specific phases of disaster preparedness, response or recovery that the envisioned system would support, helped to put the previous creative process into imagined practice and generated concrete ideas for the application of SCDT as part of DM. However, during the first workshop, empty timelines were used and filled with a co-created scenario on the fly. This approach was very inefficient and for subsequent workshops, coherent scenario timelines were already prepared beforehand, leading to efficient and focused discussions during the workshops.

For the final participatory design method, the concrete ideas generated before were opened up again to formulate the vision of a tool covering the identified features. This was done as part of an adapted version of the *Walt Disney method* (see Schawel & Billing, 2014). Participants were first asked to discuss their vision as *dreamers*, without any resource-related or logistical constraints, before taking the exact opposite stance of *sceptics*, discussing any reasons that might prevent the actual development or minimise the impact of the system. Swinging between the optimistic dreams and sceptical realism of stakeholders and end-users, the workshop hereby revealed both the potential of and the possible obstacles for the PANTHEON system, while generating specific design ideas and recommendations based on participatory knowledge.

The tools that are used should always be adapted according to the specific application areas, end-users and stakeholders, and practical restrictions. While methods like the design charrette (a thorough planning session where developers, designers, and others work together to create a development vision (Tamarck Institute, n.d.)) can for instance be a very useful tool which offer a maximum of participation from the end-user's side, for many potential participants, they may be too work- or time-intense. In these cases, workshops or interviews may be more appropriate. Participatory design should always ensure that the most vulnerable within a society are not overlooked. A potential way to involve especially vulnerable or marginalized groups comes in the form of social network analyses: these can identify isolated groups within a community, which can then be targeted for participative events (Ryan et al., 2020). Ryan et al. (2020) offer a comprehensive literature review on community engagement methods used in disaster preparedness, and the Office for Coastal Management (2015) of the United States of America offers a list focusing particularly on stakeholder engagement.

In general, we recommend to continuously involve stakeholders and end-users in the development process to refine recommendations and requirements for the specific applications targeted with any project. Since users possess the necessary practical understanding, but may lack insight into technical possibilities that technicians and designers can provide, participatory design can cut across the boundaries of different professional sectors and harness the practical knowledge and creative potential of citizens through community participation, rather than treating them as mere consumers (Sanoff, 2011). It is our hope that the participatory design process alongside the requirements and design criteria for an effective disaster management tool described in this paper can inspire other endeavours and help in bridging the gap between different stakeholder groups. The approach used was successful in capturing real end-user needs through a combination of contextual analysis and exploratory workshop methods - a process that can be applied outside the PANTHEON project where diverse stakeholder perspectives are required for ongoing design orientation. In addition, such an approach can ensure an inclusive approach which considers the needs of the most vulnerable within the population. By addressing and taking the actual needs of end users seriously and translating them into concrete design criteria recommendations to be adapted by technology suppliers, we also hope to contribute to an improved disaster management in the focus regions of the PANTHEON project and beyond.

Acknowledgements

The project PANTHEON has received funding from the European Union's Horizon Europe Programme under Grant Agreement No 101074008. We thank Prof. Marc Bonazountas for his constructive internal review of the text.

References

Bittner, S., Grabmaier, I., Kainz, S., Barrado, C., Bagiyan, V., Bakunts, S., & Zamora, D. (2024). D2.5 PARTICIPATORY GOVERNANCE MODEL AND RECOMMENDATIONS (PANTHEON - Community-Based Smart City Digital Twin Platform for Optimised DRM Operations and Enhanced Community Disaster Resilience). Horizon Europe PANTHEON, Grant Agreement N°101074008.

Bittner, S., Grabmaier, I., Kainz, S., Shtefchyk Tatarchuck, K., Barrado, C., Zamora, D., Bakunts, S., Triantafyllou, I., Tsaloukidis, J., & Petropoulos, N. (2023). D3.2 REPORT ON PARTICIPATORY DESIGN PROCESS (PANTHEON - Community-Based Smart City Digital Twin Platform for Optimised DRM Operations and Enhanced Community Disaster Resilience). Horizon Europe PANTHEON, Grant Agreement N°101074008.

Conceptboard. (2024). Conceptboard. https://conceptboard.com/de/

Interaction Design Foundation. (2024). Participatory Design. Interaction Design Foundation. https://www.interaction-design.org/literature/topics/participatory-design

Kainz, S., Geyer, C., Kirilova, S., Grabmaier, I., Apostolopoulou, V., Bakunts, S., Tatarchuck, K. S., Triantafyllou, I., Bagiyan, V., Kazantzidou, D., Tsaloukidis, J., & Valles, L. (2024). D2.3 COMMUNITY VULNERABILITY AND CAPACITY ASSESSMENTS (PANTHEON - Community-Based Smart City Digital Twin Platform for Optimised DRM Operations and Enhanced Community Disaster Resilience). Horizon Europe PANTHEON, Grant Agreement N°101074008.

Karamousadakis, M., Stergiopoulos, G., Schuster, B., & Tsaloukidis, J. (2024). D3.6 USE CASE SCENARIOS (PANTHEON - Community-Based Smart City Digital Twin Platform for Optimised DRM Operations and Enhanced Community Disaster Resilience). Horizon Europe PANTHEON, Grant Agreement N°101074008.

Mayring, P. (2002). Einführung in die qualitative Sozialforschung [Introduction to qualitative social research]. Beltz: Weinheim, Basel.

Office for Coastal Management. (2015). Introduction to Stakeholder Participation. https://coast.noaa.gov/data/digitalcoast/pdf/stakeholder-participation.pdf

Ryan, B., Kim Andrea, J., Taylor, M., & McAndrew, R. (2020). Community engagement for disaster preparedness: A systematic literature review. International Journal of Disaster Risk Reduction, 49, 101655. https://doi.org/10.1016/j.ijdrr.2020.101655

Sanoff, H. (2011). Multiple Views of Participatory Design. Focus, 8(1). https://doi.org/10.15368/focus.2011v8n1.1

Schawel, C., & Billing, F. (2014). Walt-Disney-Methode. In C. Schawel & F. Billing (Eds.), Top 100 Management Tools: Das wichtigste Buch eines Managers Von ABC-Analyse bis Zielvereinbarung (pp. 273–275). Gabler Verlag. https://doi.org/10.1007/978-3-8349-4691-1_91

Stadt Wien, MA 45 - Wiener Gewässer. (2017). Donauhochwasserschutz Wien / Flood Control on the Danube, Vienna (5). Stadt Wien, MA 45 - Wiener Gewässer. https://www.wien.gv.at/umwelt/gewaesser/hochwasserschutz/donau/entlastungsgerinne.html

Tamarck Institute. (n.d.). Index of Community Engagement Techniques. https://www.tamarackcommunity.ca/library/index-of-community-engagement-techniques?gclid=Cj0KCQjw_5unBhCMARIsACZyzS335QQZjmN3yTsT0Cto0QEOq5468nnJKZr5iss9EhvuYAnQhJEolmUaAomcEALw_wcB

Triantafyllou, I., Tsaloukidis, J., & Apostolopoulou, V. (2024). D2.2 MULTI-HAZARDS/RISK DATA AND ASSESSMENT REPORT (PANTHEON - Community-Based Smart City Digital Twin Platform for Optimised DRM Operations and Enhanced Community Disaster Resilience). Horizon Europe PANTHEON, Grant Agreement N°101074008.

Tsaloukidis, J., Kazantzidou-Firtinidou, D., Condomines, J.-P., Mykoniatis, G., Barrado, C., Apostolopoulou, V., Nakos, S., Grabmaier, I., Shtefchyk Tatarchuck, K., Chronopoulos, P., & Bagiyan, V. (2023). D2.1 COMMUNITY BASED DRM ANALYSIS AND REGIONAL ECOSYSTEM (PANTHEON - Community-Based Smart City Digital Twin Platform for Optimised DRM Operations and Enhanced Community Disaster Resilience). Horizon Europe PANTHEON, Grant Agreement N°101074008.

VanDerHorn, E., & Mahadevan, S. (2021). Digital Twin: Generalization, characterization and implementation. Decision Support Systems, 145, 113524. https://doi.org/10.1016/j.dss.2021.113524

ETHICAL INTEGRITY OF RESEARCH IN AI

ETHICAL PERSPEKTIVES AND THE ISSUE OF AUTHORIAL INTEGRITY IN THEUSE OF AI IN CONTEMPORARY ACADEMIC PRACTICE

Anton Lisnik, Milan Majerník, Ján Salanci

Institute of Management
Slovak University of Technology in Bratislava
anton.lisnik@stuba.sk, milan.majernik@stuba.sk, jan.salanci@stuba.sk

DOI: 10.35011/IDIMT-2024-405

Keywords

AI, ethics, integrity of academic activity, data

Abstract

The existence of AI has created indescribable conditions for the collection and use of information in the area of science and academic activity. The very ability of the virtual space to search for information and the ability to compile it into consistent texts opened the way to the use of AI in scientific practice. The very fact of using AI raises questions of ethics, integrity and authorship. The text of the contribution has the ambition to explore the interrelationships between these three points of view on the use of AI in academic practice.

1. Introduction

The growth of neural networks and the use of their ability to learn created prerequisites for the creation of texts, contents and other outputs where data is processed. The use of their ability created the prerequisites for working with the text. Predictive writing was created, which was supposed to make it easier to write texts either in individual editors or in social networks, until it turned into a self-learning entity, which nowadays has many forms and is built on different platforms and is called AI. It generates texts that are also used in academia. The meaning of academic activity consists in the creation of ideas and the investigation of their truth by verification using the principles of the methodology of science.

The integrity of scientific activity in connection with the authorship of texts (sources of information and attribution of authorship) and the originality of the work itself becomes an open question. Another serious problem is the veracity or verification of the sources used in the scientific text in connection with the algorithmization of the weights of individual information in the virtual space. In today's era of booming viral space and the sharing of information in the viral space, we are attacked by information mosaics, which in the final version I evaluate to the objective truth as untruth, i.e. hoax.

2. Introduction to the issue

After the invention of the printing press and then after the publication of the viral space, that is, the Internet, the use of neural networks can be considered a new revolution in working with information. Hellman (Hellman, 2019) refers to ChatGPT as the most sophisticated and publicly available version of the AI tool co-creation. This statement describes two-way communication and two-way influencing of content. The content changes according to the acceptance of the intensity of the mutual weights of the arguments that are found in the most relevant relationships and that will offer the AI as a result of the work. Mutual influence raises the question of the truth of the fact.

These kinds (types) of GPT will suggest sentences that develop the topic relevantly and coherently, leaving the writer only to choose sentences until the task is finished. Some tools, such as OpenAI's Text Completion feature, can even write an entire essay in response to a simple prompt.

Completing a word or word combinations is already a matter of local experience (knowledge). We do not know if they also affect the outer space, or if it is just a local individualized experience of the Neural Network, which is created for a specific user.

The foundation of ChatGPT (OpenAI (https://openai.com/api/) ChatGPT created November 2022) is GPT-3 or Generative Pre-trained Transformer, version 3, a deep learning neural network model used by artificial intelligence (AI) for training computers to generate human-like texts based on data sets from the Internet and other sources (Floridi, Chiriatti, 2020); (Zhang, Li, 2021).

The widespread availability of such tools is a serious problem in education (Godwin-Jones, 2021) because it raises questions about authorship and academic integrity in assessment.

Their use is of concern to education in general, especially as educators strive to equip students with 21st century competencies and values such as critical and inventive thinking, information literacy skills, resilience and integrity. While these tools can take the pain out of the research and writing process, they deprive students of the opportunity to develop academic writing skills and the cognitive, language, and social-emotional competencies that they could acquire through employment and experiencing authentic academic writing processes. Above all, students are denied the opportunity for self-expression and creativity (Yeo, 2023).

The use of texts generated by a neural network threatens the author's integrity in the context of the possibility of checking the originality of the text itself. It depends on which version one uses. Because the cheap free versions are learning and more accurate day by day, but still the paid version will provide a citation service where it is difficult to detect plagiarism. Biermann(2022) in his study: Writers want AI collaborators to respect their personal values and writing strategies: A human-centered perspective on AI co-writing, even states that the autonomy, ownership and integrity of authorship are threatened.

The very work with text AI is very demanding, because it requires quality language skills and also an orientation in expertise, the ability to perform so-called brainstorming. The aspect of using AI in the so-called less developed languages, such as Slovak. But in addition to the fact that AI is improving in these languages, preparation and competence For these reasons, the use of AI creation tools, especially the new generation of co-creation GPT, is an urgent problem in TESOL and requires urgent attention. This is particularly so because the response of educators to the release and subsequent use of ChatGPT has been largely negative, with some institutions reverting to traditional assessment methods such as proctored pen and paper exams (Cassidy, 2023a, 2023b; Yang, 2023).

Still, using such tools can bring a host of benefits to language learners. For example, students have the opportunity to create a focus on the meaning of the written output so that the GPT generates appropriate texts. In addition, the vast amount of generated texts offers a range of meaning-focused, accurate and authentic language models, as well as sources of content knowledge.

3. Questions of an ethical nature in the use of AI in academic practice

Although text generators (transformers) and predictors are capable learning forms of artificial intelligence, they are still the author's tool of the text writer or computer user. Their originality comes from learning from the author of the text and is confirmed by his work and original ideas. The computer only learns what the author teaches it. So it is very difficult to say afterwards that someone (something) created the content.

AI itself is already an absolutely different reality in terms of autonomy, originality and scope. Although it is built on the deep learning ability of the author of the text, he enriches these learned abilities and contents - supplements them with the help of algorithms choosing from many characteristics according to which the author of the text is classified into groups and areas. It affects e.g. group on social networks, intrusions from groups related to work, entertainment, friends. Here we can state that if we chose two AIs and gave them the same task, we would come to a conclusion where there would be various and qualitatively diverse differences.

So which one is true and authentic? Both! The second serious group of questions is the authorship of the text. It is difficult for us to deal with this, because it is either plagiarism, or it is the impossibility of identifying the originator of the source. Meanwhile, there is nothing that should be definitively attributed to the author of the text, who submits it as his work or cited work.

The third important fact is truthfulness. This fact is influenced by the setting of the learning network and the amount of information, which is influenced by the algorithmization of the selection factor according to which the information is selected. It can be e.g. frequency of information, number of views or shares. It can be a kind of rating scale of correctness and value based on belonging to an institution or professional group. But who sets the algorithmization and selection rules? Affiliations and characteristics of the author of the text or the searcher? Belonging to a group? Or will the neural network decide it?

Correct setting of knowledge acquisition and appropriate popularization creates prerequisites for a suitable form of education of respondents and thus also creations of prerequisites for correct decision-making in the professional field. (Katrenčík et al. 2023),(Zátrochová, 2023).

4. Correct knowledge of the principles of AI functioning and the level of knowledge among respondents working and potentially working with AI

From the evaluation of guided interviews, we identified the prerequisites for the research, which took place in January 2024 on a sample of 60 respondents. 60 respondents were interviewed at 2 universities in Bratislava. They were final year students and doctoral students. From the entire battery of questions, we monitored the correctness of the answers in relation to the awareness (knowledge) of the correctness of the answers in connection with the mechanism, the methodology of AI work.

We were interested in the extent to which the respondents' ideas about their knowledge in the given issue coincided with the reality. For the answer a- I know my way around very well, we expected an average rate of agreement (percentage of correct answers) of more than 85%, on the other hand, for the answer c- I don't know my way around, we expected a level of awareness of less than 30%.

H0: In the monitored groups (a, b, c) the level of awareness is the same. Average values are equal. m1=m2=m3

H1: In the monitored groups (a, b, c), the degree of awareness is different. . Average values are not equal. There is at least one pair where: $mi \neq mj$, i.j = 1,2,3

The ANOVA single factor function at the significance level of 0.05 was used for verification. We present the test results in Table 1.

Table 1. Concordance of the respondents' idea of their knowledge in the given issue with the reality (Anova: single Factor).

Groups	Count	Sum	Average	Variance		
a	5	300	60	244,898		
b	51	3342,857	65,54622	136,1345		
c	4	242,8571	60,71429	187,0748		
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	210,284	2	105,142	0,717	0,492	3,158
Within Groups	8347,539	57	146,448	H0		
Total	8557,823	59				

Source: own survey processing.

Based on the values of F and F_{crit}, we recommend accepting the hypothesis H0. A different degree of awareness was not confirmed. The real knowledge of the respondents does not correspond to their own ideas about knowledge in the given issue. An important factor that distorts the given results is the abundance in individual classes a,b,c. Differences in groups are statistically insignificant.

5. Conclusion

Ethical questions are closely related to confusion. Because correctly, I need to have enough correct knowledge to be able to make the right decisions from the field of expertise, at the same time also from the field of morality.

From the conducted research, we found that the main problem is the fact of incorrect awareness of the principles of AI operation. Ethical questions about the use of AI are essentially questions about the nature of academic expertise.

If we wanted to present the old moral rule: in medio stat virtus, we would be left dependent on the rational decision of a human, who would set the rules and teach them to the AI as well. Just who will be the rule maker? Human or AI? Can I determine this today without AI?

The conclusion of the research clearly shows us that it is first necessary to work very intensively on acquiring a lot of correct knowledge about the principles of AI.

Acknowledgement

Article supported by a grant: 023STU-4/2023.

References

Biermann, O. C. (2022). Writers want AI collaborators to respect their personal values and writing strategies: A human-centeredperspective on AI co-writing (doctoral dissertation). University of British Columbia, Vancouver, Canada. DOI: 10.14288/1.0420422

Cassidy, C. (2023a, January 10). Australian schools "flying blind" on use of ChatGPT and other learning technology. The Guardian. https://amp.thegu ardian.com/austr alia-news/

2023/jan/10/austr alia-lagging-behind-on-ai-learning-tech-for-classrooms-report-suggests

Cassidy, C. (2023b, January 10). Australian universities to return to "pen and paper" exams after students caught using AI to write essays. The Guardian. https://www.thegu ardian.com/austr alia-news/2023/jan/10/universities-to-return-to-pen-and-paper-exams-after-students-caught-using-ai-to-write-essays

Floridi, L., Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681–694. https://doi.org/10.1007/s1102 3-020-09548-1

Godwin-Jones, R. (2021). Big data and language learning: Opportunities and challenges. Language Learning, 25(1), 16. http://hdl.handle.net/10125/ 44747

Hellman, M. (2019, April 30). Augmented writing technology: A writer's friend or foe? The Seattle Times. https://www.seatt letim es.com/busin ess/techn ology/ augme nted-writing-technology -a-writers-friend-or-foe/

Katrečík I., Zatrochová M., Mucha B.: Ethical implications of artificial intelligence data usage: a cae staudy of Slovakia and global perspectives. In: IDIMT 2023: New Challenges for ICT and Management - 31st Interdisciplinary Information Management Talks. Pages 365 - 3722023 31st Interdisciplinary Information Management Talks, IDIMT 2023. ISBN: 978-399151176-2. DOI: 10.35011/IDIMT-2023-365.

Yang, M. (2023, January 6). New York City schools ban AI chatbot that writes essays and answers prompts. The Guardian https://www.thegu ardian.com/us-news/2023/jan/06/new-york-city-schools-ban-ai-chatbot-chatgpt

Yeo, M.A. (2023) Academic integrity in the age of Artificial Intelligence (AI) authoring apps. In: TESOL Journal. 2023;14:e716. P. 1-12. DOI: 10.1002/tesj.716

Zátrochová Z., Katrenčík I.: Investment Evaluation Methods for Business Performance. In: Studies in Systems, Decision and Control. Volume 466, Pages 211 – 244. 2023. ISSN: 21984182. DOI: 10.1007/978-3-031-27506-7_9.

DIFFERENCES IN AI UTILIZATION BETWEEN STUDENTS AND TEACHERS IN HIGHER EDUCATION INSTITUTION: AN EMPIRICAL STUDY

Ján Závadský, Zuzana Závadská, Zuzana Osvaldová

Matej Bel University
Institute of Management Systems
jan.zavadsky@umb.sk, zuzana.zavadska@umb.sk, zuzana.osvaldova@umb.sk

DOI: 10.35011/IDIMT-2024-411

Keywords

AI Utilization; Higher Education; Students; Teachers; References

Abstract

This study investigates the utilization of Artificial Intelligence (AI) within the context of higher education, focusing on the divergent patterns of AI adoption among students and teachers at Matej Bel University. Employing a questionnaire as the primary research method, the objectives were twofold: to ascertain the extent of AI usage among the university's students and teachers and to examine the scope of AI references within academic materials. The findings reveal significant differences in AI utilization between students and teachers. This empirical study contributes to the burgeoning discourse on AI ethics in higher education, emphasizing the need for a comprehensive understanding of AI's role in shaping teaching methodologies and learning experiences. Empirical research was carried out on a sample of teachers at Matej Bel University and on a sample of students at the Faculty of Economics. The representativeness of the sample was verified by the chi-square test. The research confirmed greater use of AI by students, but on the contrary, a more ethical approach by teachers.

1. Introduction

Studies in the Literature review underscore the multifaceted nature of AI ethics, spanning from theoretical frameworks to practical applications, and highlight the critical need for ongoing dialogue, research, and policy development to navigate the ethical complexities of AI technologies. The literature review focuses on the burgeoning field of artificial intelligence (AI) ethics, highlighting various dimensions of the discourse from bibliometric analyses to specific challenges in education, healthcare, and broader societal impacts. This review synthesizes the findings from recent research, offering a comprehensive overview of the current state of AI ethics. In the rapidly evolving domain of artificial intelligence (AI), ethical considerations have become increasingly significant. This literature review synthesizes research findings that span various facets of AI ethics, including ethical frameworks, the dual use of generative AI, the ethical management of human-AI interactions, and the privacy and ethical implications in specific sectors like healthcare and digital transformation.

2. Literature review

Graf & Bernardi (2023) embark on a discussion about the balance between ethics, transparency, and advancement within the context of ChatGPT's use in research. This study underlines the intricate balance required to foster innovation while adhering to ethical standards. Slota et al. (2023) explore the dynamics of AI ethics work, emphasizing the role of personal values and professional commitments. Through interviews with stakeholders in AI research, law, and policy, they advocate for incentivizing ethical considerations in AI development and application. Franzke (2022) delves into AI ethics guidelines through a qualitative lens, analyzing 70 documents to understand the promoted values and the translation of ethical concepts into practice. This exploration reveals a tendency towards utilitarian values and a call for greater reflexivity in ethical guidelines. Grinbaum & Adomaitis (2024) discuss the dual-use concerns of generative AI and large language models, suggesting the application of the Dual Use Research of Concern framework to improve governance and societal awareness of generative AI's impact. Lyons et al. (2023) propose a shift from attributing ethics to AI itself towards ensuring the ethical use of AI, highlighting key research areas such as education, transparency, human-centered design, and ethical monitoring. Zhang et al. (2021) offer a bibliometric analysis to map the ethical and privacy concerns in AI, identifying leading contributors and evolving themes, thereby bridging AI techniques with ethical considerations. Majerník et al. (2023) and Lysá et al. (2019) bring selected aspects of quality methods applicable in ethical attitudes in higher education institutions.

Hauer (2020) challenges the division of AI ethics into the creation of ethical standards and the autonomous ethical behavior of AI, critiquing philosophical approaches that hinder the development of morally capable machines. Saurabh et al. (2022) present a framework for AI-led ethical digital transformation, emphasizing the integration of ethics into digital transformation processes and the identification of ethical pillars for guiding this integration. Heyder et al. (2023) review the ethical management of human-AI interaction from a socio-technical perspective, proposing a framework that reconciles duty and virtue ethics in the management of AI systems. Klenk (2024) addresses the manipulation risks of generative AI, advocating for a design-oriented research agenda that encompasses conceptual, empirical, and design considerations to mitigate these risks. Miao (2019) investigates human rights ethics within AI research, proposing management and supervision at every stage of AI's lifecycle to ensure its beneficial impact on humanity. Vakkuri et al. (2021) introduce ECCOLA, a method for implementing ethically aligned AI systems, developed through cyclical action research and collaboration with both researchers and practitioners. Tamburrini (2022) explores the ethical implications of AI's carbon footprint and the distribution of responsibilities among involved actors, suggesting nudging interventions to promote climate-aware AI research. Morley et al. (2020) review publicly available AI ethics tools, methods, and research, aiming to bridge the gap between ethical principles and practical applications in AI development. Mirbabaie et al. (2022) critique the unstructured discourse on AI ethics within information systems research, applying a citation-based technique to identify and analyze fundamental manuscripts in AI ethics. Maouche (2019) discusses the opportunities, risks, and ethical challenges associated with Google AI, highlighting the controversies and ethical debates surrounding emerging technologies and their governance.

Gao et al. (2024) provide a bibliometric analysis tracing the evolution of AI ethics research over two decades, identifying a tripartite progression from initial concepts to the development of human-centric AI systems. They highlight key issues in AI ethics, such as transparency, privacy, and the challenge of superintelligence, while also pointing out research gaps in the large ethics model and AI identification. Complementing this, Chuang, Chang, Chen, Selvamani, and Shia (2022) conduct a worldwide bibliometric analysis, underscoring the global interest in AI ethics and pinpointing the regions and researchers most active in this field. Their work illustrates certain countries' significant contributions and the discourse's evolution over seventy years. The role of explainable AI (XAI) in fostering ethical AI practices is examined by Vainio-Pekka et al. (2023), who stress the lack of a common framework and conceptual clarity within AI ethics and XAI research. They argue for systematic mapping to identify research gaps and empirical studies to advance the field. Bouhouita-Guermech, Gogognon, and Belisle-Pipon (2023) delve into the specific challenges AI poses to research ethics, highlighting issues such as responsibility and transparency. Their scoping review reveals that research ethics boards (REBs) lack the necessary knowledge and tools to evaluate AI research ethically. Hagendorff (2020) critiques the proliferation of AI ethics guidelines, analyzing their overlaps and omissions. His evaluation seeks to bridge the gap between ethical principles and their implementation in AI research. Hallamaa and Kalliokoski (2022) propose enhancing AI ethics by adopting methodologies from fields like systems theory and safety research, arguing that current ethical codes have limited influence due to their abstract nature. The importance of integrating AI ethics into medical education is emphasized by Weidener and Fischer (2024), who propose a principle-based approach to prepare medical professionals for the ethical challenges posed by AI in healthcare. Choi, Yang, and Goo (2024) explore the effects of an AI ethics education program on middle school students, demonstrating significant improvements in ethical awareness and attitudes toward AI. This study underscores the value of early education in AI ethics. This literature review reveals a diverse and evolving field of AI ethics research characterized by a blend of bibliometric analyses, thematic studies, and educational initiatives. The works of Gao et al. (2024), Vainio-Pekka et al. (2023), Bouhouita-Guermech et al. (2023), and others collectively underscore the critical need for systematic frameworks, empirical research, and education in navigating the ethical challenges posed by AI. As AI technologies continue to permeate various aspects of human life, the discourse on AI ethics remains crucial for guiding ethical AI development and implementation. The research was carried out in January and February 2024.

3. Research methodology

This research aims to explore the extent and nature of Artificial Intelligence (AI) utilization among students and teachers at Matej Bel University in Banská Bystrica. Through the examination of research objectives and hypotheses, the study aims to shed light on the dynamics of AI utilization in higher education, particularly focusing on the differences between student and teacher engagements with AI technologies and the ethical dimensions of AI integration into academic practices. Research.

3.1. Research objectives and variables

The first objective seeks to understand the scope of AI usage among the student and teacher populations within the university, examining how these two groups incorporate AI into their academic and educational practices.

The second objective investigates the extent to which AI-generated content is referred to in academic materials by students and teachers, aiming to identify patterns and norms in the citation of AI-produced texts. To guide the inquiry, the study posits two hypotheses.

Hypothesis 1 (H1) suggests that students significantly more utilize AI for writing and creating their academic works compared to their teachers, reflecting a generational or technological gap in the adoption of AI tools for academic purposes.

Hypothesis 2 (H2) proposes that teachers exhibit a more ethical attitude towards the utilization of AI, particularly in how they refer to or cite text generated by AI in their academic materials, indicating a higher level of awareness or concern for ethical considerations in AI usage within educational settings. Research variables are:

1. %TU: percentage of teachers using AI to generate texts out of all teachers in the sample;
2. %TR: percentage of teachers who mention AI in the references out of all teachers;
3. %SU: percentage of students using AI to generate texts out of all students;
4. %SR: percentage of students who mention AI in the references out of all students;
5. %NET: percentage of teachers who do not cite AI of teachers who use AI;
6. %NES: percentage of students who do not cite AI of students who use AI.

3.2. Research sample representativeness

We calculated the representativeness of the sample using Pearson´s chi-squared test ($\chi2$ – test. It tests a null hypothesis, stating that the frequency distribution of certain events observed in a sample (ni) is consistent with a particular theoretical distribution (npi) at the level of statistical significance (α) for appropriate degrees of freedom (k-1), where k is the number of fitted parameters. The frequencies observed, and the expected, (theoretical) frequencies due to teachers` position, are shown in Table 1. Representativeness of the sample due to study degree is shown in Table 2.

Table 1. $\chi2$ - test due to teachers` position.

Teachers` position	np_i		n_i		$(n_i-np_i)^2$	$(n_i-np_i)^2/np_i$
	No.	%	No.	%		
Professor	58	11.37	9	10.47	0.82	0.072
Associate Professor	128	25.10	21	24.42	0.46	0.018
Assistant (PhD.)	294	57.65	51	59.30	2.74	0.048
Assistant	2	0.39	0	0.00	0.15	0.392
Lector	11	2.16	2	2.33	0.03	0.013
Researcher	17	3.33	3	3.49	0.02	0.007
Σ	**510**	100.00	**86**	100.00		**0.551**

Source: (author)

Matej Bel University in Banská has 510 teachers and the research sample consisted of 86 teachers. The questionnaire was anonymous. The $\chi2$ value we achieved is lower than the critical $\chi2$ value at the level of statistical significance $\alpha = 0.05$ for 5 degrees of freedom (6 - 1), which in particular presents the value of 1.14 (value in statistical tables). Since $0.551 < 1.14$, we accept the null hypothesis and we state that the sample file of teachers represents their theoretical distribution.

Table 2. $\chi2$ - test due to study degree (Faculty of Economics)

The study degree	np_i		n_i		$(n_i-np_i)^2$	$(n_i-np_i)^2/np_i$
	No.	%	No.	%		
Bachelor's degree – full-time students	720	54.26	58	48.33	35.10	0.65
Bachelor's degree – part-time students	112	8.44	14	11.67	10.41	1.23
Master's degree – full-time students	406	30.60	36	30.00	0.35	0.01
Master's degree – part-time students	89	6.71	12	10.00	10.84	1.62
Σ	**1327**	100.00	**120**	100.00		**3.51**

Source: (author)

Faculty of Economics of Matej Bel University in Banská has 1327 students (Academic Year 2023/2024; except for mobility students) and the research sample consisted of 120 students. The questionnaire was anonymous. The χ^2 value we achieved is lower than the critical χ^2 value at the level of statistical significance $\alpha = 0.05$ for 3 degrees of freedom (4 - 1), which in particular presents the value of 7.815 (value in statistical tables). Since $3.51 < 7.815$, we accept the null hypothesis and we state that the sample file of students at the Faculty of Economics represents their theoretical distribution.

3.3. Research results

Hypothesis 1 (H1) suggested that students significantly more utilize AI for writing and creating their academic works compared to their teachers. As Figure 1. shows, 69.77% of the teachers in the sample reported using AI to generate texts, denoted as %TU. In contrast, a higher percentage of students, 93.33%, indicated utilizing AI for text generation, represented as %SU. These findings highlight a significant difference in the adoption rates of AI for text generation between the two groups, with students showing a notably higher propensity towards using AI technologies for this purpose. We accept hypothesis H1.

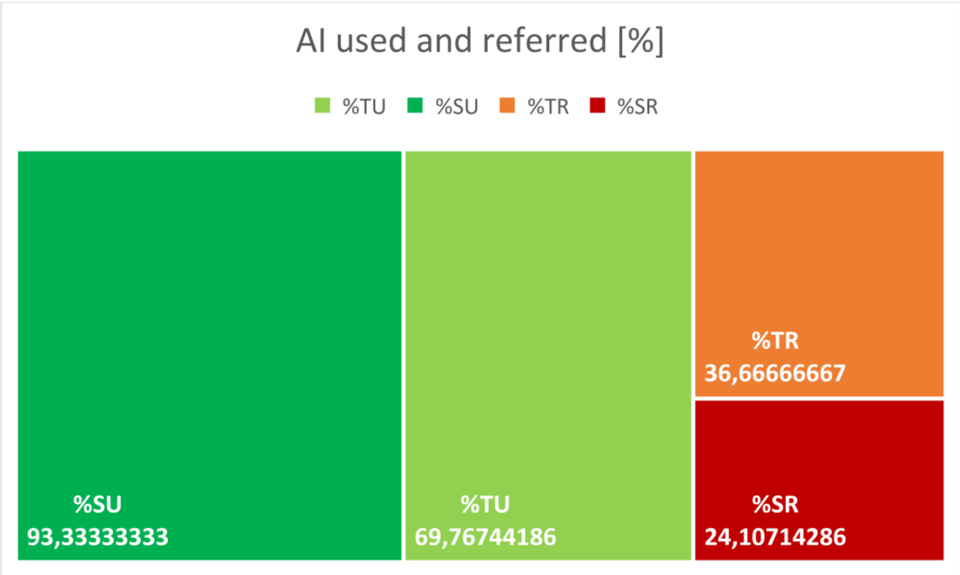


Figure 1. The scope of AI utilization and citation Source: (author)

Hypothesis 2 (H2) proposed that teachers exhibit a more ethical attitude towards the utilization of AI. The results indicate a discrepancy in the citation practices of AI-generated content between teachers and students at Matej Bel University. Specifically, 36.67% of teachers who utilize AI for generating texts acknowledge this by mentioning AI in their references. In contrast, a lower proportion, 24.11%, of students who use AI for text generation do the same. This difference of 12.56% suggests that teachers demonstrate a higher tendency towards ethical practices in AI utilization by more frequently acknowledging the use of AI in their academic outputs. This finding supports the hypothesis that teachers possess a more ethical attitude towards the use of AI, particularly in the correct citation of AI-generated texts. We also accept hypothesis H2.

4. Discussion

The research findings reveal a notable disparity in the citation practices concerning AI-generated content among teachers and students at Matej Bel University. This significant difference underscores a more conscientious approach among teachers regarding the ethical considerations of AI usage in academic publications. The data suggests that teachers exhibit a higher degree of ethical responsibility in their scholarly outputs compared to students, reinforcing the observation that teachers are more ethical in their publication practices than students. Based on the research findings regarding the utilization and citation practices of AI-generated content among teachers and students at Matej Bel University (MBU), several recommendations are proposed to enhance AI ethics in academic settings:

1. Implementing comprehensive training programs for both students and teachers that cover the ethical implications of using AI in academic work. These programs should focus on the importance of transparency and integrity, emphasizing the necessity of citing AI-generated content to maintain academic honesty.
2. MBU should develop and disseminate clear guidelines on how to ethically use and cite AI-generated content in academic works.
3. Guarantors of study programs should integrate AI ethics into existing syllabus to ensure all students and faculty understand the ethical considerations of using AI technologies.
4. Supporting interdisciplinary studies that explore the benefits and challenges of AI use in academic settings, promoting a culture of ethical AI utilization.

By adopting these recommendations, MBU can lead the way in fostering an ethical approach to AI utilization in academic settings, ensuring that all members of the academic community are equipped to use AI technologies responsibly and transparently.

5. Conclusion

The study conducted at Matej Bel University (MBU) in Banská Bystrica aimed to illuminate the dynamics of Artificial Intelligence (AI) utilization within higher education, specifically examining differences in AI use and citation practices between students and teachers. Through a detailed methodology incorporating anonymous questionnaires and statistical analysis, the research assessed the extent of AI-generated content usage and its acknowledgment in academic works. *The findings confirmed both proposed hypotheses: students are significantly more inclined to use AI for text generation, while teachers demonstrate a more ethical approach in citing AI-generated content*. The disparity in AI citation practices revealed through the research underscores a greater ethical consideration among teachers in their scholarly endeavors. A considerable percentage of students, however, neglect to acknowledge AI contributions in their academic outputs, highlighting a critical area for ethical education and policy development at MBU. Given these insights, the conclusion of the paper emphasizes the imperative for MBU to pioneer in fostering ethical AI usage within academic settings. Recommendations include the implementation of comprehensive training programs on AI ethics, the establishment of clear guidelines for ethically using and citing AI-generated content, the integration of AI ethics into curricula, and the encouragement of interdisciplinary research on AI's academic implications. By embracing these strategies, MBU can enhance the ethical standards of AI utilization among its community, ensuring transparency, integrity, and academic honesty in the burgeoning era of AI-driven education.

Acknowledgement

The paper is supported by the KEGA project no. 023STU-4/2023 Teaching financial literacy using information technology as a means of increasing the quality of life and eliminating the effects of the economic crisis.

References

Hu, Y., Wood, J. F., Smith, V., & Westbrook, N. (2004). Friendships through IM: Examining the relationship between instant messaging and intimacy. Journal of Computer-Mediated Communication, 10(1), 38-48.

Tidwell, L. C., & Walther, J. B. (2002). Computer-mediated communication effects ondisclosure, impressions and interpersonal evaluations: Getting to know one another a bit at a time. Human Communication Research, 28(3), 317-348.

Underwood, H., & Findlay, B. (2004). Internet relationships and their impact on primary Computer Interaction, Idea Group Reference, Hershey, London, Melbourne, Singapore, 2006, ISBN 1-59140-562-9

Bolte, L., Vandemeulebroucke, T., & Van Wynsberghe, A. (2022). From an Ethics of Carefulness to an Ethics of Desirability: Going Beyond Current Ethics Approaches to Sustainable AI. Sustainability, 14(8), 4472. https://doi.org/10.3390/su14084472

Bouhouita-Guermech, S., Gogognon, P., & Belisle-Pipon, J.C. (2023). Specific challenges posed by artificial intelligence in research ethics. Frontiers in Artificial Intelligence, 6(1149082). https://doi.org/10.3389/frai.2023.1149082

Choi, J.I., Yang, E.J., & Goo, E.H. (2024). The Effects of an Ethics Education Program on Artificial Intelligence among Middle School Students: Analysis of Perception and Attitude Changes. Applied Sciences-Basel, 14(4), 1588. https://doi.org/10.3390/app14041588

Chuang, C.W., Chang, A., Chen, M.C., Selvamani, M.J.P., & Shia, B.C. (2022). A Worldwide Bibliometric Analysis of Publications on Artificial Intelligence and Ethics in the Past Seven Decades. Sustainability, 14(18), 11125. https://doi.org/10.3390/su141811125

Franzke, A.S. (2022). An exploratory qualitative analysis of AI ethics guidelines. Journal of Information Communication & Ethics in Society, 20(4), 401-423. https://doi.org/10.1108/JICES-12-2020-0125

Graf, A., & Bernardi, R.E. (2023). ChatGPT in Research: Balancing Ethics, Transparency and Advancement. Neuroscience, 515, 71-73. https://doi.org/10.1016/j.neuroscience.2023.02.008

Gao, D.K., Haverly, A., Mittal, S., Wu, J.M., & Chen, J.D. (2024). AI Ethics: A Bibliometric Analysis, Critical Issues, and Key Gaps. International Journal of Business Analytics, 11(1), 338367. https://doi.org/10.4018/IJBAN.338367

Grinbaum, A., & Adomaitis, L. (2024). Dual use concerns of generative AI and large language models. Journal of Responsible Innovation, 11(1), 2304381. https://doi.org/10.1080/23299460.2024.2304381

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines, 30(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8

Hallamaa, J., & Kalliokoski, T. (2022). AI Ethics as Applied Ethics. Frontiers in Computer Science, 4(776837). https://doi.org/10.3389/fcomp.2022.776837

Hanssen, A.B., & Nichele, S. (2019). Ethics of Artificial Intelligence Demarcations. Nordic Artificial Intelligence Research and Development, 1056, 133-142. https://doi.org/10.1007/978-3-030-35664-4_13

Hauer, T. (2020). Machine Ethics, Allostery and Philosophical Anti-Dualism: Will AI Ever Make Ethically Autonomous Decisions? Society, 57(4), 425-433. https://doi.org/10.1007/s12115-020-00506-2

Heyder, T., Passlack, N., & Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. Journal of Strategic Information Systems, 32(3), 101772. https://doi.org/10.1016/j.jsis.2023.101772

Hildt, E., Laas, K., & Sziron, M. (2020). Editorial: Shaping Ethical Futures in Brain-Based and Artificial Intelligence Research. Science and Engineering Ethics, 26(5), 2371-2379. https://doi.org/10.1007/s11948-020-00235-z

Klenk, M. (2024). Ethics of generative AI and manipulation: a design-oriented research agenda. Ethics and Information Technology, 26(1), 9. https://doi.org/10.1007/s10676-024-09745-x

Lysá, L., Jenco, M., Droppa, M., & Lisnik, A. (2019). Entropy - A Tool for Measuring the Quality of Human Potential. Quality-Access to Success, 20(170), 94-99.

Lyons, J.B., Hobbs, K., Rogers, S., & Clouse, S.H. (2023). Responsible (use of) AI. Frontiers in Neuroergonomics, 4(1201777). https://doi.org/10.3389/fnrgo.2023.1201777

Majerník, M., Zatrochová, M., Lisnik, A., & Lysá, L. (2023). Application Of Quality Management in The Production of Glucans in The Food and Pharmaceutical Industry in Slovakia. Quality-Access to Success, 24(193), 238-247. https://doi.org/10.47750/QAS/24.193.27

Maouche, S. (2019). Google AI: Opportunities, Risks, and Ethical Challenges. Contemporary French and Francophone Studies, 23(4), 447-455. https://doi.org/10.1080/17409292.2019.1705012

Miao, Z.Y. (2019). Investigation on human rights ethics in artificial intelligence researches with library literature analysis method. Electronic Library, 37(5), 914-926. https://doi.org/10.1108/EL-04-2019-0089

Mirbabaie, M., Brendel, A.B., & Hofeditz, L. (2022). Ethics and AI in Information Systems Research. Communications of The Association for Information Systems, 50(38), 726-753. https://doi.org/10.17705/1CAIS.05034

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2023). Operationalising AI ethics: barriers, enablers and next steps. AI & Society, 38(1), 411-423. https://doi.org/10.1007/s00146-021-01308-8

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. Science and Engineering Ethics, 26(4), 2141-2168. https://doi.org/10.1007/s11948-019-00165-5

Saheb, T., Saheb, T., & Carpenter, D.O. (2021). Mapping research strands of ethics of artificial intelligence in healthcare: A bibliometric and content analysis. Computers in Biology and Medicine, 135(104660). https://doi.org/10.1016/j.compbiomed.2021.104660

Saurabh, K., Arora, R., Rani, N., Mishra, D., & Ramkumar, M. (2022). AI led ethical digital transformation: framework, research and managerial implications. Journal of Information Communication & Ethics in Societ, 20(2), 229-256. https://doi.org/10.1108/JICES-02-2021-0020

Slota, S.C., Fleischmann, K.R., Greenberg, S., Verma, N., Cummings, B., Li, L., & Shenefiel, C. (2023). Locating the work of artificial intelligence ethics. Journal of The Association for Information Science and Technology, 74(3), 311-322. https://doi.org/10.1002/asi.24638

Tamburrini, G. (2022). The AI Carbon Footprint and Responsibilities of AI Scientists. Philosophies, 7(1), 4. https://doi.org/10.3390/philosophies7010004

Vainio-Pekka, H., Agbese, M.O.O., Jantunen, M., Vakkuri, V., Mikkonen, T., Rousi, R., & Abrahamsson, P. (2023). The Role of Explainable AI in the Research Field of AI Ethics. ACM Transactions on Interactive Intelligent Systems, 13(4), 26. https://doi.org/10.1145/3599974

Vakkuri, V., Kemell, K.K, Jantunen, M., Halme, E., & Abrahamsson, P. (2021). ECCOLA - A method for implementing ethically aligned AI systems. Journal of Systems and Software, 182(111067). https://doi.org/10.1016/j.jss.2021.111067

Weidener, L., & Fischer, M. (2024). Proposing a Principle-Based Approach for Teaching AI Ethics in Medical Education. JMIR Medical Education, 10(e55368). https://doi.org/10.2196/55368

Yu, L.H., & Yu, Z.G. (2023). Qualitative and quantitative analyses of artificial intelligence ethics in education using VOSviewer and CitNetExplorer. Frontiers in Psychology, 14(1061778). https://doi.org/10.3389/fpsyg.2023.1061778

Zhang, Y., Wu, M.J., Tian, G.Y., Zhang, G.Q., & Lu, J. (2021). Ethics and privacy of artificial intelligence: Understandings from bibliometrics. Knowledge-Based Systems, 222(106994). https://doi.org/10.1016/j.knosys.2021.106994

WATERMARK AS A TOOL TO ADDRESS ABUSE OF LARGE-SCALE LANGUAGE MODELS

Miroslav Vacura

Prague University of Economics and Business
Department of Philosophy
vacuram@vse.cz

DOI: 10.35011/IDIMT-2024-419

Keywords

Large language models, LLM, watermark, abuse

Abstract

This paper deals with the problem of identifying texts generated by large language models (LLM) using digital watermarking technology. Advances in artificial neural networks, which are now capable of producing texts comparable to human written speech, have led to an increased risk of misuse of these technologies for ethically questionable purposes such as spreading misinformation or generating academic papers. This paper focuses on the development of methods to incorporate digital watermarks into generated texts, thus enabling their subsequent automatic identification. A suitable watermark should, among other things, be resistant to text editing and easily detectable by special software, while at the same time not being easily removable. Watermarks can take various forms, from the simple insertion of metadata to the use of complex cryptographic methods. This study provides an overview of existing methods, while presenting a new approach that could improve detection capability while keeping computational resource requirements to a minimum.

1. Introduction

Recent advances in the development of complex artificial neural networks focused on language production, called large language models (LLMs), have led to systems that can produce fluent grammatically and semantically error-free text. OpenAI's ChatGPT system beats the Turing test (Biever, 2023), so it can be said to communicate at the level of a native speaker of the language. A number of other systems from other vendors have similar results, such as Claude from Anthropic or Llama from Meta.

However, this progress has also led to the development of risks that texts generated in this way will be used for ethically unacceptable or questionable purposes. Examples of such uses include:

- cheating on academic writing assignments or theses,
- creation of texts containing fake news and disinformation,
- writing spam or phishing emails,
- automated bot accounts on social media (e.g. spreading political propaganda).

Other ethically problematic uses of large language models concern the ability of these systems to generate computer code. These include:

- generating code for academic and other IT-focused theses.
- deliberate creation of computer viruses and other malware,
- the generated code may contain security flaws or vulnerabilities that can be exploited for attacks (if the code is not audited).

In all cases of generated content, whether it is text or code (but also video or images, which we do not deal with here), there are also issues related to copyright, licensing, and generally legal and ethical responsibility for the generated content in a situation where we have no real trustworthy artificial intelligence and no real guidelines (Stamboliev & Christiaens, 2024). If we attribute to large language models the role of advisors in some areas, as suggested, for example, by Špecián and Císař Brown (2024) it is useful to be able to determine which of the advisory texts have been generated by an automatic LLM-based advisor.

Another problem concerning the identification of LLM-generated texts, while not ethical, is that the unrestricted dissemination of these texts impairs the further development of LLM. When collecting the text corpus used for training LLM, it is advisable to use texts produced by human speakers, not artificial texts generated by LLM.

For all these reasons, it is important to have tools to determine whether a particular text was generated by an LLM. One such tool is a digital watermark. In the next section, we first characterize digital watermarking in general and identify what requirements a watermark for large language models should meet. In the third section, we characterize the different types of digital watermarks and compare them to show which is the most appropriate to use in the context of language models. In the next section, we describe the most important current approaches to generate watermarks for LLMs, and the fifth section presents our developed method for watermarking LLMs as a follow-up. The last section of the text presents a conclusion.

2. Digital watermark

The term watermark is usually used to refer to a mark or design element that is embedded in material to identify the author, confirm authenticity or prevent unauthorized copying and distribution of the content. Watermarks can take different forms and can be visible or almost invisible, depending on their purpose. The original watermarking techniques involved physical artefacts and were mainly used as a protective feature of deeds and other tangible documents (Christie-Miller, 2021).

In the environment of information technology and digital media, the term watermark is used for a hidden data pattern that is imperceptible to humans but carries machine extractable information. Digital watermarks are used to protect copyright and prevent unauthorized distribution of digital content such as text documents, photographs, audio recordings, videos or software (Lin & Abdulla, 2015). Digital watermarks can take the form of a simple logo, can contain text with legal information, or more complex cryptographically protected information embedded directly into the data structure of a file so that it is not visible, either because it is not intrusive or easily identifiable. However, these watermarks can be detected and decoded by special software (Shih, 2008).

In the context of LLM, the basic required function of a watermark is to enable the identification of a given text as an LLM product. In the future, it is possible to consider including additional information in such a watermark.

2.1 Watermarking requirements

For the use of watermarking in the context of LLM, it is appropriate that the proposed digital watermarking system meets certain additional requirements (Kirchenbauer et al., 2023). First of all, it would be advisable that the watermark be robust to some degree to minor text editing. Obviously, no system can recognize text as LLM-generated if there is almost complete rewriting of the text, word substitution and modifications to the grammatical structure. However, if minor edits are made, for example only to individual words in some sentences, it would still be advisable to retain the digital watermark. Recognition can then be not binary (yes/no), but can result in some scale, for example indicating that the text was generated 70% by an automated tool.

Another related requirement is that the digital watermark must remain identifiable even if some of the text is removed. That is, its readability should not depend on the presence of the entire generated text block but should be recognizable even from some separated part of such text.

A third obvious requirement is that the digital watermark should not be easily removed from the text, for example by changing just a few specific words in a large volume of text. Obviously, any watermark can be removed, and after rewriting a large portion of the text (as in the first point), it is unlikely to be retained for most methods, but the essential point is that even with knowledge of the digital watermarking algorithm, removal is not easy and possible without major modification of a large portion of the text.

The last requirement is that watermark detection does not require demanding human expert work. The amount of text that is generated is large and it is not feasible for the inspection to depend on limited human resources. Thus, for practical applicability, it is essential that the watermark is algorithmically identifiable and that its application is not significantly computationally intensive, since it is likely to be applied frequently and to large amounts of text in the future as part of standard inspection applications.

As a result, these requirements can be summarized as follows – the first two requirements are based on (Zhao et al., 2023): A digital watermark for LLM should:

- be resistant to text modification (attack robustness),
- its rules should not be easily discoverable (security robustness),
- to be readable from part of the text,
- not be easily removable,
- be algorithmically readable with low computational complexity.

3. Methods of inserting watermarks into texts

Several alternative methods are used for text identification: metadata watermarks, digital image-based watermarks and methods based on stylometric analysis. In the following sections, we evaluate the applicability of these methods for recognizing texts generated by language models.

3.1 Metadata watermarks

A common approach may be to embed special metadata in text files that are in a format that allows the embedding of metadata containing information about the origin of the text (e.g. PDF). This method is straightforward and can be easily implemented, but it is also more easily detected and the metadata can potentially be removed by unauthorized users. Therefore, in such a case, it is advisable to combine metadata embedding with cryptographic protection (locking) of the text document. Although some data formats may not officially support metadata embedding, it is possible to embed suitable metadata, e.g. after the end symbol of the data in the file.

3.2 Digital image-based watermarks

As with images and videos, digital watermarks can be used for digital text. These require the text to be saved in a format that allows additional information to be inserted and stored (e.g. PDF). The watermark can then be added, for example, in the form of graphic elements of imperceptibly small size, imperceptible adjustments to the shape of letters, variations in grey scale, etc. Such information may then be preserved even if printed on a printer with a sufficiently fine resolution.

3.3 Stylometric analysis

The role of a watermark does not have to be only an explicit mark embedded in a digital medium, in this case text. It can also be a specific writing style or grammatical pattern in the text that can be analyzed and compared. A similar method is used to identify the author for historical texts where authorship information is not preserved. In the case of large character models, we can deliberately create such watermarks by adjusting parameters of the language model, training dataset or generation algorithm. The goal is that the generated texts contain specific patterns of words, punctuation or syntactic forms (e.g., Meral et al., 2007) that are unrecognizable to humans but can be detected by specially designed software. This approach still sometimes distinguishes between syntactic, semantic and structural methods. These approaches are generally very challenging and require sophisticated techniques in order not to violate the naturalness and readability of the generated text and thus reduce the quality of the LLM output.

3.4 Evaluation: suitable methods for LLM watermarks

We now evaluate the above methods in terms of their applicability for identifying LLM-generated text through watermarking.

Metadata watermarks are not usable because LLMs usually generate plain text that is then transmitted through one-step copy-paste procedures. Often only fragments of such generated text are worked with, and LLMs are rarely used to produce full text documents that include metadata structures.

Also, digital image-based watermarks are not usable for identifying LLM-generated text for the same reason. The text produced by language models is usually in the format of a plain ascii/unicode data fragment. It is not generated in graphical form, so image-based watermarks cannot be used.

The most appropriate, or only reliably applicable, method is the stylometric approaches, which can be applied to plain text and do not require additional added data structures, either metadata or graphical.

4. Existing watermark designs for LLM

We first briefly describe the principle of language models in simplified form and then discuss some of the existing methods for identifying LLM-generated texts on which our proposed approach is based.

4.1 Large language models

Large language models have a vocabulary V of element count v containing a list of individual words or fragments, which are referred to as tokens $V = \{t_1, \dots t_v\}$. If a language model implemented as chat is given a sentence as an input query (prompt), the first step is to split this sentence into a sequence p of length n tokens: $p = <p_1 , \dots , p_n >$.

The language model for predicting the next word is a function f, implemented by a neural network, which takes as input a sequence of known tokens and its output is a vector v of values, one for each word in the dictionary (typical dictionaries contain more than 50,000 tokens, hence the length of the output vector).

These values are then converted to a normalized probability distribution. From which the most likely following token can be selected (using methods such as standard multinomial sampling, beam search or greedy sampling). This token is added to the input sequence, so the new sequence has $n+1$ tokens and is again given as input to the language model and the process is repeated.

4.2 Existing methods for watermarking LLM-generated texts

The most common watermarking paradigm for LLM-generated text divides the vocabulary into green and red lists (V_G and V_R) and then prefers to select tokens from the green list when generating the text.

An example of this approach is provided by Kirchenbauer et al. (2023) who proposed an algorithm that computes the hash of the last token of an input sequence, uses the result as the seed of a pseudo-random number generator, and labels the tokens in the V dictionary randomly as green and red. This results in two disjunctive sets of tokens V_G and V_R . Then the most probable token from the green set V_G is selected from the probability distribution that is a product of the LLM. Repeating this procedure results in an output sequence consisting of only tokens from the green set V_G . Checking whether the text is generated by the LLM then consists of checking whether the tokens in the given parsed text are randomly from the green and red sets, then the text was not generated, or are from the green set and then the product of the LLM. The disadvantage of this method is low security robustness because the frequency of tokens from the green set is higher than those from the red set. At the same time, attack robustness is also low because the watermark can be invalidated by replacing words with synonyms.

Other authors have proposed a similar approach with various variations. For example, Zhao (2023) randomly divides the vocabulary into green and red sets. Subsequently, the logits (the raw, unnormalized scores that LLM produces before any form of probability distribution is applied) of the language model for the tokens in the green list are increased, while the logits for the tokens in the red list are left unchanged. Machine-generated texts can then be detected based on a higher occurrence of green list tokens than would correspond to a random distribution.

5. Improved method for watermarking LLM-generated texts

In this section, we propose an improved version of digital watermarking for LLM-generated texts by building on the existing approaches of Kirchenbauer et al. (2023) and Zhao (2023).

For a language model L with a vocabulary V of elements v containing a list of tokens, we create a secret signature vector $s = \{s_1 , \dots, s_v\}$. The individual entries of this vector take one of three possible values $s_x \in \{0,1,2\}$. The values 1 and 2 correspond approximately to the split into a green and a red list, but we use a numeric label because neither value is preferred.

The signature vector s thus assigns one of these values to each token t. The starting assignment method is to randomly assign a value of 1 or 2 to each token. For example, the resulting signature vector might look like this $s = \{1, 1, 2, 1, 2, ...\}$, for a dictionary $V = \{$car, nose, wire, walk ....$\}$.

The next stage is tuning the signature vector. In the V dictionary we identify tokens with synonymous meaning, for example we can use any commonly available thesaurus. We then modify the signature vector so that synonymous tokens are always assigned the same value.

Another optimization is that some tokens can be assigned a value of 0, which means they will be ignored. This is particularly useful for very frequently used tokens (conjunctions like "and", etc.).

The essence of our proposed watermark is that the output text generated by the language model L, will consist of a sequence of tokens $o = <o_1, ..., o_m>$ whose associated signature vector values will alternate in a certain pattern, for example: $t_1 = <1, 2, 1, 2, 1, 2, ...>$ or $t_2 = <1, 1, 2, 2, 1, 1, 2, 2, ...>$, or other more complex variants from the list of patterns T. The pattern can be different in different generated texts, randomly chosen from a given list T of regular or mathematically defined patterns, making it impossible to reverse-analyze to determine which tokens are assigned which signature vector values and thus increasing security robustness. Tokens with a value of 0 may also be present in the token sequence, these are then ignored. Thus, $t = <1, 2, 0, 1, 2, 0, 1, 2, ...>$ is interpreted as $t = <1, 2, 1, 2, 1, 2, ...>$.

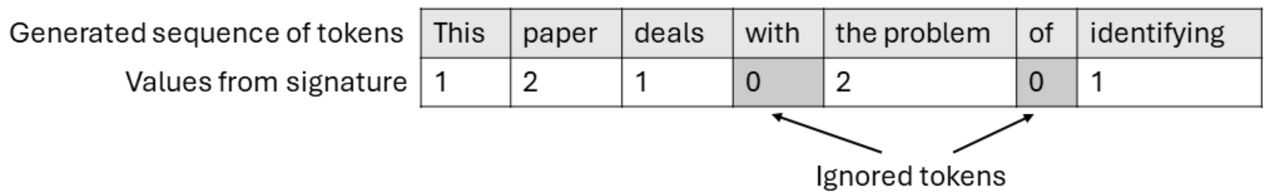


Figure 1. Generating tokens Source: (author)

We will now explain how to machine generate the text to contain the desired watermark. First, we choose a target pattern from a list T, e.g. $t_1 = <1, 2, 1, 2, 1, 2, ...>$. Then we run the language model L to generate the target token based on the prompt $p = <p_1, ..., p_n>$ as described in Section 0. In the first step, the language model generates a normalized probability distribution from which the most likely first output token x is then selected. The signature vector s assigns a value of 1 to this token, so the output corresponds to the target pattern and can be included as the first token of the output sequence $o_1 = x$. The next output token can then be generated.

In the second step, we run the language model L again, specify $<p_1, ..., p_n, o_1>$ as the input vector, and let it generate the target token again. In the second step, the language model again generates a normalized probability distribution, in which the most probable token is y. The signature vector s assigns this token a value of 1, which does not correspond to the target pattern. In the generated probability distribution, the second most likely token is y_2. The signature vector s assigns a value of 2 to this token y_2, which does correspond to the target pattern and can be included as the second token of the output sequence $o_2 = y_2$. We now have two tokens of output sequence $o = <o_1, o_2>$. Subsequently, we can continue with another cycle to generate next output token o_3 in the same manner.

5.1 Assessment of compliance with requirements

In Section 2.1, we listed several requirements that a suitable watermark for LLM-generated text should meet.

- Resistance to text modification – the presented method is resistant to substitution of words for their synonyms after tuning the signature vector to synonymy of tokens.
- Readable even from part of the text – the watermark is readable even from small text fragments.
- Not easy to remove – the watermark is difficult to remove without extensive editing and replacing large sections of text.
- Algorithmic readability with low computational complexity – the watermark is easily detectable, but only using producer's tools, as it requires knowledge of the secret signature vector.

In general, the proposed approach has the desired characteristics.

The presented approach also has advantages over the approaches of Kirchenbauer et al. (2023) and Zhao (2023), these work with green and red sets and the resulting generated text must have a larger proportion of tokens from the green set. In contrast, our approach has no such feature, resulting in higher security robustness. Since our approach works with a fixed signature vector that can be semantically tuned with respect to synonymy, it has higher security robustness and is robust to replacement of words for synonymous terms.

6. Conclusion

The presented text discusses the problem of detection of machine-generated texts using digital watermarking technology. It presents the main existing results and the actual design of the method, which has a number of advantages. A specific feature of the proposed method is the necessity to keep the signature vector secret. Thus, the approach is not applicable to fully open-source language models, but rather to proprietary ones such as ChatGPT from OpenAI or Claude from Anthropic, where the producer can embed a watermark in the generated texts and at the same time offer watermark detection as a service without making the signature vector public. In this respect, our approach belongs to the so-called white-box methods, in contrast to black-box methods, which can be used even in the case of open source or only need API access to LLM models (Tang et al., 2024).

In the future, we plan to create a prototype and test it, which is quite costly in terms of time and money. At the same time, we plan to elaborate in more detail the method of tuning the signature vector with respect to synonymy and other parameters, and to identify a suitable set of tokens to be in the ignored set.

Acknowledgement

The research reported in this paper has been supported by The Czech Science Foundation (GACR) grant n. 24-11697S.

References

Biever, C. (2023). ChatGPT broke the Turing test—The race is on for new ways to assess AI. Nature, 619, 686–689. https://doi.org/10.1038/d41586-023-02361-7

Christie-Miller, I. (2021). Revealing watermarks: How to enhance the security of hand-made paper items and reveal hidden data. Academic Studies Press.

Gehrmann, S., Strobelt, H., & Rush, A. (2019). GLTR: Statistical Detection and Visualization of Generated Text. In M. R. Costa-jussà & E. Alfonseca (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 111–116). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-3019

Hovy, D. (2016). The Enemy in Your Own Camp: How Well Can We Detect Statistically-Generated Fake Reviews – An Adversarial Study (p. 356). https://doi.org/10.18653/v1/P16-2057

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A Watermark for Large Language Models (arXiv:2301.10226). arXiv. https://doi.org/10.48550/arXiv.2301.10226

Lin, Y., & Abdulla, W. H. (2015). Audio watermark: A comprehensive foundation using MATLAB. Springer.

Meral, H. M., Sevinç, E., Ünkar, E., Sankur, B., Özsoy, A. S., & Güngör, T. (2007). Syntactic tools for text watermarking. Security, Steganography, and Watermarking of Multimedia Contents IX, 6505, 339–350. https://doi.org/10.1117/12.708111

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature (arXiv:2301.11305). arXiv. https://doi.org/10.48550/arXiv.2301.11305

Shih, F. Y. (2008). Digital watermarking and steganography: Fundamentals and techniques. CRC Press.

Špecián, P., & Císař Brown, L. (2024). Give the machine a chance, human experts ain't that great…. AI & SOCIETY. https://doi.org/10.1007/s00146-024-01910-6

Stamboliev, E., & Christiaens, T. (2024). How empty is Trustworthy AI? A discourse analysis of the Ethics Guidelines of Trustworthy AI. CRITICAL POLICY STUDIES. https://doi.org/10.1080/19460171.2024.2315431

Tang, R., Chuang, Y.-N., & Hu, X. (2024). The Science of Detecting LLM-Generated Text. Communications of the ACM, 67(4), 50–59. https://doi.org/10.1145/3624725

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending Against Neural Fake News. Advances in Neural Information Processing Systems, 32. https://www.semanticscholar.org/paper/Defending-Against-Neural-Fake-News-Zellers-Holtzman/ad7129af0644dbcafa9aa2f111cb76526ea444a1

Zhao, X., Ananth, P., Li, L., & Wang, Y.-X. (2023). Provable Robust Watermarking for AI-Generated Text (arXiv:2306.17439). arXiv. https://doi.org/10.48550/arXiv.2306.17439

CORPORATE SOCIAL RESPONSIBILITY OF BANKS IN THE CZECH REPUBLIC FOR THE YEAR 2023

Ladislav Luc

Prague University of Economics and Business
Faculty of Informatics and Statistics
Lucl01@vse.cz

DOI: 10.35011/IDIMT-2024-427

Keywords

Corporate social responsibility; human rights; sustainable finance; environmental, social and corporate governance (ESG)

Abstract

This paper updates the status of the application of social responsibility issues in the banking sector of the Czech Republic for the year 2023 and compares it with 2015. Using the recommendations of the internationally recognized ISO 26000 standard, the author's main objective in the field of corporate social responsibility was to process data for the year 2023 and then to compare it with the data for 2015 and to make a basic evaluation based on this comparison. With this post, the author follows up on his earlier analysis, which he conducted in years 2014 - 2015. The topic is complemented by a brief description of the initiative of the Czech Banking Association in the form of the implementation of the Memorandum for Sustainable Finance. This initiative currently illustrates the relevance and importance of the topic of social responsibility in the banking sector. The author used methods of collecting, analysing and comparing relevant information that is available on the topic. For the banks surveyed, this was primarily information published on their websites.

1. Introduction

Over the past decade, much has changed in the area of banks' social commitment. There has been a significant shift in banks' perception of corporate social responsibility towards its active implementation in banking practice. Banks have recognised the importance of applying these principles to all areas of their business, particularly those related to economic, social and environmental impacts. The economic area is characterised by concepts such as rejection of corruption, transparency, good relations with customers, shareholders, business partners or protection of intellectual property. The social sphere can be represented by the concepts of philanthropy, communication with owners (stakeholders), strict respect for human rights and compliance with labour standards. Finally, the environmental area, represented for example by sustainable production, environmental policy and the protection of natural resources. It can be observed with certainty that the active application of all the above-mentioned concepts in practice in accordance with the ethical convictions of the management and employees brings to the banks themselves a long-term sustainable competitive advantage and no less significant stability in the fight in today's highly competitive banking market. The subject of this paper is an analysis of the status of individual CSR topics within the Czech banking sector in 2023 and their comparison with 2015 (Luc, 2015). In terms of time, this is defacto one decade that has passed since the last survey conducted by the author in the past (Luc, 2014).

2. Methodology

For the purposes of this research, the author used the non-binding rules of the international standard ISO 26000 - Guidelines for Corporate Social Responsibility, which he applied to the Czech banking sector. The analysis focused on all banks operating on the Czech market, i.e. 46 entities for the year 2023. For these purposes, the author used a list of individual banks, which is freely available on the Czech National Bank's website (CNB, 2024). The research was further based on the analysis and comparison of the data obtained for the period 2023 and 2015 in the basic seven themes defined by the ISO standard 26000 for social responsibility (ISO 26000, 2010): organizational governance, human rights, labor practices, the environment, fair operating practices, consumer issues, and community involvement and development. Each of the above ISO 26000 themes further includes specific sub-questions, which the author has compared with the actual data and information provided on the banks' CSR websites for the year 2023. If a particular bank presented its commitment to the issue on its website, it received one point for this activity; if not, it received zero. The totals for each sub-question are then converted into percentages. The author compared the results thus obtained for the year 2023 with the previous values of 2015. The structure of the paper always corresponds to one given topic and its division into individual sub-questions, which fully correspond to the structure of the ISO 26000 standard.

3. Research results - the state of corporate social responsibility in czech banks

The following text already fully addresses the conclusions of the analysis carried out in January 2024. Basically, it contains a comparison of the data obtained in the area of application of social responsibility in the banking sector of the Czech Republic for the previous period from 2015 to 2023. The breakdown of the individual topics is based on the structure corresponding to the ISO 26000 standard.

3.1 Organizational governance

The first compared topic was the management of the organization as a decision-making system and the implementation of key decisions in the area of social responsibility.

Compared to 2015, when 27 banks decided to introduce individual topics of social responsibility (it was 27 banks out of 46, in percentage 58.70%), this number significantly increased during the following decade. In the year 2023, all banks (46 out of 46) applied CSR issues, i.e. 100%.

At this point it is necessary to state the fact that the overall portfolio of banks has changed over the years. In total, 10 entities ceased or terminated their active activity on the Czech banking market. 10 banking licences were newly activated and 5 banking institutions changed ownership.

3.2 Human rights

The second area of comparison, in line with the recommendation of the ISO standard 26000, was the issue of human rights, which mainly touches on the following points: due diligence, situations threatening human rights, avoiding complicity, addressing grievances, discrimination and groups at risk, civil and political rights, economic, social and cultural rights, and fundamental rights at work. In this area, the proportion of the subject matter in total CSR has increased from 19.25% in 2015 to 23.95% in 2023.

3.2.1 Due diligence

In the context of social responsibility, due diligence includes the fact that an organisation conscientiously and methodically considers the actual and potential negative impacts of its activities and manages them so as to minimise or eliminate the risk of social or environmental damage, including the non-violation of human rights. In this area during the period under review was achieved one hundred percent exposure of engagement of banks, from 29.63% in 2015 to 100% in 2023.

3.2.2 Human rights risk situation

Situations that threaten human rights are in particular those activities that may have a negative impact on children, addressing a corrupt environment, using supply chains in a lawful manner, ensuring operational security and protecting assets. Similarly, the banks have reached 100% in 2023, up from 18.52% in 2015.

3.2.3 Avoidance of complicity

The theme of avoiding complicity, where the bank must not knowingly facilitate the violation or abuse of human rights, whether for its own benefit or silent complicity, was also fulfilled by the banks on 100%.

3.2.4 Resolving grievances

Complaint handling includes in particular the following key attributes: legitimacy, accessibility, predictability, fairness, compatibility with rights and transparency. Complaint resolution remained a dominant theme for banks, with process improvements taking them to 100% by 2023, up from 51.85% in 2015.

3.2.5 Discrimination and vulnerable groups

Discrimination is when certain individuals or certain groups are singled out to deny them the same treatment or opportunities as others and when this selection is based on prejudice rather than a legitimate reason. Basic attributes of discrimination include race, skin colour, gender, age, nationality or nationality, ethnic or social origin, caste, marital status, sexual orientation, health status such as HIV positivity, political affiliation. In this area, the state of involvement of banks in 2023 showed 100% compared to 55.56% in 2015.

3.2.6 Civil and political rights

In the banking environment, this includes in particular the right to personal security, as well as freedom of opinion and expression, the right to protection against unlawful interference with private life, family, home or correspondence, the right to privacy, access to public services and the right to participate in elections. In the area of addressing civil and political rights, there has been a significant improvement from 25.93% in 2015 to 95.65% in 2023.

3.2.7 Economic, social and cultural rights

This theme of social responsibility in the Bank's environment includes non-discriminatory access to healthcare, education, work, food, religion and culture and real opportunities to participate without discrimination in decision-making. The importance of this topic has increased over the past period from 33.33% in 2015 to 100% in 2023.

3.2.8 Fundamental rights at work

Fundamental rights at work are based on the recommendations of the International Labour Organisation (ILO) and include in particular the following areas: freedom of association and the effective recognition of the right to collective bargaining, the elimination of all forms of forced or involuntary labour, the effective elimination of child labour and the elimination of discrimination at work. The issue of fundamental rights at work reached 100% engagement of banks in 2023 compared to 33.33% in 2015.

3.3 Labour practices

The third area of comparison was banks' basic labor relations practices. This includes all policies and practices relating to work carried out within or on behalf of the Bank. This area includes employment and labour relations, working conditions and social protection, social dialogue, occupational health and safety, and human resource development and workplace training. Proportional representation of the subject on the total CSR for 2015 : 2023, 24.13% ↓ 15.05%.

3.3.1 Employment and employment relationships

The establishment of employment relationships and labour relations are crucial for banks because they address the obligations of both employees and employers and, as such, form the basis for the successful operation and management of the bank. Banks are 100% engaged in this area for 2023 compared to 55.56% in 2015.

3.3.2 Conditions of work and social protection

The scope of this topic is the bank's provision of basic working conditions for its employees (setting working hours, rest periods, holidays, reconciliation of work and family life, etc.), including social protection in cases of occupational injury, illness, maternity, parenthood, old age, unemployment, disability or any other financial hardship. In this area, as in the previous case, the banks' full exposure has been reached, i.e. to 100% in 2023 from 59.26% in 2015.

3.3.3 Social dialogue

Social dialogue between employer and employees within the bank leads to a better understanding of both parties and also creates the basic conditions for the prosperity of the bank as a whole. In the area of social dialogue as well, the banks leveled the bar from 62.96% in 2015 to 100% in 2023.

3.3.4 Health and safety at work

Another important issue for banks in the context of corporate social responsibility is the application of internal processes to ensure the health and safety of employees. A bank's responsible health and safety practices can reduce overall costs, improve morale and increase productivity as well. In this key theme, banks achieved 100% engagement in 2023 compared to 55.56% in 2015.

3.3.5 Human development and training in the workplace

In the area of human resource development and workplace learning, banks are interested in facilitating education, training and lifelong learning for employees. The key benefits for banks in implementing this process are motivation, efficiency, productivity and overall employee performance. Banks are also engaged in this area at 100% from the original 77.78% in 2015.

3.4 The environment

Another topic compared was the environment. Banks are actively involved in environmental protection. Specific themes are pollution prevention, sustainable use of resources, climate change mitigation and adaptation, and protection of the environment, biodiversity and restoration of natural habitats. Proportional representation of the subject on the total CSR for 2015 : 2023, 9.77% ↑ 10.07%.

3.4.1 Prevention of pollution

Banks are particularly active in the areas of their direct impact on the environment, i.e. in the areas of pollution prevention including air emissions, water emissions, solid and liquid waste production and other pollution from their activities, products and services. This area continues to be highly supported by banks in 2023, at 86.96%, up from 40.74% in 2015.

3.4.2 Sustainable resource use

In this area, banks focus on the sustainable use of resources, in particular by using electricity, fuels, raw materials, land and water more responsibly and by combining or replacing non-renewable resources with renewable ones. Sustainable use of resources continues to play an important role for banks, as underlined by the increase in exposure in this area from 33.33% in 2015 to 82.61% in 2023.

3.4.3 Climate change mitigation and adaptation

This topic received significant attention from banks during the decade under review, resulting in a significant increase in exposure from 3.70% in 2015 to 82.61% in 2023.

3.4.4 Protection and restoration of the natural environment

In this area, banks support more socially responsible activities to protect and restore the natural environment and its ecosystems, in particular by restoring and creating natural habitats. The share of banks' exposure increased significantly over the period under review from 68.42% in 2015 to the current 82.61%.

3.5 Fair operating practices

The fifth area compared was the issue of ethics in the business of banks. The key areas are the fight against corruption, responsible political engagement, fair competition, promotion of social responsibility in the value chain, and respect for property rights. Proportional representation of the subject matter on total CSR for 2015 : 2023, 7.47% ↑ 14.72%.

3.5.1 Anti-corruption

Banks focus on identifying corruption risks and implement and maintain policies and procedures to prevent corruption and extortion. Other possible areas include supporting training on corruption for employees, contractors and suppliers. Banks can be commended in the fight against corruption as they achieved 100% engagement for the period under review compared to 2015 which was 29.63%.

3.5.2 Responsible political involvement

Banks are involved in the promotion and development of public policy with the aim of being fully transparent and avoiding political influence. Although banks are generally considered to be non-political entities, they have also achieved a significant improvement in this area from 5.26% in 2015 to 89.13% in 2023.

3.5.3 Fair competition

As part of fair competition, banks must conduct their activities in accordance with competition law and cooperate with the relevant authorities, including taking into account the social context in which they operate (e.g. abuse of poverty to gain undue competitive advantage). In this area, there was a significant strengthening, namely by 70.37% for the period 2015 – 2023, i.e. 100% involvement of banks in the field of fair economic competition.

3.5.4 Promoting social responsibility in the sphere of influence

Banks consider the impact of selling their products and services, including the impact of their market presence. Banks strive to be responsible in the economic chain for compliance with applicable laws, regulations and their own impacts on society and the environment. The impact of engagement in this area has also reached 100% engagement from the original 11.11% in 2015.

3.5.5 Respect for property rights

The right to own property is a human right recognised by the Universal Declaration of Human Rights. These rights include both physical property and intellectual property. Banks have policies and practices in this area that promote respect for property rights and traditional knowledge. In the area of respect for property rights, the banks also achieved 100% exposure from the original 22.22% in 2015.

3.6 Consumer issues

This sixth comparative theme is related to the provision of banking services in the context of consumer health and safety. It covers fair marketing, the provision of factual and non-biased information and fair contracting practices, consumer health and safety, sustainable consumption, customer service, support and resolution of complaints and disputes, protection of personal data and consumer privacy, access to essential services, and education and awareness. Proportional representation of the subject on the total CSR for 2015 : 2023, 14.08% ↑ 20.87%.

3.6.1 Fair marketing, information and contractual practices

Banks provide clear and sufficient information on the prices, conditions and costs of the services they offer (loans, credit and other banking services). In this area, banks created 100% exposure compared to 2015, which was only 3.70%.

3.6.2 Protecting consumers health and safety

Protecting the health and safety of consumers involves providing products that are safe and that do not pose an unacceptable risk to consumers when used or consumed. Here, banks focus on the area of vulnerable groups of potential users of their services. There has again been a significant improvement in this area, up to 100% status in 2023 compared to 22.22% in 2015.

3.6.3 Sustainable consumption

Sustainable consumption is the consumption of products and resources at a level consistent with sustainable development. Sustainable consumption also includes concern for ethical behaviour regarding animal welfare. In this area, banks achieved for the period under review a score of 93.48%, compared to 11.11% since 2015.

3.6.4 Consumer service, support, and dispute resolution

Customer service, support, complaint handling and dispute resolution are mechanisms the bank uses to address consumer needs after products are purchased or delivered. Such mechanisms include warranties and guarantees, technical support regarding use, as well as return, repair and maintenance conditions. According to the results achieved, customer service has become a significant item in the social responsibility of banks, which reached a maximum in this area, i.e. 100% compared to 40.74% in 2015.

3.6.5 Consumer data protection and privacy

Data protection and consumer privacy is designed to protect consumers' privacy rights by limiting the types of information that is collected and the ways in which that information is obtained, used and secured. By definition, this area is very sensitively guarded by the banks. Personal data protection and consumer privacy for 2023 has reached 100% engagement compared to the 2015 figure of 47.37%. The fact that the new European GDPR rules were implemented in 2016 - 2018 has certainly had an impact in this area.

3.6.6 Access to essential services

Banks support projects of various organizations that are aimed at ensuring and respecting the right to basic needs, but also essential services such as the right to electricity, gas, water or even telephone. In terms of access to basic services, banks achieved 100% engagement in 2023 compared to 40.74% in 2015. The energy crisis of 2022-2023 certainly contributed to this result.

3.6.7 Education and awareness

Banks support any education and awareness initiatives. They focus on supporting disadvantaged consumers in rural and urban areas, which in some cases include low-income consumers and consumers with low or no literacy skills. In the area of education and outreach, the banks have achieved 100% engagement from the initial 29.63% in 2015.

3.7 Community involvement and development

The last compared topic of corporate social responsibility, is one of the key issues for banks. Banks are aware of their position within society and also seek to contribute to its development in this space. The cornerstones here are community engagement, education and culture, job creation and upskilling, technological development and access to technology, wealth and income creation, health and social investment. Proportional representation of the subject matter on the total CSR for 2015 : 2023; 25.28% ↓ 15.31%.

3.7.1 Community involvement

For banks, community engagement means not only providing financial assistance but also actively creating various forms of non-profit organisations to support different civic groups. There was a further upward shift in community involvement to 73.91% from 66.67%.

3.7.2 Education and culture

In the field of education and culture, banks are one of the leaders who do their best to ensure all social and economic development. Culture is an important component of a community's life and identity, and the banks are fully aware of this. In this area, the banks have maintained a high level of engagement of 76.09% in 2023 compared to 73.37% in 2015.

3.7.3 Employment creation and skills development

In the area of job creation and skills development, banks seek to leverage their position within the overall financial sector, which also means that they are not only the vehicle for these opportunities, but can also help other organisations to do so. Overall, banks' engagement in this area has increased from 29.63% in 2015 to 69.57% in 2023.

3.7.4 Technology development

The demands of contemporary life are characterised by the use of information and communication technologies, which are also a valuable basis for many other economic activities. In this respect, banks contribute to better access to these technologies through training, various partnerships and other actions. In the area of use and development of information technology, banks have significantly strengthened to 71.74% by the end of 2023 compared to 33.33% in 2015.

3.7.5 Wealth and income creation

Banks are one of the main engines of wealth creation in the economy. They are involved in the design and implementation of business development programmes, along with support in the employment of women. Bank tax compliance is also part of this theme as a necessary condition for governments to generate revenue and use it to address critical development issues. Wealth and revenue creation has also become one of the significant engagements of banks, where from an initial very low position 7.41%, banks have reached a status of 69.57% by 2023.

3.7.6 Health

Health is a fundamental component of society and a recognised human right. Banks also do a lot for this area in the form of contributing to health promotion and prevention, including supporting public health. In this area over the period under review, banks have seen a noticeable upward shift from 48.15% of the 2015 status to a level of 73.91% in 2023.

3.7.7 Social investment

In this area, banks are involved by investing their own financial resources to support infrastructure or other programmes designed to improve social aspects of community life, such as sanitation, safe drinking water, health, housing and food security. Banks' social sentiment did not change much over the decade, but still maintained a decent level at 73.91% compared to 70.37% in 2015.

4. CBA sustainable finance memorandum

The Memorandum was initiated by the Czech Banking Association in 2021. The framework of this document represents the bank's commitment to conduct all its activities in accordance with the rules of sustainable development. Key commitments include, for example, assessing operations through the lens of environmentally and socially responsible business requirements and aligning objectives with those defined by relevant UN global agreements, EU programmes and policies, and national sustainability frameworks; and cooperating with public administration in co-financing projects implemented with the help of EU funds; applying sustainability principles not only in the management of its business activities, but also in its relations with clients, suppliers, shareholders and other partners; cooperating at international and national level through the CBA, especially with the European Banking Federation; and Transparently and regularly reviewing the ESG (Environmental, Social and Corporate Governance) impacts of its business, both positive and negative, and reporting them in accordance with the currently applicable rules (CBA, 2024). Banks are also committed by this Memorandum to the responsible financing of their clients, suppliers, competent authorities and institutions, as well as to the protection of the environment.

By the end of 2023, the CBA had 33 members out of a total of 46 licensed banking subjects. Of these, 14 members of the association have committed to the Memorandum (i.e. 42.42% of the total number of CBA members). With regard to the total number of banks operating on the Czech banking market, only 30.43% of banks have signed up to this initiative.

5. Conclusion

Based on the above comparison of Czech banks' CSR engagement at the end of 2023 compared to 2015 (this is defacto a one-decade period), the following key conclusions can be drawn:

1. in the area of the number of CSR engaged banks in the Czech Republic by the end of 2023, the total ratio of engaged banks in the Czech Republic equaled an incredible 100%,

2. in individual topics of corporate social responsibility, they strengthened their role at the expense of other topics: human rights, consumer issues, community involvement and development and labour practices,
3. in the field of human rights, the banks focused on all relevant topics, the only reserve remains the issue of civil and political rights,
4. in the area of labour relations, banks realise their importance and their engagement has also reached 100%,
5. for the environment, issues of pollution prevention remained a priority, however, according to the results achieved, it can be stated that the other areas remain equivalent issues in the involvement of banks, there are sustainable use of resources, mitigation of climate change and protection of the environment, biological diversity and restoration of natural biotopes,
6. in fair operating practicies, the banks also achieved the maximum score in the issues of fighting corruption, fair economic competition, promoting social responsibility in the value chain and respecting property rights compared to the last theme of responsible political involvement,
7. in consumer issues, all monitored issues become key, except sustainable consumption,
8. and finally, for community involvement and development, the question of education and culture claimed its dominant position, followed equally by community involvement, health and social investment,
9. an interesting phenomenon, which can also be derived from the evaluated data, is the fact that new and smaller banks are profiled mainly in the following topics: human rights, practices in the field of labor relations, business ethics and consumer protection; and long-established banks are more oriented towards the environment and community engagement
10. in the area of joining to the CBA memorandum, there is still open space for banking subjects who operating on the Czech banking market (69.57% has not yet engaged).

In conclusion, it can be stated that the involvement of banks operating on the Czech banking market in the area of corporate social responsibility (CSR) has increased significantly over the last 10 years. It is probably no surprise that all banking entities operating on the Czech banking market are fully devoted to this topic. I dare to predict to the future that with regard to the currently ongoing implementation of the pan-European ESG project, as the next step in strengthening corporate social responsibility, Czech banks (not only) will be more than adequately prepared for this project as well.

References

ISO 26000 (2010), INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. Discovering ISO 26000. Ženeva: ISO, 2010. ISBN 978-92-67-10538-3

Luc, L. (2014), SOCIAL RESPONSIBILITY OF THE BANKS IN THE CZECH REPUBLIC. Systémové přístupy 2014, ISBN 978-80-245-2074-2

Luc, L. (2015), SOCIAL RESPONSIBILITY OF THE BANKS IN THE CZECH REPUBLIC OF THE YEAR 2015, System approaches´15 – Interaction of soft and hard systems, ISBN 978-80-245-2125-1

CNB (2024), Czech National Bank, Regulated institutions and registered financial market entities lists,https://apl.cnb.cz/apljerrsdad/JERRS.WEB15.BASIC_LISTINGS_RESPONSE_3?p_lang=cz&p_DATUM=26.10.2014&p_hie=HI&p_rec_per_page=25&p_ses_idx=1

CBA (2024), Czech Banking Association, https://cbaonline.cz/memorandum-cba-pro-udrzitelne-finance

WORK WITH AI-GENERATED DATA

Miroslav Warhol, Patrik Bretz

Slovak University of Technology in Bratislava
Institute of Management
miroslav.warhol@stuba.sk, patrik.bretz@stuba.sk

DOI: 10.35011/IDIMT-2024-437

Keywords

Artificial, intelligence (AI), ethics, education

Abstract

Using artificial intelligence has seen an extremely large boom in the last decade. Each segment work with data, materials, marketing, machines, give orders. Artificial intelligence brings more efficiency, more possibilities, more creativity to work. However, using artificial intelligence should have its own ethical codes, those who use it should know how to use it and where is its limits. However, working with artificial intelligence can be difficult, and the order submitter must be able to specify exactly what he expects from the artificial intelligence. But event that can change over time, and artificial intelligence will be able correctly understand the thought processes of their users. However, it still needs to enter an enormous amount of data and commands there. It is possible that over time it will be used by all sectors of the national economy, and this is a great space for increasing its using. They will make the work of employees who are *missing in every sector of the national economy and in every region of the world more efficient.*

1. Introduction

Through theoretical and practical expertise, we will look into the nature of artificial intelligence in our work. How is it to work with artificial intelligence like and whether our findings allow us to determine whether or not artificial intelligence could be useful in the work processes.

Based on the questionnaire, we will determine whether the general professional public is interested in artificial intelligence study, education, and the potential to use AI technologies.

This will allow for further evaluation of whether or not artificial intelligence study programmes and general education at schools should be given significant consideration. It's feasible that students' education will equip them with the information necessary to use artificial intelligence more effectively and thus improve their processes at work and make these processes more efficient.

2. Artificial intelligence (AI)

What is Artificial Intelligence? The term "artificial intelligence" is used so frequently these days, that most people shouldn't be unfamiliar with it. Human intelligence is defined in psychology as a collection of various components rather than as a single integrated cognitive process.

AI research currently focuses mainly on the following areas:

- learning = machine learning – searching for patterns and algorithms in data
- reasoning, logical thinking, deduction, finding solutions to problems
- acquiring knowledge and skills
- planning – setting a goal and achieving it
- reading text, processing and understanding language - ability to communicate
- perception and development of the senses - sensory perceptions
- the ability to move and manipulate objects; (Copeland Jack 2000).

Faisal R. Elali and Leena N. Rachid report that the production of AI-generated research and the falsification of work pose serious challenges to the scientific and medical community. The feasibility of producing fabricated works, together with the hard-to-detect nature of published works and the lack of AI detection technologies, creates an opportunistic atmosphere for fraudulent research. The risks of AI-generated research include using said work to change and implement. Reasons for creating research using AI-based technology include financial gain, potential fame, promotion in academia, and resume building, especially for medical students who are in increasingly competitive waters. While AI technologies can be used to streamline secular research processes, they can also be used to pollute the field of scientific research and weaken the legitimate works of other authors. (2023).

3. Research

Survey findings

On the basis of the survey, which we carried out on a sample of 500 business subjects in the Slovak Republic, we came up with interesting findings. A total of 254 subjects responded to the questionnaire. The addressed business entities were from various areas.

We asked about the use of artificial intelligence (AI) in their work, whether they utilize it, and their opinions on its integration with education in this matter. The responses did not surprise us; we expected such a high percentage of respondents who use AI in their work. In the following chapter, we will analyse the questions in detail, along with the results and responses.

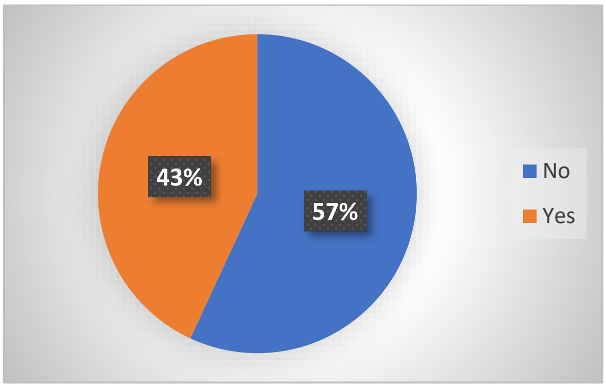


Figure 1. “Do you use artificial intelligence in your work?” Source: (author)

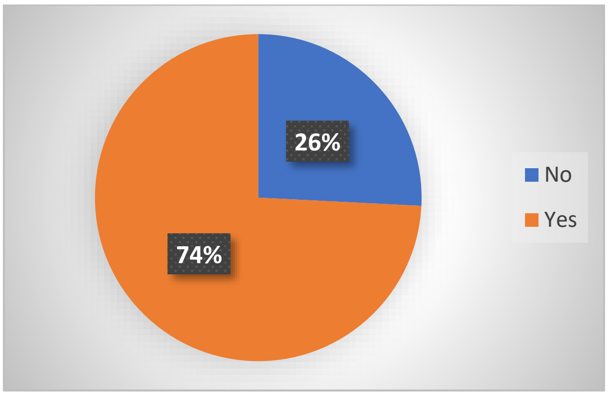


Figure 2. “Do you think that working with AI will become a part of our lives?” Source: (author)

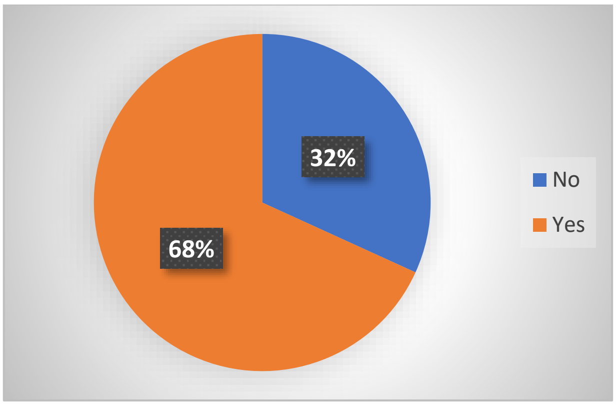


Figure 3. “Would you like to undergo education in the field of AI for working with it?” Source: (author)

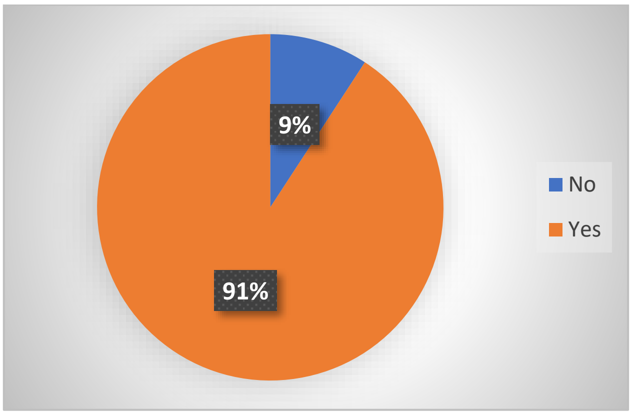


Figure 4. "Do you think it is necessary to expand education and supplement work with artificial intelligence in high schools or universities?" Source: (author)

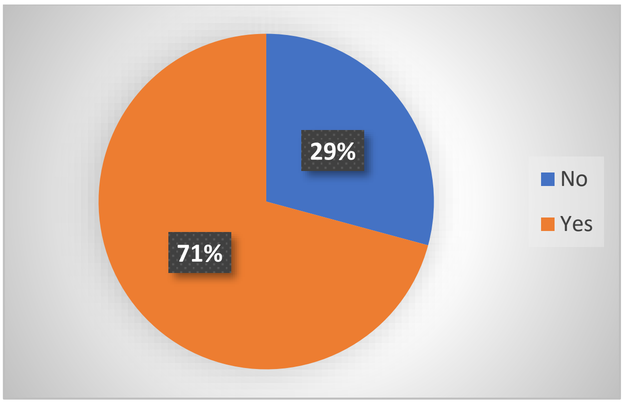


Figure 5. "Would it be beneficial for you to know how to use artificial intelligence in business management?" Source: (author)

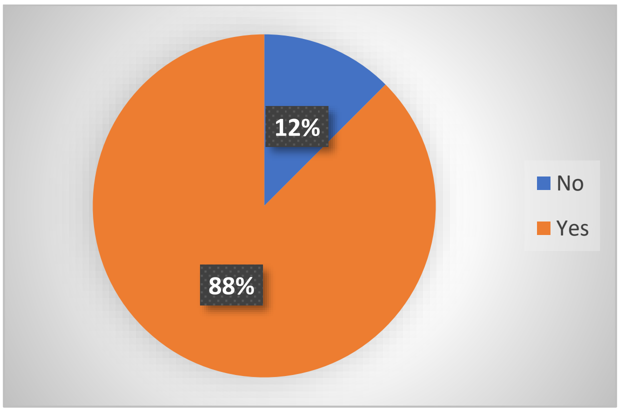


Figure 6. "Do you think that AI has applications in marketing?" Source: (author)

On a relatively simple task, we also tried working with artificial intelligence. Based on the accommodation regulations we prepared for a specific facility, we compared the AI-generated accommodation regulations. The task was defined for a specific facility that can be easily found online. Artificial intelligence created this document based on the task: "Please prepare accommodation regulations for the King's Spring guesthouse to include at least 17 points. The guesthouse is non-smoking, pets are allowed, check-in is possible from 14:00, check-out is by 10:00. Propose additional criteria necessary for the accommodation facility."

Our specifications were correctly applied and comprehended by the task. However, when we gave AI "freedom" to do the assignment in the second half of the task, it did not meet our requirements or our task. Data from the facility's website was used by artificial intelligence. It is a fact that certain data were used that are not included in the material and do not fully meet the role. It can be concluded that it is essential to provide as much detail as possible about the task. The more data we provide AI, the closer it will go to fulfilling its task.

4. Conclusion

These results support the notion that AI can be useful in our day-to-day work. Understanding how to use it correctly is crucial. Our research indicates that AI can make a variety of document creation processes more efficient. This is another argument in favour of schools beginning to emphasise AI-related educational activities. Young individuals of today, who will be entering the workforce, ought to be familiar with the basic principles of artificial intelligence.

By familiarizing students with AI technologies and their applications, they can better navigate the evolving job market and contribute effectively in their future careers.

We would like to leave you with the idea that there might be two potential contrasting scenarios in which the world may shape itself after years of automation. The positive aspect is that machines will liberate us from hard manual labour and tough physical labour, allowing us more time to pursue our creative interests, personal development, and live stress-free lives. On the other hand, the more negative interpretation, which is shared by individuals such as Elon Musk, is that unrelenting automation will eventually take over all jobs and that humans will cease to exist because we will no longer be able to provide services to each other, which lies in mutual assistance. This dichotomy prompts reflection on the ethical considerations surrounding AI and automation, including concerns about economic inequality, job displacement, and the role of humans in a technologically advanced society.

Attachment No. 1 - Accommodation regulations created by Artificial Intelligence

1. Non-smoking guesthouse: We kindly ask guests to refrain from smoking throughout the premises of the guesthouse, including public areas and rooms.
2. Pet-friendly accommodation: We welcome our four-legged companions! Pet-friendly accommodation is permitted, but we kindly ask for compliance with rules regarding responsible behaviour with animals.
3. Check-in from 14:00: Guests can arrive and check in from 14:00 onwards. If you have any special requirements regarding arrival time, please contact us in advance.
4. Check-out by 10:00: On the day of departure, guests are kindly requested to check out by 10:00 to allow us to prepare the rooms for the next guests.
5. Free parking: We provide free parking for guests of the King's Spring guesthouse. Parking spaces are available subject to availability.
6. Wi-Fi connection: Free Wi-Fi connection is available throughout all areas of the guesthouse to keep you connected while on the go.
7. Breakfast Included: Breakfast is included in the accommodation price. We offer a rich selection of meals to help you start your day full of energy.

8. 24-hour reception: Our reception is open 24 hours a day to assist you with any inquiries or requests during your stay.
9. Housekeeping service: Daily room cleaning is included in the accommodation price to ensure you have a pleasant and clean stay.
10. Room air conditioning: All our rooms are equipped with air conditioning to provide you with a comfortable stay during the warm summer months.
11. Private bathroom in each room: Each room has its own bathroom with a shower or bath, toilet, and toiletries.
12. Coffee and tea making facilities: Coffee and tea making facilities are available in the rooms so you can enjoy a refreshing cup during the day.
13. Balcony or terrace with a view: Some rooms offer a balcony or terrace with views of the surrounding nature, where you can relax and unwind.
14. Children's playground: For our younger guests, we have a children's playground where they can have fun and spend time with their peers.
15. Accessibility for people with disabilities: Our accommodation facility is accessible for people with disabilities, with appropriate adaptations and equipment.
16. Restaurant with home-cooking style: In our restaurant, we serve dishes prepared from fresh, local ingredients to bring you the taste of real home feel.
17. Luggage storage: Guests have the option to use our luggage storage in case they need to store their belongings before or after check-out.

Attachment No. 2 - Accommodation regulations created by people

1. Only the guest who has checked-in can be accommodated in the hotel. The guest is required to submit valid ID card, passport or other valid means of identification.
2. During the stay in the pension guest is identified by the pension card obtained at the front Desk at the time of check-in.
3. The room is hold for the guest until the 5 PM if not specified otherwise in the booking.
4. Check-in time is 2 PM if not specified otherwise.
5. Check-out time is 11 AM if not agreed otherwise. If the guest does not vacate the room by the check-out time pension is entitled to charge for the following night.
6. The guest who checks-in before 6 AM will pay for the previous night as well.
7. Pension is bind to provide the guest by the accommodation specified in the confirmed order. If due to the pension the accommodation is not possible in the single or double room and the guest is upgraded, pension will charge only for the room confirmed in the order.
8. Pension is responsible for the guest items brought to the hotel or possible damaged caused to them only if these items were stored at the designated place or where they are usually stored. Pension is responsible for the money and valuables only if they were stored in the pension safety deposit box.
9. The guest is not allowed to store the skiing and other sports equipment in the guest room, only at the designated areas.

10. Visits of the non-guests at the hotel room are possible only after the registration at Front Desk.
11. In case of emergency, injury or illness pension will ensure the medical assistance.
12. On the hotel premises and in the guest rooms it is not allowed to use own electrical appliances except the ones used for the personal hygiene.
13. On the hotel premises and in the guest rooms it is not allowed to leave pets unattended by the owner.
14. From 10 PM until 6 AM the guests are obliged to keep the night silence.
15. The guest is fully liable for any damages on the pension property and is obliged to pay for them in full.
16. The guest is obliged to pay in full amount for all the services ordered and used according to the valid price list at the latest at the time of the check-out.
17. All your claims and complaints or suggestions can be communicated at the Front Desk, to the hotel management or in writing.

Acknowledgement

Supported by the project: 023STU-4/2023.

References

Alameda, T. 2019. The future of AI: can machines think? BBVA. [Online] Banco Bilbao Vizcaya Argentaria, S.A., 09. Jul 2019. https://www.bbva.com/en/the-future-of-ai-can-machines-think/

Boran, M. 2009. Hi, Robot: how the future of robotics means making friends with machines. Independent. [Online] ndependent.ie, January 2009. https://www.independent.ie/business/technology/hi-robot-how-the-future-ofrobotics-means-making-friends-with-machines-26504347.html

Faisal R. E. and Leena N. R. 2023. AI-generated research paper fabrication and plagiarism in the scientific community. https://www.cell.com/patterns/pdf/S2666-3899(23)00043-0.pdf

Khusainov R. V., Lisnik A., Zatrochová M., Babiuk A. M.., Mashkov K. Y. 2023. Economic and legal bases of the Carpathian Euroregion development during the COVID-19 pandemic (Hungary, Slovakia, Poland, Ukraine). [Online]. https://app.crepc.sk/?fn=detailBiblioFormChildOAFA8&sid=ACA353D1F8B027A5CE22D5614721&seo=CREP%C4%8C-detail-%C4%8Cl%C3%A1nok

Lisnik A., Kuperová M. 2023. Ethical Aspects of Working with Data. [Online] 2023. https://app.crepc.sk/?fn=detailBiblioFormChildQAD78&sid=DA442AA97B4AE92CD28E72272932&seo=CREP%C4%8C-detail-kapitola-/-pr%C3%ADspevok

Majerník M., Zatrochová M., Lisnik A., Lysá Ľ. 2023. Application Of Quality Management in The Production of Glucans in The Food and Pharmaceutical Industry in Slovakia. [Online] 2024. https://app.crepc.sk/?fn=detailBiblioFormChildKAEPF&sid=26BAE0713B0B941B8EC626204A87&seo=CREP%C4%8C-detail-%C4%8Cl%C3%A1nok

Martindale, J. 2019. Cortana vs. Siri vs. Google Assistant. Digital Trends. [Online] Designtechnica Corporation, 18. August 2019. https://www.digitaltrends.com/computing/cortana-vs-siri-vs-google-now/.

Maynard, N. 2019. HOW AI CAN REVIVE RETAIL. Jupiter Research. [Online] April 2019. https://www.juniperresearch.com/document-library/white-papers/how-ai-can-reviveretail.

Schwarzmann, M. 2019. IBM pustí kvantový počítač na trh. K technologii se poprvé dostanou běžní zákazníci. E15. [Online] CZECH NEWS CENTER a.s., Januar 2019. https://www.e15.cz/byznys/technologie-a-media/ibm-pusti-kvantovy-pocitac-na-trhk-technologii-se-poprve-dostanou-bezni-zakaznici-1355232.

Štědroň, B. a iní. 2019. Prognostika. Praha: C. H. Beck, 2019. 978-80-7400-746-0. Tate, Karl. 2014. History of A.I.: Artificial Intelligence (Infographic). LiveScience. [Online]

Tuck, J. 2018. Artificial Intelligence: it will kill us. [Video] s.l.: TED-Ed, 2018.

Wired J. 2005. Toys R Kids: High-Tech Playmates. Wired. [Online] Condé Nast, 19. Februar 2005. https://www.wired.com/2005/02/toys-r-kids-high-tech-playmates/. XCON: An Expert Configuration System at Digital Equipment Corporation, Henry Winston, Patrick a A. Prendergas, Karen. 1986. 9780262257220: MIT Press, 1986. ISBN: 978-0262- 257-220.

Závadský J., Závadská Z., Lisnik A. 2022. Design of methodological platform to support the integration of standardized quality management systems applicable in the mining industry. [Online] 2024. Dostupné na: https://app.crepc.sk/?fn=detailBiblioFormChildMAFSA&sid=32E41B49AC27F5A156B0D7A4AA0A&seo=CREP%C4%8C-detail-%C4%8Cl%C3%A1nok

ECO-FRIENDLY AI: BALANCING INNOVATION WITH ENVIRONMENTAL RESPONSIBILITY

Nikola Kuchtíková, Milos Maryska

Prague University of Economics and Business
Faculty of Informatics and Statistics
nikola.kuchtikova@vse.cz, milos.maryska@vse.cz

DOI: 10.35011/IDIMT-2024-445

Keywords

AI, sustainability, environmental impact, energy optimization, public awareness

Abstract

This paper explores strategies for promoting environmental sustainability within artificial intelligence (AI). Through a rigorous methodology that combines research, analysis, and synthesis of current literature and practices, it identifies critical strategies to reduce the environmental impact of AI technologies. These strategies encompass enhancing energy efficiency through optimized algorithms and renewable energy sources, improving data efficiency with data-centric approaches, implementing model compression techniques to lower computational demands, and conducting comprehensive lifecycle assessments to evaluate environmental footprints. Additionally, this paper emphasizes the importance of ethical AI principles, fostering collaboration and knowledge sharing among stakeholders, establishing robust regulatory measures, and raising public awareness and education. Each strategy is elaborated with practical examples and evidence-based insights to guide AI developers, researchers, policymakers, and other stakeholders in adopting sustainable practices. By highlighting the critical intersection of AI and environmental sustainability, this paper seeks to raise awareness, provide actionable solutions, and encourage proactive steps toward building a more sustainable future in AI development and deployment.

1. Introduction

Artificial intelligence (AI) revolutionizes sectors, boosting innovation and efficiency, yet its growth brings environmental costs like high energy use and increased carbon footprint. As AI scales, addressing these concerns is vital to ensure its positive impact without worsening ecological issues.

This paper examines the intersection of AI and environmental sustainability, identifying key strategies to mitigate the ecological footprint of AI technologies. Through a comprehensive approach, it explores several methods for creating a greener AI ecosystem. By examining these strategies, the paper aims to raise awareness and provide actionable solutions for fostering sustainability in AI. It seeks to guide developers, researchers, policymakers, and stakeholders toward responsible and environmentally conscious AI practices, ultimately contributing to a more sustainable future in AI development and deployment.

2. Methodology

The methodology for writing this paper involved a combination of research, analysis, and synthesis of information. The methodology is aimed to provide readers with a comprehensive overview of the strategies for promoting environmental sustainability in AI, supported by evidence-based research and analysis. Here is an overview of the methodology in four steps:

1. **Research:** Extensive research was conducted to understand the intersection of AI and environmental sustainability. This involved studying academic literature, industry reports, news articles, and reputable online sources to gather information on current trends, challenges, and best practices in sustainable AI development.

2. **Identification of Key Strategies:** Based on the research findings, critical strategies for making AI more sustainable were identified. These strategies encompassed various aspects such as energy efficiency, data efficiency, model compression, hardware innovation, renewable energy, lifecycle assessment, ethical AI principles, collaboration, regulatory measures, and public awareness.

3. **Analysis and Evaluation:** Each identified strategy was analyzed regarding its relevance, effectiveness, and potential impact on promoting environmental sustainability in AI. Consideration was given to both the technical feasibility and practical implications of implementing these strategies.

4. **Synthesis and Organization:** The information gathered from research and analysis was synthesized to develop coherent and structured content for this paper. Each strategy is elaborated upon, providing explanations, examples, and insights to help readers understand its significance and practical implications.

3. State of the art

In recent years, integrating artificial intelligence (AI) technologies with environmental sustainability efforts has gained significant traction, offering innovative solutions to address pressing ecological challenges. From climate change mitigation to biodiversity conservation, AI is revolutionizing how we monitor, manage, and protect the environment. Here is an overview of key developments and applications in this rapidly evolving field:

- **Environmental Monitoring and Conservation:** AI-driven systems enable real-time data analysis for monitoring changes in land cover, biodiversity, and habitat loss, supporting conservation and restoration initiatives (Onyebuchi et al., 2024)
- **Climate Change Modeling and Adaptation:** AI techniques enhance climate modelling accuracy, informing adaptive strategies to mitigate risks and build resilience in vulnerable communities (Jain et al., 2023).
- **Renewable Energy Optimization:** AI optimizes the generation, distribution, and consumption of renewable energy resources, improving grid efficiency and enabling better resource forecasting (Nzubechukwu et al., 2023)
- **Sustainable Agriculture and Food Systems:** AI-powered precision agriculture tools optimize resource use, increase crop yields, and promote sustainable farming practices for food security and environmental conservation (Bhagat et al., 2022)

- **Conservation and Wildlife Protection:** AI aids in species identification, habitat monitoring, and anti-poaching efforts, supporting conservation initiatives and protecting endangered species (Tuia et al., 2022).

Challenges include data quality, algorithm bias, and ethical considerations, but collaboration between AI researchers, environmental scientists, and policymakers drives innovation and responsible deployment for a more sustainable future.

4. Results

Making AI sustainable involves considering relevance, effectiveness, and potential impact. Now, we will highlight some strategies to make AI more sustainable.

4.1 Energy efficiency and the use of renewable sources

Energy efficiency is critical in developing and deploying artificial intelligence (AI) technologies, with significant implications for environmental sustainability. AI developers can reduce energy consumption, minimize carbon emissions, and mitigate the environmental impact by optimizing algorithms to create leaner AI models and reduce computational needs for training and interference (Omar, 2023), as well as hardware innovations, to enable faster, energy-efficient computations, lowering overall energy consumption and computational processes (Zhou, 2024).

Many developers will use cloud-based services and hardware for most of their work, and they should use the most energy-efficient hardware and servers for their tasks. This means to:

- Use data centres powered by renewable energy, such as Digital Realty, Schneider Electric, or Google Cloud (Swallow, 2023).
- Use specialized HW for AI for quicker processing and less energy usage, such as Tensor Processing Units (TPUs) or graphics processing units (GPUs).
- Deploy AI models on edge devices or IoT devices whenever possible to reduce the need for data transmission and cloud-based processing.
- Match HW to specific tasks: While faster processor cores may be needed for particular jobs, high memory bandwidth may be more advantageous for others.

For many AI developers, significant energy savings can be made by developing software frameworks and libraries specifically designed to minimize AI energy consumption. Implementing techniques such as optimized runtime scheduling, task parallelization, and resource-aware programming can maximize software performance while reducing energy requirements (N3xtcoder, 2023).

In the context of edge computing, energy-efficient algorithms optimize AI tasks on devices with limited resources. By dynamically adjusting resource allocation, these algorithms reduce energy consumption while maintaining performance. For example, in innovative city applications, such algorithms enable longer device battery life and contribute to overall energy savings, advancing environmental sustainability (Hua et al., 2023).

Not only mentioned data centres but whole organizations can reduce carbon emissions, minimize environmental impact, and advance sustainability goals by harnessing renewable energy sources such as solar, wind, and hydroelectric power and biomass to power AI infrastructure instead of fossil fuels (Zhang et al., 2024).

An example of renewable energy's potential impact on AI is using wind power to support computing operations. Companies like Facebook and Amazon have invested in wind farms to power their data centers, reducing carbon emissions and promoting environmental sustainability while supporting AI technologies (Cacciuttolo et al., 2024).

An example of the potential impact of hardware innovation on AI is the development of neuromorphic computing chips, such as IBM's TrueNorth chip. These chips mimic neural networks, offering faster and more energy-efficient computations than traditional architectures. Neuromorphic chips have the potential to revolutionize AI, accelerating research and development while minimizing environmental impact (Li et al., 2023b).

4.2 Data efficiency

Data efficiency in AI development significantly affects environmental sustainability. Optimizing data usage while maintaining high performance can reduce the ecological footprint associated with data collection, storage, and processing (Adadi, 2021). Various data-centric approaches promote environmentally friendly machine-learning algorithms, including active learning, knowledge transfer, dataset distillation, augmentation, and curriculum learning (Salehi and Schmeink, 2024).

We also can develop techniques for training AI models with less data or using synthetic data, reducing the need for extensive data collection, which can have environmental and privacy implications.

Data efficiency in AI, like federated learning in healthcare and cross-lingual transfer learning in natural language processing, optimizes model training with decentralized data and less reliance on large datasets. These approaches enhance AI while preserving privacy, reducing environmental impact, and widening technology access (Hassan et al., 2024).

4.3 Model compression

Model compression is a critical technique in AI development that reduces the computational resources required for training and inference, promotes environmental sustainability, and enhances data efficiency. By compressing AI models, developers can minimize energy consumption, lower hardware requirements, and mitigate the ecological impact of AI technologies (Li et al., 2023a).

As the name suggests, model compression helps reduce the neural network size without compromising accuracy too much. Many model compression techniques can be used to reduce the model size. Some widely used ones are pruning, quantization, knowledge distillation and low-rank factorization (Pokrhel, 2022).

Model compression techniques enable deploying deep learning models on resource-constrained devices like smartphones and edge devices. This facilitates real-time inference and intelligent decision-making at the network's edge, improving efficiency and privacy while reducing reliance on centralized infrastructure (Singh and Gill, 2023).

4.4 Lifecycle assessment

Life Cycle Assessment (LCA) is a methodology used to evaluate the environmental impact of a product or system throughout its entire lifecycle, from raw material extraction, development and training to deployment and disposal. In AI, LCA is essential for understanding and mitigating the environmental footprint of AI technologies and systems (Oduque De Jesus et al., 2021).

An example of the potential impact of lifecycle management on AI is the implementation of predictive maintenance systems in industrial settings. These systems use AI to monitor equipment health in real-time, predicting failures before they occur. Organizations improve efficiency and reduce environmental impact by avoiding costly downtime and extending lifespan (Ucar et al., 2024).

4.5 Ethical AI

Ethical AI refers to the principles, practices, and policies that govern the responsible development, deployment, and use of artificial intelligence technologies. While ethical considerations in AI often focus on issues such as fairness, transparency, accountability, and privacy, they also play a crucial role in promoting environmental sustainability. Ethical AI guides responsible decision-making throughout the AI lifecycle, focusing on reducing impact and optimizing resource usage (Siau and Wang, 2020). AI ethics are complex, and these could be the starting points for related discussions:

- Developing a global repository of AI strategies and policies.
- Developing a governance structure or platform for ensuring accountability and transparency in the development of AI.
- Encouraging greater knowledge sharing among different states and stakeholders.
- Creating opportunities for states and other actors to collaborate on developing a global ethical framework for AI.
- Including diverse stakeholders in the development of AI policies and strategies.
- Investing in studying and comparing AI's social, ethical, political and environmental implications and its security and economic implications.

Unless we develop AI policies and regulations in a collaborative environment, AI is unlikely to foster collaboration and will instead reinforce competition norms (Ebadi, 2018).

An example of ethical AI could be privacy-preserving techniques, like federated learning and differential privacy, adopted by companies such as Apple and Microsoft, prioritizing user privacy while enabling valuable insights for AI systems. This fosters trust and encourages the broad adoption of AI technologies (Williamson and Prybutok, 2023).

4.6 Collaboration and Knowledge Sharing

Over the past few years, several countries worldwide have started to develop national artificial intelligence (AI) policies and strategies. AI, the digital economy and the future of work remain priorities (Ebadi, 2018). The AI community can address complex environmental challenges, develop innovative solutions, and advance sustainability goals by fostering collaboration among researchers, developers, policymakers, and stakeholders to share best practices, tools, and resources (Yu et al., 2024). Collaboration is necessary for ethical AI, as discussed in the previous section.

Despite the wide-reaching impacts of AI on various industries and sectors, no mechanism or body is charged with assessing national AI strategies, policies or ethics. The subject — which increasingly impacts day-to-day life internationally — is worth a severe assessment (Ebadi, 2018). There is also no up-to-date overview comparing national AI strategies.

Collaboration in AI, like in the development of AI-based medical imaging, brings together researchers, medical professionals, and imaging specialists to share expertise and data. This ensures accurate algorithms, improving diagnoses and patient outcomes (Pinto-Coelho, 2023).

4.7 Regulatory Measures

Governments and regulatory bodies can incentivize responsible AI development and deployment practices that minimize environmental impact, promote energy efficiency, and advance sustainability goals by establishing policies, regulations, and standards (De Almeida et al., 2021).

One example of regulatory impact on AI is the European Union's GDPR, enforced in May 2018. Designed to safeguard individuals' privacy and data rights, the GDPR affects AI systems handling personal data. It imposes stringent data collection, processing, and storage rules, prompting AI innovation to respect privacy and hold organizations accountable for data breaches. (Li et al., 2019).

4.8 Public Awareness and Education

By raising awareness about the environmental impact of AI technologies and educating the public about sustainable practices, individuals can make informed decisions, advocate for change, and contribute to collective efforts to address environmental challenges (Kandlhofer et al., 2023).

The aim is to empower individuals to reduce their environmental impact and promote ecological balance. AI platforms offer training, fostering sustainability leaders who drive innovation. Through green building and Corporate Social Responsibility (CSR) initiatives, AI education empowers professionals to lead transformative change (Nadeem and Arsalan, 2024).

The Climate Change AI (CCAI) initiative mobilizes a global community to use AI to combat climate change. It fosters collaboration and innovation through workshops, developing AI solutions for environmental challenges like deforestation monitoring and energy optimization (Cowls et al., 2021).

5. Discussion

The comparative analysis of eco-friendly AI practices reveals that integrating various strategies significantly enhances environmental sustainability in AI technologies. Enhancing energy efficiency, improving data efficiency, implementing model compression techniques, conducting lifecycle assessments, adhering to ethical AI principles, fostering collaboration, establishing robust regulatory measures, and raising public awareness collectively contribute to reducing AI's ecological footprint.

Empirical evidence supports the effectiveness of these strategies in promoting sustainable AI. For instance, optimized algorithms and renewable energy integration demonstrably reduced the carbon footprints of major tech companies. Model compression techniques have successfully deployed AI on resource-constrained devices, reducing energy consumption. Federated learning and other data-centric approaches have maintained model performance while minimizing environmental impact.

Despite advancements, challenges persist in achieving eco-friendly AI. Technologically, the scalability of renewable energy to meet AI demands remains problematic and ensuring model accuracy while reducing size requires compression techniques, which pose adoption barriers. Economically, transitioning to energy-efficient hardware and renewable energy sources involves high costs, deterring smaller companies. Additionally, sustained investment in eco-friendly AI research is challenging. Societally, more awareness about AI's environmental impact and the benefits of sustainable practices is needed, hindering support. Furthermore, implementing robust regulatory measures may require more stakeholder concern about operational costs and compliance burdens.

To overcome these challenges, a multifaceted strategy is essential. Technological advancements mean continuous innovation in renewable energy sources to enhance scalability for AI's energy needs, alongside refining model compression techniques for broader adoption. Economic incentives, such as government subsidies and grants, can ease the financial burden on companies adopting sustainable AI practices, while public-private partnerships leverage combined resources and expertise for innovation. Societal engagement involves targeted educational campaigns to raise awareness about AI's environmental impact and inclusive policy development, ensuring regulatory measures resonate with diverse stakeholders and garner widespread support.

6. Conclusion

Through articulating the mentioned strategies, this paper aims to raise awareness about the environmental impact of AI technology, offering actionable solutions for mitigating the ecological footprint of AI development and deployment. It strives to foster a culture of sustainability and responsible innovation within the AI community while guiding policymakers, researchers, developers, and stakeholders seeking to promote environmental stewardship in AI.

This paper serves as a resource for individuals and organizations interested in understanding and implementing sustainable practices in the field of AI. It empowers readers to consider the environmental implications of AI technology and encourages them to take proactive steps toward building a more sustainable future.

Acknowledgement

Paper was processed with support from institutional-support fund for long-term conceptual development of science and research at the Faculty of Informatics and Statistics of the Prague University of Economics and Business (IP400040) and the project IG 409033.

References

Adadi, A. (2021). A survey on data-efficient algorithms in big data era. Journal of Big Data, 8(1), 24. https://doi.org/10.1186/s40537-021-00419-9

Bhagat, P. R., Naz, F., & Magda, R. (2022). Artificial intelligence solutions enabling sustainable agriculture: A bibliometric analysis. PLOS ONE, 17(6), e0268989. https://doi.org/10.1371/journal.pone.0268989

Cacciuttolo, C., Cano, D., Guardia, X., & Villicaña, E. (2024). Renewable Energy from Wind Farm Power Plants in Peru: Recent Advances, Challenges, and Future Perspectives. Sustainability, 16(4), 1589. https://doi.org/10.3390/su16041589

Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021). The AI Gambit — Leveraging Artificial Intelligence to Combat Climate Change: Opportunities, Challenges, and Recommendations. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3804983

De Almeida, P. G. R., Dos Santos, C. D., & Farias, J. S. (2021). Artificial Intelligence Regulation: A framework for governance. Ethics and Information Technology, 23(3), 505–525. https://doi.org/10.1007/s10676-021-09593-z

Ebadi, B. (2018, August 20). Collaboration Is Necessary for Ethical Artificial Intelligence. Centre for International Governance Innovation. https://www.cigionline.org/articles/collaboration-necessary-ethical-artificial-intelligence/

Hassan, A., Prasad, V., Bhattacharya, P., Dutta, P., & Damaševičius, R. (Eds.). (2024). Federated learning and AI for healthcare 5.0. IGI Global

Hua, H., Li, Y., Wang, T., Dong, N., Li, W., & Cao, J. (2023). Edge Computing with Artificial Intelligence: A Machine Learning Perspective. ACM Computing Surveys, 55(9), 1–35. https://doi.org/10.1145/3555802

Jain, H., Dhupper, R., Shrivastava, A., Kumar, D., & Kumari, M. (2023). AI-enabled strategies for climate change adaptation: Protecting communities, infrastructure, and businesses from the impacts of climate change. Computational Urban Science, 3(1), 25. https://doi.org/10.1007/s43762-023-00100-2

Kandlhofer, M., Weixelbraun, P., Menzinger, M., Steinbauer-Wagner, G., & Kemenesi, Á. (2023). Education and Awareness for Artificial Intelligence. In J.-P. Pellet & G. Parriaux (Eds.), Informatics in Schools. Beyond Bits and Bytes: Nurturing Informatics Intelligence in Education (Vol. 14296, pp. 3–12). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-44900-0_1

Li, E.-P., Ma, H., Ahmed, M., Tao, T., Gu, Z., Chen, M., Chen, Q., Li, D., & Chen, W. (2023a). An Electromagnetic Perspective of Artificial Intelligence Neuromorphic Chips. Electromagnetic Science, 1(3), 1–18. https://doi.org/10.23919/emsci.2023.0015

Li, H., Yu, L., & He, W. (2019). The Impact of GDPR on Global Technology Development. Journal of Global Information Technology Management, 22(1), 1–6. https://doi.org/10.1080/1097198X.2019.1569186

Li, Z., Li, H., & Meng, L. (2023b). Model Compression for Deep Neural Networks: A Survey. Computers, 12(3), 60. https://doi.org/10.3390/computers12030060

N3xtcoder. (2023). How can AI and AI developers help reduce the energy usage of AI? https://n3xtcoder.org/developers-energy-impact-of-ai

Nadeem, W., & Arsalan, H. (2024). Promoting Environmental Sustainability through AI-driven Education Initiatives. Unpublished. https://doi.org/10.13140/RG.2.2.24109.09443

Nzubechukwu, C. O., Adebayo Olusegun Aderibigbe, Emmanuel Chigozie Ani, Peter Efosa Ohenhen, Donald Obinna Daraojimba, & Bukola A. Odulaja. (2023). AI-driven solutions in renewable energy: A review of data science applications in solar and wind energy optimization. World Journal of Advanced Research and Reviews, 20(3), 401–417. https://doi.org/10.30574/wjarr.2023.20.3.2433

Oduque De Jesus, J., Oliveira-Esquerre, K., & Lima Medeiros, D. (2021). Integration of Artificial Intelligence and Life Cycle Assessment Methods. IOP Conference Series: Materials Science and Engineering, 1196(1), 012028. https://doi.org/10.1088/1757-899X/1196/1/012028

Omar, R. (2023). AI And Energy Efficiency. 2023 IEEE 20th International Conference on Software Architecture Companion (ICSA-C), 141–144. https://doi.org/10.1109/ICSA-C57050.2023.00040

Onyebuchi, N. C., Preye Winston Biu, Aniekan Akpan Umoh, Bartholomew Obehioye Obaedo, Abimbola Oluwatoyin Adegbite, & Ayodeji Abatan. (2024). Reviewing the role of AI in environmental monitoring and conservation: A data-driven revolution for our planet. World Journal of Advanced Research and Reviews, 21(1), 161–171. https://doi.org/10.30574/wjarr.2024.21.1.2720

Pinto-Coelho, L. (2023). How Artificial Intelligence Is Shaping Medical Imaging Technology: A Survey of Innovations and Applications. Bioengineering, 10(12), 1435. https://doi.org/10.3390/bioengineering10121435

Pokhrel, S. (2022, January 19). 4 Popular Model Compression Techniques Explained. Xailient. https://xailient.com/blog/4-popular-model-compression-techniques-explained/

Salehi, S., & Schmeink, A. (2024). Data-Centric Green Artificial Intelligence: A Survey. IEEE Transactions on Artificial Intelligence, 5(5), 1973–1989. https://doi.org/10.1109/TAI.2023.3315272

Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. Journal of Database Management, 31(2), 74–87. https://doi.org/10.4018/JDM.2020040105

Singh, R., & Gill, S. S. (2023). Edge AI: A survey. Internet of Things and Cyber-Physical Systems, 3, 71–92. https://doi.org/10.1016/j.iotcps.2023.02.004

Swallow, T. (2023, May 3). Energy Digital. Top 10: Green Energy Data Centres. https://energydigital.com/top10/top-10-green-energy-data-centres

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., Mathis, A., Mathis, M. W., Van Langevelde, F., Burghardt, T., Kays, R., Klinck, H., Wikelski, M., Couzin, I. D., Van Horn, G., Crofoot, M. C., Stewart, C. V., & Berger-Wolf, T. (2022). Perspectives in machine learning for wildlife conservation. Nature Communications, 13(1), 792. https://doi.org/10.1038/s41467-022-27980-y

Ucar, A., Karakose, M., & Kırımça, N. (2024). Artificial Intelligence for Predictive Maintenance Applications: Key Components, Trustworthiness, and Future Trends. Applied Sciences, 14(2), 898. https://doi.org/10.3390/app14020898

Williamson, S. M., & Prybutok, V. (2024). Balancing Privacy and Progress: A Review of Privacy Challenges, Systemic Oversight, and Patient Perceptions in AI-Driven Healthcare. Applied Sciences, 14(2), 675. https://doi.org/10.3390/app14020675

Yu, P., Mulli, J., Syed, Z. A. S., & Umme, L. (Eds.). (2024). Facilitating Global Collaboration and Knowledge Sharing in Higher Education With Generative AI: IGI Global. https://doi.org/10.4018/979-8-3693-0487-7

Zhang, L., Ling, J., & Lin, M. (2022). Artificial intelligence in renewable energy: A comprehensive bibliometric analysis. Energy Reports, 8, 14072–14088. https://doi.org/10.1016/j.egyr.2022.10.347

Zou, C. (2024). Revolutionizing machine learning: Harnessing hardware accelerators for enhanced AI efficiency. Applied and Computational Engineering, 47(1), 141–146. https://doi.org/10.54254/2755-2721/47/20241256

HOW DIGITIZATION AND AI ARE TRANSFORMING AGRICULTURAL MANAGEMENT FOR RESILIENCE AND SUSTAINABILITY

Dagmar Skokanova

Czech Technical University in Prague
Masaryk Institute of Advanced Studies
Dagmar.Skokanova@cvut.cz

DOI: 10.35011/IDIMT-2024-455

Keywords

Agricultural management, AI ethics, artificial intelligence; food security; food waste reduction; precision agriculture; sustainability.

Abstract

The Food and Agriculture Organization of the United Nations projects that the global population will surpass 9 billion by 2050. With shrinking farmland, depleting natural resources, unpredictable climate shifts, and evolving market demands, the agricultural production system faces a transformative phase. To meet these challenges, agriculture must boost productivity, enhance efficiency, adapt to climate variability, and prioritize sustainability. However, artificial intelligence (AI) emerges as a promising tool to tackle these demands head-on. The current study aims at presenting the contemporary Czech young generation and its interest and attitude towards food security, development of Czech agriculture and the impact of agriculture on the environment. The study uses a structured questionnaire and analyzes the primary data to verify the hypothesis. The study finds out that current younger generation in the Czech Republic displays minimal interest in topics such as food security, contemporary trajectory of Czech agriculture, and environmental effects of agricultural practices, and thereby the hypothesis is confirmed. The practical value of the study lies in the empirical assessment of the awareness about the Food Security in the Czech Republic and in the world and the awareness of food self-sufficiency in the Czech Republic, which proves the need for increased awareness. On the other hand, the recognition of hunger as a pivotal issue within the UN's Sustainability Goals is widespread. It reflects a collective understanding of the urgency to address food insecurity as a fundamental aspect of sustainable development.

1. Introduction

The agricultural sector is confronted with myriad challenges that threaten its resilience and sustainability, exacerbated by climate change, population growth, and resource scarcity. Additionally, pest and disease management, food waste, supply chain vulnerabilities, and socioeconomic factors contribute to these difficulties. As Alexander et al. (2023) articulate, “factors such as increasing regulations and worsening environmental conditions are stressing agricultural systems and are opening windows of opportunity for technological solutions” (p. 146).

Shifts in weather patterns, the increased frequency of extreme weather events, and rising temperatures pose significant challenges to agricultural productivity. Events such as droughts, floods, heatwaves, and storms can damage crops, disrupt supply chains, and threaten food security. Resource scarcity, including the limited availability of essential resources such as water, arable land, and nutrients, presents significant hurdles for agricultural sustainability. Inefficient resource use can lead to soil degradation, water pollution, and the loss of biodiversity.

Moreover, pest infestations and crop diseases can devastate yields, resulting in significant economic losses for farmers. Traditional pest and disease management methods often rely on the indiscriminate use of chemical pesticides and herbicides, which can harm beneficial organisms and pollute the environment. The issue of food waste is also critical, with a substantial portion of the global food supply lost or wasted at various stages of the supply chain, from production to consumption. Food waste contributes to environmental degradation, squanders valuable resources, and exacerbates food insecurity.

Complex and interconnected supply chains are vulnerable to disruptions caused by natural disasters, geopolitical conflicts, and economic shocks. Inefficient logistics, lack of transparency, and market volatility can compromise the resilience of agricultural systems. Socioeconomic factors further complicate the landscape. Inequities in access to resources, technology, and markets can hinder the resilience and sustainability of agricultural systems, particularly in developing countries. Small farmers often lack access to credit, information, and infrastructure needed to adopt sustainable practices.

Addressing these challenges necessitates innovative approaches that enhance the resilience and sustainability of agricultural systems. AI technologies offer promising solutions by enabling data-driven decision-making, optimizing resource use, and facilitating more efficient and sustainable farming practices, ultimately contributing to the long-term viability of agriculture and food security.

2. Theory and Hypothesis

2.1 Digitization and Artificial Intelligence (AI)

Digitization is the integration of digital technologies into various sectors of society, aiming to replace numerous routine human activities with digitized processes. This transformation relies on both technical means (computers, the Internet, micro-sensors, data storage) and software tools (for storing, processing, and retrieving data), all interconnected in cyberspace and secured against losses, leaks, and cyber-attacks. The foundation of digitization is data captured digitally through sensors, cameras, etc., stored in large repositories (big data), and promptly retrieved, processed, and used.

Artificial Intelligence (AI) refers to automated decision-making processes that operate independently of humans and can reflect on experiences stored in data. AI uses mathematics and logic to simulate human learning and decision-making. Key elements of AI include large data sets, sufficient computing power, and appropriate programs to process and utilize the data. AI can be divided into software AI, such as virtual assistants and image recognition software, and product-related AI, like autonomous vehicles and drones. AI is useful for transforming raw data into usable information, optimizing complex systems, improving prediction, and discovering new contexts and insights.

2.2 Environmental Protection and Sustainability

Sustainability, originally termed sustainable development, gained prominence with the Club of Rome's 1972 publication "The Limits to Growth," which highlighted the unsustainable nature of current economic and population growth patterns. The book influenced the Stockholm Declaration, the first UN Conference on the Environment, and the Brundtland Commission's report "Our Common Future," which defined sustainable development as meeting present needs without compromising future generations' ability to meet theirs.

The 1992 Rio de Janeiro Summit established the three pillars of sustainable development: economic growth, environmental protection, and social equity. Subsequent global efforts, including the ban on harmful chemicals and the reduction of greenhouse gases, culminated in the adoption of the "17 Sustainable Development Goals" (SDGs) by the UN in 2015. These goals include ending hunger, achieving food security, and promoting sustainable agriculture.

2.3 Challenges and Goals in Sustainability

Despite numerous steps towards sustainability, significant challenges remain. The 2021 UN Climate Change Conference (COP26) highlighted discrepancies between commitments and actions, as exemplified by the participants' use of private jets. The focus on limiting global warming sometimes overshadows other critical issues, such as ending hunger, which remains a significant challenge due to economic and climatic factors. Approximately 690 million people are hungry, with 135 million suffering from acute hunger due to conflicts, climate change, and economic downturns.

The Global Hunger Index (GHI) measures progress in combating hunger based on undernourishment, child stunting, child wasting, and child mortality. Since 2015, progress has stagnated due to the COVID-19 pandemic, ongoing climate change, and conflicts, including the Russia-Ukraine war. There are 43 countries at risk of severe hunger, and 58 countries are unlikely to achieve low hunger by 2030. EU legislative actions, such as restricting arable land use, have faced protests and led to policy relaxations.

2.4 Digitalization and AI for Sustainability in Agriculture

The application of digitalization and AI in sustainability, particularly in agriculture and food production, offers promising solutions. AI can optimize resource use, enhance crop yields, and improve food security by analyzing large data sets and making informed decisions. Digital tools can monitor environmental conditions, predict crop performance, and manage supply chains efficiently, contributing to sustainable agricultural practices and food production systems.

Next, let us focus on the use of digitalization and artificial intelligence for sustainability, particularly in the context of agriculture and food production.

2.5 Hypothesis

Digitalization has penetrated nearly every industry, including agriculture. Technologies such as precision farming and smart irrigation systems enable farmers to optimize resources, enhance productivity, and minimize environmental impact. AI further revolutionizes agricultural practices by analyzing vast data and providing actionable insights.

This study examines the interests and attitudes of the younger generation in the Czech Republic towards food security and self-sufficiency in food production. Food Security, as defined by the 1996 World Food Summit, is the state where "all people at all times have physical and economic access to sufficient, safe, and nutritious food" (Adra, 2020). According to the WHO, over 822 million people worldwide suffer from hunger, with most cases in Asia and Africa. Despite significant progress, achieving the UN Sustainable Development Goal of Zero Hunger by 2030 remains a challenge.

In Europe, agriculture is a priority for the EU, focusing on food security, environmental impact reduction, and promoting local farming. The Common Agricultural Policy (CAP) for 2023-27 aims to support farmers, ensure food quality, and protect the environment (European Commission, 2023). Food security levels and food waste vary across EU countries. In the Czech Republic, food waste is relatively low, with 80% of Czechs wasting little to no food, though Generation Z wastes the most. This generation will influence future behavior patterns and shows interest in reducing waste.

The issue of food security is relevant for all countries, but it does not have a universal solution. (Kushniruk at al. 2022) This is evident in the European Union countries, where despite sharing a common food policy, varying levels of food security are apparent. In addition, the levels of food waste also differ. The situation in the Czech Republic suggests that the Czech Republic is one of the most successful EU countries in terms of food waste in households. According to Vesela (2023) from Mendel University in Brno, "almost 80% of Czechs do not waste food at all or waste only up to 5% of the food they buy." However, Vesela states (2023) that Generation Z (people born between 1996 and 2009) wastes the most food. Generation Z is the generation that will be responsible for passing on patterns of behaviour to the generations to come. It is clear though that young people are interested in relevant information and want to reduce waste.

Therefore, we hypothesize the following:

H1: The current younger generation in the Czech Republic displays minimal interest in topics such as food security, contemporary trajectory of Czech agriculture, and environmental effects of agricultural practices.

3. Methodology and Research Results

3.1 Respondent Selection Process

The research aimed to gauge the attitudes of the contemporary young Czech generation towards food security. To achieve this, a systematic and detailed respondent selection process was employed, ensuring a representative and diverse sample. Below is an explanation of the respondent selection process.

The target demographic for this study comprised young Czech individuals aged 18 to 39. This age range was selected because it includes Generation Z (born approximately between 1996 and 2010) and younger Millennials (born approximately between 1981 and 1996). These groups are particularly relevant as they are expected to play a significant role in future food security practices and policies, both as consumers and as potential policymakers.

The selection criteria was based on age, residency and diversity. Only individuals between 18 and 39 years old were included to focus on the attitudes of the younger generation. Respondents needed to be residents of the Czech Republic to ensure that their perspectives were relevant to the local context of food security. Efforts were made to include respondents from various educational, occupational, and geographical backgrounds to capture a wide range of attitudes and opinions.

The selection process aimed to ensure a representative sample of the young Czech population by targeting diverse groups through educational, professional, and community channels. This approach provided a comprehensive understanding of the attitudes of the young generation towards food security in the Czech Republic.

3.2 Recruitment Method

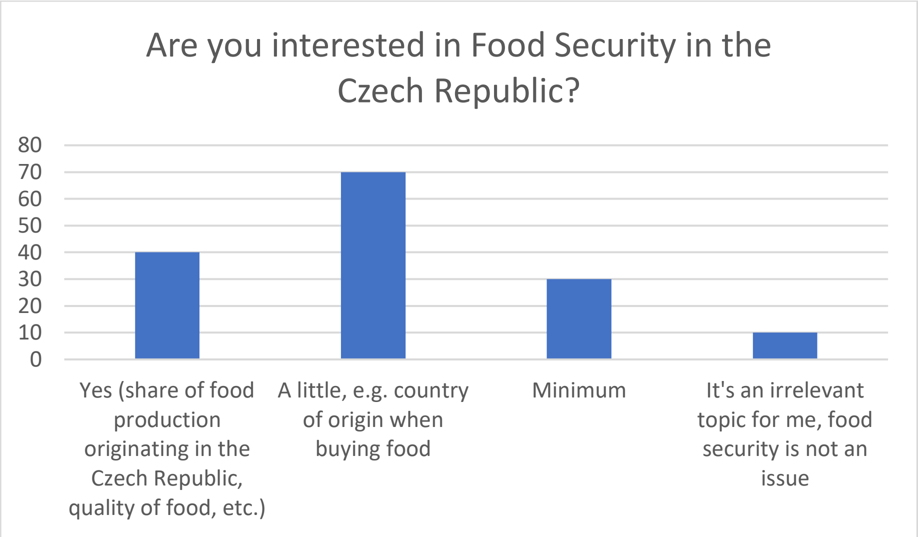
To recruit participants, a structured questionnaire was distributed electronically via email. The emails were sent to more than 300 individuals using the following strategies. Emails were sent to students and young professionals associated with universities and colleges. Educational institutions were chosen because they have a concentrated population of young people who are likely to be informed and interested in contemporary issues, including food security. Emails were also distributed through professional networks and organizations that include young professionals. Community and social groups, including online forums and local community centres, were utilized to reach a broader audience within the target demographic.

Out of the more than 300 individuals contacted, 151 responses were received, resulting in a response rate of approximately 50%. The data collection took place in March 2024. The questionnaire, designed based on existing literature, was verified for its relevance and clarity before distribution.

3.3 Research Results

The first question surveyed respondents' awareness of food security in the Czech Republic. The results showed that 70% of respondents expressed a moderate interest in food security, such as wanting to know the country of origin when buying food, 40% of respondents were interested in knowing the share of food production originating in the Czech Republic and the quality of food. Nearly one-third indicated minimal interest in food security and 10% considered food security an irrelevant topic.

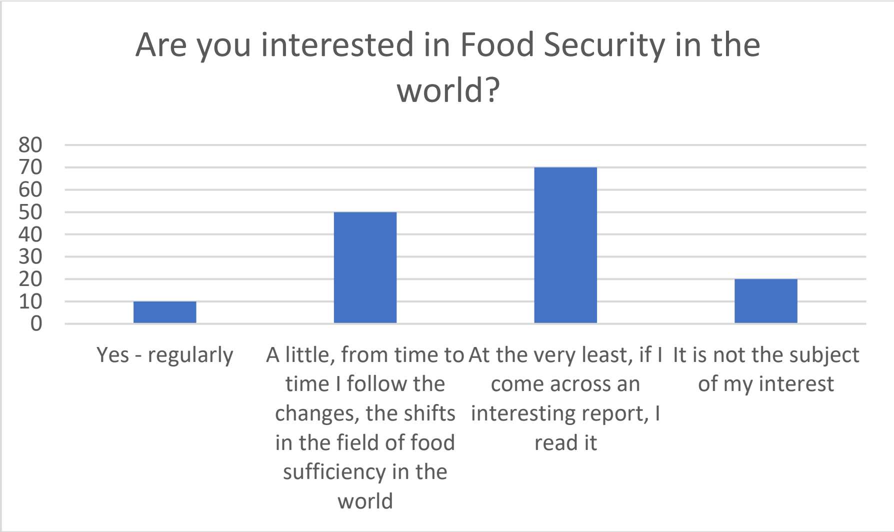
Table 1. Interest in Food Security in the Czech Republic



Source: (Own processing)

Another question focused on respondents' interest in global food security. The research results indicate that a significant majority (70%) of respondents have at least a basic interest in the topic, with many reading reports when they find them engaging. Approximately 50% of respondents follow developments in global food security intermittently, keeping track of occasional changes and shifts in the field. Only 10% of respondents expressed a regular and active interest in global food security, while 20% reported that the subject does not interest them at all.

Table 2. Interest in Food Security in the world

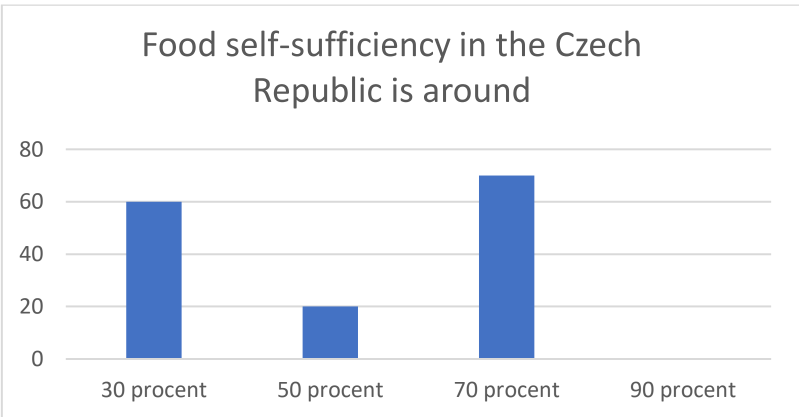


Source: (Own processing)

The next question assessed respondents' awareness of food self-sufficiency in the Czech Republic, which is approximately 70 percent. Table 3 shows that the majority of respondents accurately identified the food self-sufficiency level, with most believing it to be around 70 percent. This indicates a relatively high level of awareness about the country's food self-sufficiency among the surveyed group. About 60 percent of respondents mistakenly believed that food self-sufficiency is 30 percent. Despite the accurate awareness among some, about 60 percent of respondents mistakenly believed that the food self-sufficiency level is only 30 percent. This substantial misconception suggests that a considerable number of young Czechs may underestimate the country's capability to produce its own food. This could be due to a lack of access to accurate information or misunderstanding of the term 'food self-sufficiency'.

20 percent of respondents estimated the food self-sufficiency level to be 50 percent. While this estimate is incorrect, it is closer to the actual figure compared to those who believed it to be 30 percent. This suggests a moderate level of understanding among this group. Interestingly, no respondents believed the food self-sufficiency level to be as high as 90 percent. This indicates that overestimation of food self-sufficiency is not a prevalent issue among the surveyed population, which could imply a general awareness of the limitations and challenges faced by the Czech agricultural sector.

Table 3. Food self-sufficiency in the Czech Republic

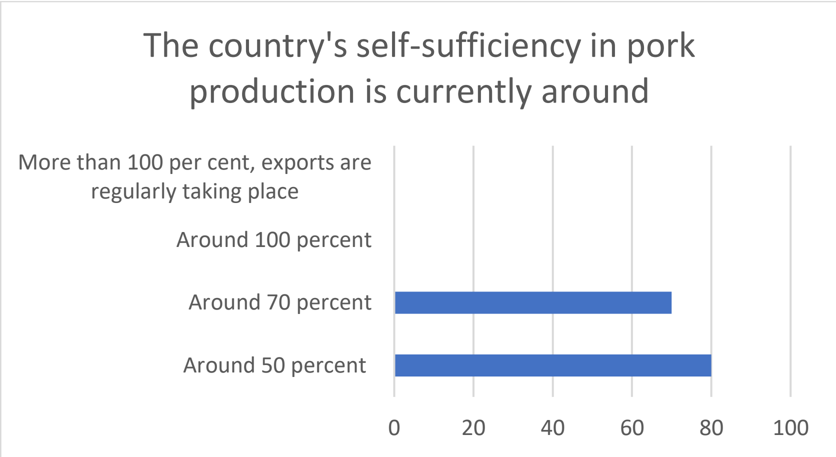


Source: (Own processing)

Another question assessed the respondents' perceptions of the Czech Republic's self-sufficiency in pork production. Historically, the Czech Republic has not been self-sufficient in this commodity, with the self-sufficiency rate standing at 51.2% in 2021, a decline from previous years. Over the long term, self-sufficiency in pork production has decreased by 18.8% since 2011.

Table 4 presents a notable disparity between respondents' beliefs and the actual data. A significant majority, 70% of respondents, estimated that the country's self-sufficiency in pork production is around 70 percent. Furthermore, 80% of respondents believed the self-sufficiency rate to be approximately 50 percent. This overestimation highlights a widespread misconception regarding the current state of pork production self-sufficiency in the Czech Republic.

Table 4. The country's self-sufficiency in pork production



Source: (Own processing)

Hunger is widely recognized as a critical issue within the United Nations' sustainability goals. Ending hunger is not only a fundamental human right but also a vital component of achieving sustainable development. Lack of access to nutritious food adversely impacts health and well-being and hinders the ability to break the cycle of poverty.

The United Nations' Sustainable Development Goal 2 (Zero Hunger) explicitly aims to end hunger, achieve food security, improve nutrition, and promote sustainable agriculture by 2030. This goal acknowledges the interconnectedness of food systems, poverty, and environmental sustainability. Addressing hunger necessitates a multifaceted approach that includes poverty alleviation, enhancing agricultural productivity, improving food distribution systems, and mitigating climate change.

The last question whether the problem of hunger is currently one of the UN's sustainability goals reveals that all respondents recognize hunger as a current UN sustainability goal. This unanimous acknowledgment underscores the global urgency of addressing hunger. It highlights the imperative for coordinated efforts from governments, international organizations, civil society, and the private sector to address the root causes of hunger and ensure universal access to sufficient, safe, and nutritious food.

4. Comparison of findings with similar studies

The findings of this study were compared with the "Food Safety in the EU" report, which highlights distinct yet related emphasis in the field of agriculture and ensuring food security.

The "Food Safety in the EU" report from the Special Eurobarometer Wave EB97.2 conducted between March and April 2022, centers on the perceptions and attitudes of EU citizens towards food safety. The report explores various themes including awareness and concerns about food safety topics, factors influencing food-related decisions, and levels of trust in different sources of information about food risks. It reveals that cost, taste, and food safety are primary considerations for Europeans when purchasing food, and highlights the high level of awareness about food safety topics among EU citizens. The report also discusses the public's concern about pesticide residues, antibiotic residues in meat, and additives, which are seen as major food safety issues.

While this study focuses on presenting the contemporary Czech young generation and its interest and attitude towards food security, the Eurobarometer report emphasizes the consumer perspective on food safety and the importance of addressing public concerns to maintain trust in the food system. Both articles highlight the importance of ongoing innovation and clear communication to maintain the sustainability and safety of the food supply chain.

5. Robotics and Automation in Agriculture

5.1 Robotics in Plant Protection

The rapid advancement of new technologies and the evolving digital landscape present unique opportunities for the development of automated and robotic systems in agriculture and forestry. Technological progress in machine vision, GPS, laser technology, and mechatronics has facilitated the implementation of robotic systems and smart technologies for precision agriculture, both indoors and outdoors.

In recent years, significant development has focused on advanced systems that integrate wireless sensor networks, machine learning, machine vision, and artificial intelligence for automated agriculture (Radova, 2021). Intelligent technologies employing machine vision and learning are being developed for a range of tasks, including planting, irrigation, weeding, pruning, harvesting, and the detection and identification of plant pests. The challenge of disease and pest detection remains a critical and intriguing area of research.

The interest in agricultural robots has surged due to their potential applications and the industry's drive to incorporate robotics. Process automation is particularly valuable in tasks such as soil preparation, sowing, fertilization, and harvesting. Repetitive, time-sensitive tasks represent ideal applications for robotic systems.

Automated weeding robots, designed for navigation and economic efficiency, are a significant focus of research. These robots leverage machine vision to distinguish weed species from cultivated crops (Fennimore et al., 2019). The accurate identification of weeds is crucial for effective weed management and conservation. The detection and identification process in agriculture parallels the anomaly screening processes in industrial production, underscoring its importance in precision agriculture.

6. Future Outlook and Conclusions

The transformative potential of digitization and artificial intelligence (AI) in agricultural management is evident in addressing the complex challenges facing global agriculture. With the world population growing and environmental pressures intensifying, the need for sustainable and resilient agricultural practices is increasingly urgent. AI provides innovative solutions to enhance productivity, reduce waste, and adapt to changing climatic conditions, thereby promoting food security and environmental sustainability.

However, our study highlights a concerning lack of interest among the contemporary young generation in the Czech Republic regarding critical issues such as food security, agricultural development, and environmental impact. This finding underscores the importance of raising awareness and fostering engagement among youth to secure a sustainable future for agriculture.

Moving forward, bridging the gap between technological innovation and public awareness is essential. Educating and empowering individuals, particularly the younger generation, about the significance of sustainable agriculture and the role of AI in fostering resilience can lead to more informed decision-making and meaningful action. By leveraging the potential of digitization and AI while prioritizing ethical considerations and sustainability, we can pave the way for a more resilient and sustainable agricultural future.

References

Alexander, C. S., Yarborough, M., & Smith, A. (2023). Precision agriculture. Springer. 25(1), 146-185. doi: 10.1007/s11119023-10063-3

Fennimore, S., A., Cutulle, M. (2019): Robotic weeders can improve weed control options for specialty crops. https://doi-org.ezproxy.techlib.cz/10.1002/ps.5337

Global Hunger Index, Concern Worldwide a Welthungerhilfein 2023, https://www.globalhungerindex.org/

Kushniruk, V., Kulinich, T., Roik, O., Lushchyk, M. (2022). Sustainable Development:Strengthening of Food Security in EU Countries, 24 (11), 85-91.

Meadows, D.H., Meadows, D., L., Randers, J. & Behrens, W., W. (1972). The Limits to Growth, Universe Books, New York, ISBN 978-0876632222

Ministerstvo zemědělství zmírní požadavky na vyčleňování neprodukčních ploch. Letos zemědělci nemusí nechat pole ležet ladem, MZm 15.2.2024, https://eagri.cz/public/portal/mze/tiskovy-servis/tiskove-zpravy/ministerstvo-zemedelstvi-zmirni-pozadavky-na-vyclenovani-neprodukcnich-ploch-letos-zemedelci-nemusi-nechat-pole-lezet-ladem

Murta, F., T. (2023): Artificial Intelligence for Sustainability: what is the role of AI in advancing targets for sustainability, in: https://tema-project.eu/articles/artificial-intelligence-sustainability-what-role-ai-advancing-targets-sustainability

Nařízení vlády č. 73/2023 Sb. o stanovení pravidel podmíněnosti plateb zemědělcům, https://www.zakonyprolidi.cz/cs/2023-73

Radova, S. (2021). Současný trend v robotizaci ochrany rostlin, https://www.agromanual.cz/cz/clanky/technologie/precizni-zemedelstvi/soucasny-trend-v-robotizaci-ochrany-rostlin

Report from the Commission to the European Parliament and the Council, Summary of CAP Strategic Plans for 2023-2027: joint effort and collective ambition, Brussels, 2023, https://agriculture.ec.europa.eu/document/download/6b1c933f-84ef-4b45-9171-debb88f1f757_en?filename=com-2023-707-report_en.pdf

Report of the World Commission on Environment and Development: Our Common Future, Oxford University Press, 1987, in: https://sustainabledevelopment.un.org/content/documents/5987our-common-future.pdf

The 17 Sustainable Development Goals (SDGs), United Nations, 2015, https://sdgs.un.org/goals#

The SDGS in Action, United Nations Development Programme, 2015, https://www.undp.org/sustainable-development-goals

Vesela, L. (2023): Generation Z wastes the most food, experts want to find a way to change that, https://mendelu.cz/en/generation-z-wastes-the-most-food-experts-want-to-find-a-way-to-change-that/?psn=1321.5999755859375

Weinert, J.: What is the role of artificial intelligence in achieving the SDGs? INGLOSUS Foundation, 9/2023. https://www.linkedin.com/pulse/what-role-artificial-intelligence-achieving-sdgs/

Williams, O., A (2021): 118 Private Jets Take Leaders To COP26 Climate Summit Burning Over 1,000 Tons of CO2, 2021, Forbes

EARLY CAREER & STUDENT SHOWCASE

EXAMINING ENTERPRISE ARCHITECTURE SUPPORT IN THE SCALED AGILE FRAMEWORK

Daniel Remta, Alena Buchalcevova

Prague University of Economics and Business
Faculty of Informatics and Statistics
xremd03@vse.cz, alena.buchalcevova@vse.cz

DOI: 10.35011/IDIMT-2024-467

Keywords

Scaled Agile Framework; Enterprise Architecture; TOGAF; Maturity Assessment.

Abstract

Frameworks for scaling Agile were introduced to help enterprises achieve a better response to change and benefit from Agile in the large. Among several frameworks that have been developed for scaling Agile, the Scaled Agile Framework is the one most used. Concurrently, the Enterprise Architecture (EA) discipline can help enterprises achieve their current and future objectives. The reflection of Enterprise Architecture concepts in frameworks for scaling Agile has not been researched yet. This paper aims to analyze and assess the level of EA support in the Scaled Agile Framework (SAFe). The presented results are beneficial for enterprises that are considering the adoption of SAFe and enterprise architecture concepts.

1. Introduction

To stay competitive in global and unforgiving environments, enterprises are adopting different agile methodologies (Dikert et al., 2016). Additionally, enterprises must keep track of new technologies, new heights of data swell, complexity, and never-ending changes, whilst all the organizations' parts should be kept aligned to reach the organization's goals. This is where Enterprise Architecture (EA) can help. Enterprise Architecture supports business and information technology alignment by providing a coherent whole of principles, methods, and models that are used in the design and realization of an enterprise's organizational structure, business processes, information systems, and infrastructure (Lankhorst et al., 2010). Concurrently, to support the usage of Agile methods beyond the small, single-team projects (Boehm & Turner, 2005), several frameworks for scaling Agile were developed, e.g., Scaled Agile Framework (SAFe), Large Scale Scrum (LeSS), Scrum@Scale, and Nexus.

While the application of EA concepts into enterprises is considered beneficial (Hanschke et al., 2015; Kotusev, 2019), and research papers with a focus on blending EA frameworks with Agile exist (Buckl et al., 2011; Hanschke et al., 2015), it seems that the reflection of EA in the frameworks for scaling Agile has not been researched yet. Therefore, there is no clear understanding of how scaled agile frameworks contribute to the implementation of EA concepts in the enterprise. This paper aims to analyze and assess the level of EA support in the Scaled Agile Framework (SAFe). The presented results are beneficial for enterprises that are considering the adoption of both SAFe and enterprise architecture concepts. The main contribution lies in assessing and presenting the EA maturity level that can be achieved by implementing the SAFe framework by the book.

2. Background

Enterprise Architecture. Enterprise Architecture (EA) describes the fundamental artifacts and their interrelationships in an aggregate model for business and IT management (Canat et al., 2018). The ISO/IEC/IEEE 42010:2011 standard defines EA as managing an enterprise's fundamental organization, its components, their relationships, and the principles governing its design and evolution (ISO/IEC/IEEE, 2011). Unlike IT-focused architecture, EA includes business artifacts like goals, products, services, processes, and performance indicators (Winter & Fischer, 2006). EA's layers—Business, Process, Integration, Software, and Technology architecture—provide a holistic view (Winter & Fischer, 2006). A common mindset and architectural view intertwining business and IT plans are essential (Li et al., 2016). EA addresses this need and offers several benefits to enterprises: Bridge the communication gap between business and IT stakeholders (Kotusev, 2019); Improve Business and IT alignment (Kotusev, 2019); Increase enterprise agility through extended capability to respond to external changes (Fallmyr & Bygstad, 2014); Improve decision-making with regards to IT principals and investments (Hanschke, 2009); Develop roadmaps for modernization, orderly replacement of obsolete technology and infrastructure (Bellman & Griesi, 2015); Develop roadmaps for acquiring skills and specialties in the workforce (Bellman & Griesi, 2015).

EA aims to achieve enterprise objectives through boundaryless information flow, effective decision support, and alignment of business, application, data, and technical architectures (Bellman & Griesi, 2015). Various EA frameworks help organizations implement EA by providing vocabulary, artifacts, concepts, principles, templates, and reference models for communication and implementation (Hanschke, 2009). The most influential framework is TOGAF, offering comprehensive methods and guidelines (OMG, 2018). However, TOGAF is often seen as too heavy, slow, and documentation-driven, and it may not be suitable for off-the-shelf adoption by specific enterprises (Gill et al., 2014).

Enterprise Architecture in Agile Environments. The Agile Manifesto states that the best architectures, requirements, and designs emerge from self-organizing teams, contrasting with traditional upfront architecture (Beck et al., 2001). Agile and architecture are not opposites, but there's no clear understanding of how architecture works in agile development (Kotusev, 2019; Rost et al., 2015). Agile and EA can be combined, though architecture at lower organizational levels may not be worth the investment due to rapidly changing circumstances (Canat et al., 2018). EA is more beneficial at higher levels, while applying it at lower levels may be over-architecting. Agile's original focus on small teams implies that comprehensive upfront specifications are unnecessary (Boehm & Turner, 2005). However, as Agile expands into large companies, scaling introduces challenges like complex coordination, lack of architecture, and poor requirement analysis (Paasivaara & Lassenius, 2016). Integrating agile methods into traditional top-down organizations is difficult and can cause conflicts in development and business processes (Boehm & Turner, 2005). Thus, EA's holistic approach becomes important again.

SAFe. The Scaled Agile Framework (SAFe) is a freely revealed knowledge base of proven, integrated patterns for enterprise-scale Lean-Agile development (Scaled Agile, 2023) currently available in its 6.0 version. Research presented in this paper has been done on SAFe version 5.0, for which the "Big Picture" is still accessible (Scaled Agile, 2020). It has three configurations: Essential, Large Solution, and Portfolio. Essential configuration encompasses Team and Program levels, where the Portfolio has the Value stream level added. Activities are performed on each level, and in addition all levels are tied together (Alqudah & Razali, 2016). Some benefits of SAFe in relation to architecture have been described in (Canat et al., 2018). Framework claims to provide a recipe for adopting agile at the enterprise scale (Ebert & Paasivaara, 2017). SAFe is quite a complex framework that is being considered by some practitioners to be evolving into a new waterfall (Ebert & Paasivaara, 2017). Some other concerns, like SAFe's strong emphasis on process rather than on people, were mentioned (Remta et al., 2020).

EA Maturity Assessment. Maturity assessment has been proven as a tool to foster improvement in various disciplines; EA is not an exception. Maturity assessment in the Enterprise Architecture discipline can act as a governance instrument for analysis and evaluation of the current state, as well as identification of possible areas for improvement (Proença & Borbinha, 2017). Models for maturity assessment typically consist of several "maturity levels", and the definition of the process of the assessment conducted to make it repeatable and available for benchmarking (Proença & Borbinha, 2017). There are several maturity models for EA, like the NASCIO EA Maturity Model (NASCIO, 2003), GAO Organizational Transformation – A Framework for Assessing and Improving EA Management (Version 2.0), or US DoC ACMM Framework (Version 1.2) (United States Department of Commerce, 2007). TOGAF lists US DoC ACMM Framework as an example of how to measure the maturity of EA. US DoC ACMM provides a framework that represents the key components of a productive enterprise architecture process. US DoC ACMM contains nine architecture elements, each of which can reach different maturity levels ranging from 0 to 5 where 5 represents the most mature one. The elements are: " (1) Architecture process (2) Architecture development (3) Business linkage (4) Senior management involvement (5) Operating unit participation (6) Architecture communication (7) IT security (8) Architecture governance (9) IT investment and acquisition strategy" (United States Department of Commerce, 2007).

3. Research Methods

The research was carried out in the following steps, which are in more detail described in this section. (1) Selection of the framework, (2) Research question definition, (3) Literature review, (4) Preliminary analysis of SAFe from the viewpoint of EA support, (5) EA maturity assessment tool selection, (6) EA Maturity Assessment, (7) Results presentation.

We selected the Scaled Agile Framework (SAFe) for our research. The main selection criterion was SAFe being the most widely used framework for scaling agile (digital.ai, 2022). Additionally, SAFe also claims to provide a recipe for adopting agile at the enterprise scale (Ebert & Paasivaara, 2017) and explicitly declares enterprise architecture support. Still, how could SAFe contribute to the implementation of EA concepts in the enterprise is not clear. Thus, the main research question was defined: "To what level is EA discipline covered in the Scaled Agile Framework (SAFe)?"

The literature review was conducted to understand the existing knowledge in the area. The search for publications was carried out in AMC Digital Library, eResources of the Czech National Library of Technology, covering multiple databases (i.e., SpringerLink, Wiley Online Library, Science Direct, IEEE/IET Electronic library), and Google Scholar. Moreover, the publicly available information on the frameworks' official websites was analyzed (OMG, 2018; Scaled Agile, 2023).

Next, an analysis of the SAFe was conducted to explore any potential EA reflection. The intent was to find: (1) Considerations of any architecture concepts; (2) Descriptions of units responsible for any architecture; (3) Roles dedicated to architecture; (4) Definitions of responsibilities of architects; (5) Incorporation of architectural processes in software development; (6) Described relations between architecture and business. The US DoC ACMM Framework (Version 1.2) (United States Department of Commerce, 2007) was used for the maturity assessment. For each of the nine enterprise architecture elements described in the US DoC ACMM Framework, the corresponding maturity level was assigned using an ACMM evaluation matrix. SAFe's description was continuously mapped to the matrix, and composite scores for each element were recorded. The evidence level assigned is provided in the Results section, together with a listing and brief description of the architectural elements. Last, the results were presented and discussed.

4. Results

Preliminary Analysis of SAFe from the Viewpoint of EA support. The aim was to identify signs of incorporating EA processes and roles based on the official SAFe version 5.0 descriptions (Scaled Agile, 2020). The examination consisted of a search to find: (1) Considerations of any architecture concepts in the framework; (2) Descriptions of units responsible for any architecture; (3) Roles dedicated to architecture; (4) Definitions of responsibilities of architects; (5) Incorporation of architectural processes within software development processes; (6) Described relations between architecture and business. The results are presented in Table 1.

Table 1. Preliminary Analysis Results

EA Aspect	Presence in SAFe
Considerations of architecture concepts	Architecture is considered on all levels of the framework, and relations to the development process are described.
Descriptions of units responsible for architecture	Have architectural groups, as well as architectural gatherings defined.
Roles dedicated to architecture	Dedicated architecture role for each level.
Definitions of responsibilities of architects	Architects drive architectural initiatives via enabler epics, participate in their analysis, modelling, and design. Architects also facilitate reusing code, components, and proven patterns whilst remaining responsible for architectural decisions.
Incorporation of architectural processes within software development processes	Providing "architectural runway and governance enables supporting current and future needs" (Scaled Agile, 2020). Architecture runways and features are needed for further development.
Described relations between architecture and business	Considers connection to business, including feedback loops, and prescribes collaboration on architecture with different stakeholders at all levels.

Source: (author)

EA Maturity Assessment of SAFe. We conducted the EA maturity assessment of SAFe to understand how EA is reflected in the framework. As a maturity assessment base, The US DoC ACMM Framework (Version 1.2) (United States Department of Commerce, 2007) was used. During the assessment, we were continuously comparing the descriptions of SAFe as available on the official webpage (Scaled Agile, 2020) with the scoring model from US DoC ACMM Framework (Version 1.2) (United States Department of Commerce, 2007) and assessed each of its nine architecture elements (United States Department of Commerce, 2007). In Table 2, for each architecture element, we present the level description from ACMM, assigned maturity level, and mapping to the ID of assessment result explanation, which is then provided in Table 3.

Table 2. Architecture Element, ACMM Level Description, Assigned Maturity and Explanation Mapping

Element	US DoC ACMM Level Description	Maturity	Mapping
Architecture Process	Basic enterprise architecture process is documented based on OMB Circular A-130 and Department of Commerce Enterprise Architecture Guidance. The architecture process has developed clear roles and responsibilities	2	A

Element	US DoC ACMM Level Description	Maturity	Mapping
Architecture Development	IT vision, principles, business linkages, Baseline, and Target Architecture are identified. Architecture standards exist, but are not necessarily linked to Target Architecture. Technical Reference Model (TRM) and Standards Profile framework are established"	2	B
Business Linkage	Explicit linkage to business strategies	2	C
Senior Management Involvement	Management awareness of architecture effort	2	D
Operating Unit Participation	The entire operating unit accepts, and actively participates in, the enterprise architecture process	4	E
Architecture Communication	Architecture documents updated regularly on DoC enterprise architecture web page	3	F
IT Security	IT security considerations are ad hoc and localized	1	G
Architecture Governance	Explicit documented governance of majority of IT investments	3	H
IT Investment and Acquisition Strategy	Little or no involvement of strategic planning and acquisition personnel in the enterprise architecture process. Little or no adherence to existing standards"	1	I

Source: (author)

Table 3. Assigned Maturity Assessment Explanation

ID	Assessment Explanation
A	SAFe defines roles and responsibilities for the Enterprise Architecture process across different levels. At the portfolio level, Enterprise Architects set the strategic technical direction, leverage emerging opportunities, mitigate threats, and create a technology strategy roadmap supporting business capabilities. They lead design, engineering, reuse, and pattern application (Scaled Agile, 2020). At the program level, Solution Architects ensure technical alignment across teams by defining and communicating the technical vision. At the team level, System Architects ensure solutions fit their intended purpose and coordinate with Enterprise and Solution Architects. While SAFe's EA processes are not documented by OMB Circular A-130 and Department of Commerce Enterprise Architecture Guidance, it still provides detailed role descriptions, although it lacks detailed business architecture prescriptions.
B	Vision, as a description of the future state, is present in SAFe. Solutions are defined by Solution Context and Solution Intent, the latter guiding decisions, demonstrating compliance, and recording system architecture decisions (Scaled Agile, 2020). Solution Context identifies aspects impacting Solution Intent. SAFe includes agile architecture principles, recommending a minimum viable ("just enough") architecture over big upfront designs. Architects support business alignment by optimizing architecture to achieve business goals quickly. Enterprise, Solution, and System Architects ensure program and product strategies align with business objectives, maintaining consistency with the overall enterprise strategy (Scaled Agile, 2020).
C	EA architects and Portfolio Managers collaborate on a high-level vision of enterprise solutions and development initiatives. Next, by close collaboration with Business Owners and Product Managers, architects on all levels ensure architecture supports current and future business goals. An enterprise architecture strategy, driven by the Enterprise Architect, enables embracing fast organizational changes to help increase the competitive advantage of the enterprise.
D	Enterprise Architects create "Enabler Epics for the architectures" within a portfolio, making them visible in Portfolio Kanban systems (Scaled Agile, 2020). They guide these epics through Kanban, ensuring necessary features and capabilities support architectural initiatives. Architects collaborate with Product Management to prioritize and balance new functionality with technical work. They also

ID	Assessment Explanation
	communicate architectural strategy and business drivers to non-technical stakeholders and work with Portfolio Management to ensure awareness of architecture efforts.
E	SAFe has well-defined processes for communicating and incorporating architecture at all levels, covering top-down and bottom-up information flows. Epics, capabilities, features, and stories represent the work at the portfolio, program, and team levels. Stakeholders must incorporate these into roadmaps and backlogs, prioritizing them among other tasks. Dedicated architects ensure architectural alignment, facilitate code reuse, and mentor teams. Teams provide feedback on architectural decisions while building features, balancing intentionality and emergence (Scaled Agile, 2020). This feedback influences the overall architecture strategy set by Enterprise Architects. Alignment on planned work occurs during PI Planning.
F	Although SAFe does not specify a location for storing architecture documents, it provides detailed architecture artifacts and processes. SAFe requires "a single source of truth regarding the intended and actual architecture" (Scaled Agile, 2020), regularly updated. The Solution Intent repository contains knowledge of current and intended Solution behavior, defined by solution requirements. These requirements and designs are represented by Capabilities, Features, Stories, and Nonfunctional requirements. Architects and teams collaborate to define enablers in roadmaps, explore technical options, and provide early feedback on the architecture, which is reflected in solution requirements.
G	SAFe lacks detailed guidelines for building or maintaining IT security. It suggests that Enterprise Architects provide technical guidance, including security for the portfolio solution, and synchronize system and data security and quality. However, beyond using nonfunctional requirements, there are no specific standards or integration guidelines. SAFe includes Security specialists who contribute to architecture by working on Architectural runways for near-term development initiatives, but their role is not clearly defined.
H	There are roles accountable for governance (i.e., Portfolio Management, Epic Owners, Enterprise Architect), events (i.e., Portfolio Sync, Participatory Budgeting, Strategic Portfolio Review), and artifacts for the governance of IT investments defined. SAFe has a concept of Lean Budget Guardrails. Every portfolio operates within an approved budget.
I	There are Strategic Themes with specific, itemized business objectives that connect the portfolio to the enterprise business strategy (Scaled Agile, 2020). Essential guidance on how to define Strategic Themes using Objectives and Key Results (OKRs) is provided. Enterprise executives and portfolio stakeholders define specific, differentiated business goals representing the Strategic Intent. These have a direct impact on budgets and guardrails. No other direct connection to EA has been identified, nor have any prescriptions for acquisition strategies.

Source: (author)

5. Discussion

The research that was conducted shows that SAFe considers architecture on all organizational levels, defines dedicated roles and has processes that align architecture with business needs. It aligns with the criticisms of SAFe for its top-down approach and the strong process emphasis (Remta et al., 2020). Additionally, it supports the findings that decentralization is a major issue with the implementation of existing EA frameworks in agile environments (Speckert et al., 2013). The top-down approach and rigorous processes enable centralization and architecture governance throughout the organization. Hence, EA practices can be applied. It contradicts one of the principles from the Agile Manifesto, stating, "The best architectures, requirements, and designs emerge from self-organizing teams" (Beck et al., 2001), and even here, it seems to support some of the SAFe criticisms for evolving into "the new waterfall" (Ebert & Paasivaara, 2017). Both SAFe and TOGAF are seen as process-heavy and documentation-driven, providing a huge set of templates, process elements, and roles (Ebert & Paasivaara, 2017; Gill et al., 2014), which seem to go against the very first value in the Agile Manifesto: “Individuals and interactions over processes and tools” (Beck et al., 2001). Yet, it eventually makes the frameworks easier to combine but harder to align with the basic ideas of Agile.

The Enterprise Architecture Process Maturity Levels Assessment has been conducted for SAFe. The assessment revealed that even a robust framework such as SAFe couldn't provide enough support to fully reach the second level of EA processes maturity, as described in TOGAF's AMM example. According to the assessment results, SAFe can support EA processes' strongest maturity for elements of Operating unit participation, architecture governance, and architecture communication. This supports the reported SAFe's benefits, such as bringing developers and architects closer together (Canat et al., 2018). However, SAFe seems only partially cover IT security and IT investment and acquisition strategy processes. The relatively low supported maturity levels relate to TOGAF's requirements for extensive documents, including different models, relationship diagrams, etc. SAFe doesn't specifically dictate the documents' format and advocates for introducing "just enough" architecture concepts. The findings support theories that EA frameworks do not work in current organizations' environments (Kotusev, 2019) and question the assumption that existing EA frameworks can be used or adopted off-the-shelf for any specific organization (Gill et al., 2014). Enterprises attempting to benefit from EA concepts can’t rely on the processes that are prescribed by SAFe. Our research shows that reaching higher levels of EA maturity can’t be achieved by just following the SAFe by the book but will require a combination with different frameworks, like TOGAF. Reaching higher EA maturity levels will require i.e. more emphasis on the documentation of architectural processes, the introduction of a standard profile framework, the integration of additional standards for IT security, and the inclusion of acquisition strategies into SAFe. Generally, it seems to go beyond SAFe's focus, and thus, we consider SAFe's explicit declaration of enterprise architecture support as only partially true.

Conclusion. We addressed the research question, “To what level is EA discipline covered in the Scaled Agile Framework (SAFe)”?, using TOGAF as a reference for EA best practices. The assessment was conducted using the US DoC ACMM Framework (Version 1.2) (United States Department of Commerce, 2007). Findings indicate that while SAFe supports initial levels of EA process maturity, full alignment with TOGAF EA is not confirmed. To achieve higher EA maturity levels alongside SAFe, organizations may need to blend this agile scaling framework with other established standards.

Limitations and further research. Additional empirical research is to be conducted to validate the presented results. In the SAFe analysis and maturity assessment, various SAFe configurations were not considered, and the results were generalized to match the available description of Portfolio SAFe. The maturity results were obtained by using the US DoC ACMM Framework. Thus it is possible that using a different maturity model can lead to different maturity results. Additional research comparing the results obtained through various maturity models will increase the reliability of the assessed maturity of EA processes in SAFe.

Acknowledgment

This work was supported by an internal grant funding scheme (F4/35/2022) administered by the Prague University of Economics and Business.

References

Alqudah, M., & Razali, R. (2016). A Review of Scaling Agile Methods in Large Software Development. International Journal on Advanced Science, Engineering and Information Technology, 6(6).

Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Martin, R. C., Mellor, S., Schwaber, K., Sutherland, J., & Thomas, D. (2001). Manifesto for Agile Software Development. https://agilemanifesto.org/

Bellman, B., & Griesi, K. (2015). Enterprise architecture advances in technical communication. IEEE International Professional Communication Conference (IPCC), 1–5.

Boehm, B., & Turner, R. (2005). Management Challenges to Implementing Agile Processes in Traditional Development Organizations. IEEE Software, 22(5), 30–39.

Buckl, S., Matthes, F., Monahov, I., Roth, S., Schulz, C., & Schweda, C. M. (2011). Towards an agile design of the enterprise architecture management function. Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOC, 322–329.

Canat, M., Català, N. P., Jourkovski, A., Petrov, S., Wellme, M., & Lagerström, R. (2018). Enterprise architecture and agile development: Friends or foes? 22nd International Enterprise Distributed Object Computing Workshop (EDOCW), 176–183.

digital.ai. (2022). 15th State of Agile Report.

Dikert, K., Paasivaara, M., & Lassenius, C. (2016). Challenges and success factors for large-scale agile transformations : A systematic literature review. The Journal of Systems & Software, 119, 87–108.

Ebert, C., & Paasivaara, M. (2017). Scaling Agile. In IEEE Software (Vol. 34, Issue 6).

Fallmyr, T., & Bygstad, B. (2014). Enterprise architecture practice and organizational agility; An exploratory study. Proceedings of the Annual Hawaii International Conference on System Sciences, January, 3788–3797.

Gill, A. Q., Smith, S., Beydoun, G., & Sugumaran, V. (2014). Agile enterprise architecture: A case of a cloud technology-enabled government enterprise transformation. Proceedings - Pacific Asia Conference on Information Systems, PACIS 2014.

Hanschke, S. (2009). Strategic IT management: a toolkit for enterprise architecture management. Springer Science & Business Media.

Hanschke, S., Ernsting, J., & Kuchen, H. (2015). Integrating agile software development and enterprise architecture management. Proceedings of the Annual Hawaii International Conference on System Sciences, 2015-March, 4099–4108.

ISO/IEC/IEEE. (2011). ISO/IEC/IEEE 42010 Systems and software engineering -- Architecture description.

Kotusev, S. (2019). Enterprise architecture and enterprise architecture artifacts: Questioning the old concept in light of new findings. Journal of Information Technology, 34(2), 102–128.

Lankhorst, M. M., Proper, H. A., & Jonkers, H. (2010). The anatomy of the ArchiMate language. International Journal of Information System Modeling and Design, 1(1), 1–32.

Li, W., Liu, K., Belitski, M., Ghobadian, A., & O'Regan, N. (2016). e-Leadership through strategic alignment: An empirical study of small- and medium-sized enterprises in the digital age. Journal of Information Technology, 31(2), 185–206.

NASCIO. (2003). Enterprise Architecture Maturity Model. https://www.nascio.org/resource-center/resources/enterprise-architecture-maturity-model/

OMG. (2018). The Open Group The TOGAF® Standard, Version 9.2. https://pubs.opengroup.org/architecture/togaf9-doc/arch/

Paasivaara, M., & Lassenius, C. (2016). Scaling scrum in a large globally distributed organization: A case study. Proceedings - 11th IEEE International Conference on Global Software Engineering, ICGSE 2016.

Proença, D., & Borbinha, J. (2017). Enterprise architecture: A maturity model based on TOGAF ADM. 19th Conference on Business Informatics (CBI).

Remta, D., Doležel, M., & Buchalcevová, A. (2020). Exploring the product owner role within safe implementation in a multinational enterprise. In Lecture Notes in Business Information Processing: Vol. 396 LNBIP.

Rost, D., Weitzel, B., Naab, M., Lenhart, T., & Schmitt, H. (2015). Distilling Best Practices for Agile Development from Architecture Methodology: Experiences from Industrial Application. Software Architecture: 9th European Conference, ECSA 2015.

Scaled Agile. (2020). Scaled Agile Framework 5. https://v5.scaledagileframework.com/#

Scaled Agile. (2023). Scaled Agile Framework 6.0. https://www.scaledagileframework.com/

Speckert, T., Rychkova, I., Zdravkovic, J., & Nurcan, S. (2013). On the changing role of enterprise architecture in decentralized environments: State of the art. 17th IEEE International Enterprise Distributed Object Computing Conference Workshops, 310–318.

United States Department of Commerce. (2007). Enterprise Architecture Capability Maturity Model. 1–16.

Winter, R., & Fischer, R. (2006). Essential layers, artifacts, and dependencies of enterprise architecture. Proceedings - 2006 10th IEEE International Enterprise Distributed Object Computing Conference Workshops, EDOCW2006, January 2006, 30–38.

INCORPORATION OF AI INTO ORGANISATIONAL PERFORMANCE EVALUATION

Adéla Šrámková

Czech University of Life Sciences Prague
Faculty of Economics and Management
sramkovaad@pef.czu.cz

DOI: 10.35011/IDIMT-2024-477

Keywords

AI; performance; evaluation; controlling.

Abstract

This paper delves into the development of a sophisticated control mechanism for evaluating corporate performance, addressing the complexities of the modern business landscape marked by rapid technological advancements. Acknowledging a gap in traditional performance evaluation methods, which often neglect non-financial aspects such as innovation and CSR engagements, the author proposes a comprehensive mechanism. This tool integrates both financial and non-financial indicators, employing a scoring system to enable nuanced evaluations across organizations. The research highlights the potential of Artificial Intelligence (AI) to enhance this mechanism, leveraging AI's capacity for processing large datasets and providing predictive insights. This integration aims to improve the accuracy and efficiency of performance evaluations, facilitating strategic planning and continuous improvement. By combining a wide array of indicators with the capabilities of AI, the proposed mechanism offers a dynamic approach to navigating the complexities of the global business environment, ensuring companies can effectively manage and evaluate their performance in a rapidly evolving world. The main aim of this paper is to discuss the usage of AI in performance evaluation.

1. Introduction

In today's fast-paced and complex business world, it is crucial to have reliable methods to manage and assess how well organizations are performing. One of the reasons is that the business landscape is constantly being reshaped by globalization, environmental concerns, economic instability and other challenges. These challenges have grown due to recent events like the COVID-19 pandemic, political unrest, and unpredictable energy costs, pushing companies to monitor their spending and be cautious about their business carefully. Furthermore, rising inflation and the need to minimize risks have made it important for companies to manage their money wisely and choose their partners carefully.

The dynamic nature of today's global market, characterized by rapid technological advancements and shifting consumer expectations, further complicates the task of performance management. Companies must now navigate through a maze of digital transformation, cybersecurity threats, and the need for speed in innovation to stay competitive (Rehman et al., 2021). Moreover, the rise of remote work and the digital economy has transformed traditional business models, necessitating a reevaluation of operational strategies and workforce management. These developments underscore the importance of adaptable and forward-thinking management practices that can swiftly respond to new market realities and regulatory environments (Jáčová, 2012).

Combined with performance pressures, there is a growing focus on Corporate Social Responsibility (CSR), with companies aiming to reduce their negative impact on society and the environment. This shift improves a company's image and financial success (Fatima & Elbanna, 2023). Studies suggest that firms committed to CSR attract more investors, especially younger ones who value sustainability and ethical practices (Sharma & Kiran, 2013). Non-financial factors, such as customer satisfaction, employee well-being, and innovation, are also increasingly important (Hernandez, 2022). These factors provide a detailed picture of a company's health. Tools like the Balanced Scorecard and ESG (Environmental, Social, and Governance) criteria are becoming essential for understanding overall performance, considering both profits and societal impact (Gazi et al., 2022).

The evolving landscape necessitates a more integrated approach to performance management, where financial outcomes and societal impact are considered in tandem. This integrated perspective not only aligns with the growing ethical and environmental consciousness among consumers and investors but also supports long-term sustainability and profitability (Aly and Mansour, 2017). Companies that effectively harness the power of comprehensive performance evaluation tools and CSR initiatives are better positioned to build resilient and adaptable business models that can withstand and thrive in the face of global challenges (Nazri et al., 2020).

The recent boom in Artificial Intelligence (AI) introduces new opportunities to enhance performance management (Johnston and Cortez, 2024). AI's ability to process and analyze vast amounts of data helps companies make smarter decisions, enhancing operational efficiency and accelerating innovation. AI systems identify trends and insights imperceptible to human analysts, offering predictions that inform strategic planning and risk management (Caruso et al., 2023). Additionally, AI can automate routine tasks, freeing human resources to focus on more complex and creative business aspects. This shift towards AI-driven processes allows companies to respond dynamically to market changes and customer behavior, ensuring competitiveness (Sahoo, 2024).

The aim of this paper is to discuss the incorporation of AI into the performance evaluation mechanism, which is part of the author's broader research. As this paper is submitted to a conference section dedicated to doctoral students presenting early career results, it also outlines the author's wider research context, providing a comprehensive background for integrating AI technologies.

2. Current research

2.1 Problem definition

In the rapidly evolving business landscape characterized by continuous global challenges and technological advancements, a research gap has been identified in performance evaluation mechanisms. Traditional methods don't always cover all the important factors needed to fully understand a company's health and success, especially now that both money-related and other kinds of measures are considered crucial. This research focuses on this gap by developing a nuanced control mechanism that leverages publicly available data to offer a more holistic and adaptable approach to evaluating organizational performance. By integrating a wide spectrum of indicators, this research aims to provide a tool that is not only robust in facilitating internal and external assessments but also flexible enough to accommodate the diverse needs of modern businesses. This endeavor addresses a pressing need for methodologies that encapsulate the multifaceted nature of contemporary corporate performance, marking a significant step towards enriching the existing body of knowledge and offering practical solutions for businesses navigating the complexities of today's economy.

2.2 Research objectives

The goal of this research is to develop a controlling mechanism dedicated to evaluating an organization's performance through publicly accessible data. In this context of this research, the "performance" refers a comprehensive assessment of an organization's effectiveness, efficiency, and sustainability in achieving its strategic goals. It encompasses financial metrics such as profitability and liquidity, operational efficiency metrics, social impact indicators related to CSR activities, environmental responsibility metrics, innovation metrics, and strategic metrics. The unique value of this mechanism will be the special focus on leveraging the publicly available data and on the extended scope of involved indicators.

The above-described endeavor seeks to provide a multifaceted tool for internal and external evaluative purposes. For external stakeholders, this mechanism aims to provide a framework for assessing competitive landscapes, facilitating comparisons with peers, and conducting preliminary analyses on potential clients, partners, or acquisition targets. Internally, the mechanism offers organizations a mirror, reflecting their operational footprint in the external world and enabling consistent year-over-year performance assessments.

To achieve this, the author proposes to integrate a diverse array of indicators, including both financial metrics, like profit and EBITDA, and non-financial metrics, including employee count, engagement in corporate social responsibility (CSR) initiatives, and adherence to recognized standards. The methodology central to this mechanism is a scoring approach, which clusters these indicators based on their relevance and assigns them predetermined weights. This structure allows for a nuanced evaluation, facilitating direct comparisons across different organizations. The scoring system, along with the weighting of indicators, will be accurately crafted based on a thorough review of existing methodologies and refined through consultations with domain experts in organizational performance assessment. This scoring mechanism is supposed to be adaptable, serving as a valuable tool for future research across various domains. Furthermore, while the mechanism is tailored to the context of European Union-based companies, it retains partial applicability to organizations outside this geographical scope.

2.3 Research Questions

In pursuit of developing this control mechanism, the research is guided by four research questions:

- RQ1: What are the main financial indicators for the organizational performance evaluation?
- RQ2: What should be measured as part of the non-financial performance evaluation, and what are the main non-financial indicators?
- RQ3: Which factors are primarily impacting the organizational performance?
- RQ4: How can financial and non-financial indicators be combined to ensure the most precise evaluation of organizational performance?
- RQ5: How might the performance evaluation framework benefit from AI technologies?

2.4 Research methodology

This research begins with a comprehensive literature review to identify existing gaps, setting the stage for the study's objectives. This phase establishes the foundation for the research.

After identifying the research gap, the study gathers publicly accessible indicators for performance evaluation through a thorough review of existing literature. This includes financial and non-financial indicators relevant to organizational performance, focusing on business management, controlling practices, and metrics for measuring success. Next, desk research collects data from public resources, such as financial statement disclosures and specialized tools like Bloomberg, comparing data from international sources like the German Handelsregister and Spanish Registro Mercantil.

The study then uses an analysis-synthesis approach to categorize collected indicators into clusters based on specific criteria. Exploratory Factor Analysis (EFA) and Multiple Factor Analysis (MFA) will identify primary factors influencing organizational performance. Additional statistical methods will be selected later.

Challenges in data collection, such as inaccessible data and converting descriptive information into quantitative formats, will be addressed. The methodology includes strategies for standardizing data formats for uniform comparisons.

The mechanism will be developed in Excel, offering structured platforms with graphical dashboards, supporting data export/import functions, and accommodating multiple companies. The study will explore the mechanism's applications across different organizational contexts and industry sectors.

The export will involve also exporting of the results (data) into different formats to ensure further usage and interoperability (e.g. with AI solutions) since the Excel format is not generally suitable for further processing.

2.5 Research results

The author of this research so far focused on three areas: (i) the relationship between financial and non-financial performance, especially financial performance and corporate social responsibility (CSR), (ii) financial performance indicators, and (iii) non-financial performance indicators.

The research findings were presented at the Hradec Economic Days 2024 conference. The study critically reviewed published studies on the relationship between corporate financial performance and CSR across various industries and locations. It examined the studies' research targets, samples, periods, methodologies, and results, revealing the nuanced interplay between CSR engagements and financial outcomes.

Building on this analysis, the author reviewed non-financial performance indicators, reviewing current methodologies and classification systems. The research identified trends, gaps, and inconsistencies in non-financial performance evaluation, proposing a refined model that better aligns with modern organizational priorities. Table 1 categorizes non-financial indicators with examples such as customer satisfaction, employee retention, and environmental policy.

Table 1. Categories of non-financial indicators and their examples

Category	Exemplary indicators
Customer	Customer satisfaction, Customer retention rate, Increase of number of consumers, Consumers loyalty
Employee	Employee turnover rate, Employee retention, Level of employee engagement, Employee Turnover
Environment	Environmentally friendly production, Protection of the natural resources, Environmental policy
Innovation	Openness to innovation, Adoption of new technology, Number of innovations
Internal process	Process quality, Capacity utilization, Protection of intellectual property, Internal process performance
Learning	Number of trained employees, Professional training, Hours of employee training per employee

Category	Exemplary indicators
Product	Product and service quality, Product Cycle Time, Development of new products / services
Sales	Market performance, Sales volume trend, Percent of returned orders, Sales performance
Strategy	Time to market, Marketing, Average percent of major brand names per store, Company reputation

In corporate finance, the author categorized indicators used by investors, analysts, and managers, including profitability (gross profit margin, return on assets), liquidity (current ratio, acid-test ratio), efficiency (asset turnover ratio), financial structure (debt-to-equity ratio), and market value (earnings per share, price-to-earnings ratio). Ensuring high data quality is crucial, especially since financial indicators can be calculated differently, and handling missing or partial data is essential.

These findings highlight the importance of both financial and non-financial indicators in assessing organizational performance. They emphasize the strategic value of integrating CSR into business operations, showing the complex nature of performance evaluation and the need for adaptive approaches. This research contributes to the academic and practical discourse on performance evaluation, providing a foundation for future research and implementation strategies.

4. Discussion on AI usage in Performance Evaluation

The author of this paper identified 6 areas where AI can be used as a part of the performance evaluation. Those areas are transferred into secondary research questions which extend the research questions RQ5:

- SRO1: How can AI support aggregation and analysis of data required for the evaluation?
- SRO2: How can AI help with the application of the scoring method within the evaluation process?
- SRO3: What is the role of AI in the management of exceptions in the evaluation?
- SRO4: How can AI predict future performance trends?
- SRO5: How can AI help with the presentation of the evaluation results?
- SRO6: How can AI facilitate continuous improvement of the entire mechanism?

In the pursuit of enhancing the effectiveness of controlling mechanisms for managing organizational performance, the integration of AI emerges as a revolutionary approach. AI's ability to process extensive datasets, recognize patterns, and predict outcomes can substantially improve both the internal and external evaluation processes of a company's performance. This chapter explores the diverse roles AI could play within the existing framework for performance evaluation, emphasizing the collection and synthesis of necessary information, the application of scoring methods, management of exceptions, and the derivation of insightful conclusions.

SRO1: Data Aggregation and Analysis

AI's primary role in performance evaluation lies in its capacity to aggregate and analyze publicly available data, a critical component for both internal and external evaluations. Utilizing sophisticated algorithms, AI might facilitate the automation of collecting financial and non-financial indicators from a variety of sources, including commercial registers, financial statements, and CSR reports. This automation not only speeds up the data collection process but also guarantees a comprehensive collection of indicators, capturing data that might be missed otherwise. Additionally, AI's capabilities enhance the ability to extract pertinent data from unstructured sources, thereby enriching the dataset available for analysis (Caruso et al., 2023; Shaik et al., 2023; Prasad et al., 2023).

Furthermore, AI's ability to process large volumes of data with precision significantly reduces the risk of human error. By leveraging machine learning techniques, AI can recognize patterns and trends within the data that might not be immediately apparent to human analysts. This enhances the quality of the data aggregation process, ensuring that the subsequent analysis is based on a robust dataset. Additionally, AI-driven data aggregation can incorporate real-time data streams, providing up-to-date insights that are crucial for dynamic business environments (Prasad et al., 2023).

Moreover, the integration of natural language processing enables AI to sift through textual data, such as news articles, social media posts, and press releases, to gather qualitative insights. This qualitative data can complement quantitative indicators, offering a more holistic view of an organization's performance. For example, sentiment analysis of customer reviews can provide insights into customer satisfaction that might not be captured by traditional metrics alone (Caruso et al., 2023).

SRO2: Application of Scoring Method

Following data collection, AI applies a predefined scoring method to systematically evaluate performance. This method involves categorizing indicators into clusters and assigning weights according to their relevance. AI's dynamic adjustment of weights and clusters in response to changing market trends and organizational priorities ensures that evaluations remain up-to-date and precise. This flexibility proves invaluable for customizing evaluations to meet the specific needs of different industries or regions, thereby increasing the versatility of the evaluation mechanism (Edgengton and Kasztelnik, 2024; Kovvuri et al., 2023; Sahoo et al., 2024).

The scoring method's adaptability is particularly beneficial in volatile markets where performance indicators can shift rapidly. AI's capacity to continuously learn from new data inputs allows it to recalibrate scoring models in real-time, maintaining the relevance and accuracy of the performance evaluations. For instance, during economic downturns, certain financial indicators may become more critical, prompting AI to adjust their weights accordingly (Edgengton and Kasztelnik, 2024).

Furthermore, the use of AI in scoring methodologies facilitates the inclusion of a wider range of indicators, encompassing both traditional financial metrics and modern ESG (Environmental, Social, and Governance) criteria. This holistic approach ensures that performance evaluations are comprehensive and reflective of an organization's multifaceted nature. AI's ability to integrate diverse data points into a coherent scoring system enhances the robustness of the evaluation framework (Kovvuri et al., 2023).

SRO3: Exception Management

AI demonstrates exceptional skill in managing data exceptions and anomalies, which are common due to the variability in public disclosures and reporting standards. By identifying outliers, incomplete datasets and using predictive modelling to address data gaps or signal the need for manual review, AI maintains the integrity of the performance evaluation process (Edgengton and Kasztelnik, 2024).

In addition to identifying and managing anomalies, AI can also automate the resolution of common data issues, such as duplicate records or inconsistent formats. Machine learning algorithms can learn from historical data corrections to apply similar adjustments autonomously, reducing the need for manual intervention. This automation not only streamlines the data management process but also ensures a higher level of consistency and reliability in the data used for performance evaluations (Edgengton and Kasztelnik, 2024).

Moreover, AI can leverage advanced techniques like imputation to estimate missing values, ensuring that incomplete datasets do not compromise the accuracy of assessments. Predictive modeling can also anticipate potential future data anomalies based on historical patterns, allowing organizations to proactively address these issues before they impact the evaluation results (Sahoo et al., 2024).

SRO4: Predictive Insights and Scenario Analysis

AI's capabilities extend into offering predictive insights into future performance trends, analyzing historical data alongside current performance metrics to forecast potential changes in organizational performance. This foresight is crucial for strategic planning. Furthermore, AI supports scenario analysis, simulating the effects of various strategic decisions on performance indicators to facilitate more informed decision-making (Johnston and Cortez, 2024; Kovvuri et al., 2023; Shaik et al., 2023).

Predictive analytics driven by AI can identify leading indicators that signal future performance shifts, enabling organizations to take proactive measures. For example, if predictive models indicate a potential decline in customer satisfaction, companies can implement corrective actions before the negative trend impacts overall performance. This anticipatory capability is a significant advantage in maintaining competitive edge (Johnston and Cortez, 2024).

In scenario analysis, AI can create detailed simulations of various business strategies, helping decision-makers understand the potential outcomes of their choices. By adjusting variables and parameters within these simulations, AI can provide insights into the most effective strategies under different conditions. This capacity for detailed scenario planning supports more resilient and flexible strategic planning processes (Shaik et al., 2023).

SRO5: Presentation of Evaluation Results

Another significant contribution of AI is in the presentation of evaluation results. AI can generate customized reports and dashboards that underscore key performance insights. AI allows for prioritizing metrics that are most relevant to stakeholders and crafting presentations that cater to specific concerns or interests (Jorzik et al., 2023; Kovvuri et al., 2023).

AI-driven visualization tools can create interactive dashboards that allow users to drill down into specific data points and trends, providing a more engaging and informative experience. These dashboards can be tailored to different audiences, ensuring that each stakeholder group receives the most pertinent information in a format that is easy to understand. For example, financial analysts might receive detailed financial metrics, while board members might view high-level strategic insights (Jorzik et al., 2023).

Moreover, AI can enhance the storytelling aspect of data presentation by highlighting key trends and insights, offering contextual explanations, and even suggesting potential actions based on the data. This narrative approach helps stakeholders quickly grasp the implications of the data and make informed decisions (Kovvuri et al., 2023).

SRO6: Facilitating Continuous Improvement

Incorporating AI into the controlling promotes continuous improvement. By evaluating the outcomes of past evaluations and subsequent organizational performance, AI identifies opportunities for refining the evaluation process. This could involve adjusting used indicators, the weights assigned, or the data sources consulted, ensuring the mechanism evolves in harmony with changing business environments and organizational goals (Edgengton and Kasztelnik, 2024; Jorzik et al., 2023).

AI's continuous learning capabilities enable it to adapt and improve over time. By analyzing the effectiveness of previous evaluations and the actual performance outcomes, AI can identify patterns and suggest refinements to the evaluation framework. This iterative process ensures that the performance evaluation mechanism remains relevant and effective in a dynamic business landscape (Edgengton and Kasztelnik, 2024).

Furthermore, AI can facilitate benchmarking against industry standards and best practices, providing organizations with insights into their relative performance and areas for improvement. By continuously comparing organizational performance with peers, AI helps organizations strive for excellence and adopt best practices that enhance their competitive position (Jorzik et al., 2023).

5. Conclusion

Integrating AI within controlling mechanisms for organizational performance management presents a comprehensive approach that supports accuracy, efficiency, and strategic value. From streamlining data collection to enhancing evaluation methodologies and offering predictive insights, AI's inclusion promises to transform the controlling mechanism into a dynamic instrument. This instrument not only evaluates past and present performance but also directs future strategic decisions, representing a significant leap forward in performance management technology. From the technology perspective, the selection of an appropriate AI (type of) solution is very crucial for the successful onboarding of AI into the performance evaluation framework, and this exactly will be a subject of future steps within the research. It will also be beneficial to leave this decision to the future since the pace of development of AI technologies has been extremely rapid recently.

References

Aly, A.H. and Mansour, M.E. (2017) 'Evaluating the sustainable performance of corporate boards: the balanced scorecard approach', Managerial Auditing Journal, 32(2), pp. 167–195. Available at: https://doi.org/10.1108/MAJ-04-2016-1358.

Caruso, Simone et al. Artificial intelligence to counteract "KPI overload" in business process monitoring: the case of anti-corruption in public organizations. Business process management journal. 2023, vol. 29, no. 4, pp. 1227–1248.

Edgeington, Sean and Karina Kasztelnik. The Study on the Impact of Business Artificial Intelligence Innovation on Fair Value Investments in the United States. The journal of applied business and economics. 2024, vol. 25, no. 7

Fatima, T., & Elbanna, S. (2023). Corporate Social Responsibility (CSR) Implementation: A Review and a Research Agenda Towards an Integrative Framework. Journal of Business Ethics, 183(1), 105–121. https://doi.org/10.1007/s10551-022-05047-8

Gazi, F., Atan, T. and Kılıç, M. (2022) 'The Assessment of Internal Indicators on The Balanced Scorecard Measures of Sustainability', Sustainability, 14(14), p. 8595. Available at: https://doi.org/10.3390/su14148595.

Hernandez, D. A. (2022). Relationships Between Corporate Financial Performance and Corporate Social Responsibility When Controlling For Socially And Environmentally Conscious Investments. The Economix, 1.

Jáčová, H. (2012) 'Current tools and new trends in Enterprise Performance Measurement', XVIII (2/2012/Issue B)(2), pp. 55–63.

Johnston, Wesley J. and Roberto M. Cortez. Business-to-business digitalization, artificial intelligence, and social action. Journal of business research. 2024, vol. 172, p. 113952.

Jorzik, Philip et al. Artificial Intelligence-Enabled Business Model Innovation: Competencies and Roles of Top Management. IEEE transactions on engineering management. 2023, pp. 1–13.

Kovvuri, Veera R. R. et al. Fund performance evaluation with explainable artificial intelligence. Finance research letters. 2023, vol. 58, p. 104419.

Nazri, M. A., Omar, N. A., Aman, A., Ayob, A. H., & Ramli, N. A. (2020). Corporate Social Responsibility and Business Performance in Takaful Agencies: The Moderating Role of Objective Environment. Sustainability, 12(20), 8291.

Prasad, Sandya, Prasanna Kulkarni and Veronika. Role of artificial intelligence in business process transformation. AIP Conference Proceedings. 2023, vol. 2736, no. 1

Rehman, S.U., Kraus, S., Shah, S.A., Khanin, D. and Mahto, R.V. (2021) 'Analyzing the relationship between green innovation and environmental performance in large manufacturing firms', Technological Forecasting and Social Change, 163, p. 120481. Available at: https://doi.org/10.1016/j.techfore.2020.120481.

Sahoo, Saumyaranjan et al. Artificial intelligence capabilities, open innovation, and business performance – Empirical insights from multinational B2B companies. Industrial marketing management. 2024, vol. 117, pp. 28–41.

Shaik, Aqueeb S. et al. Artificial intelligence (AI)-driven strategic business model innovations in small- and medium-sized enterprises. Insights on technological and strategic enablers for carbon neutral businesses. Business strategy and the environment. 2023

Sharma, A., & Kiran, R. (2013). Corporate Social Responsibility: Driving Forces and Challenges. International Journal of Business Research and Development, 2(1). https://doi.org/10.24102/ijbrd.v2i1.182

EXPLORING EVIDENCE COLLECTION METHODS FOR THE INSPECTION OF CRITICAL INFRASTRUCTURE

Joel Klimont

Johannes Kepler University Linz
Institute of Networks and Security
klimont@ins.jku.at

DOI: 10.35011/IDIMT-2024-487

Keyword

forensics, security, evidence collection and storage, confidential computing, SGX

Abstract

When inspections are carried out on critical infrastructure, the operator may need to prove to the authorities at a later date that the inspections have taken place and that everything is in order. Inspections increasingly involve the use of digital devices that generate a large amount of sensor information. This data must be secured in order to qualify as trustworthy digital evidence. Two different approaches to the secure collection and storage of this forensic data are presented in this paper. One approach uses YubiKeys (SmartCards) to generate and store the necessary public/private key pairs, while the other uses SGX enclaves. By using SGX, it is possible to securely create public/private key pairs in such a way that even the operator performing the examination cannot access them. This increases the trustworthiness of the data, as the operator, or other bad actors, are unable to modify the recorded data without invalidating the signatures.

1. Introduction

In order for critical infrastructure to work reliably, it must be regularly maintained and inspected by humans. These inspections require humans to manually survey the infrastructure and identify potential problems or irregularities in the buildings and utility lines. As underground supply shafts can be very long, narrow, hot, cold or wet, performing these inspections is a tedious as well as repetitive task. Thus, automating these tasks has been of great interest, and several studies explored the possibility of automating these inspections using ground based-robots or UAV (unmanned aerial vehicles) (Wang & Yin, 2022; Lee, 2023; Jia et al., 2010; Myung et al., 2014). These projects aimed at reducing the overall cost and complexity, as well as speeding up the whole inspection process. While much thought has been given on which and how data is obtained (used sensors etc.) not much work has been performed into how the collected data can be transformed into reliable digital evidence. The INFRASPEC project aims at automating inspections in underground supply shafts using a ground-based robot and additionally focuses on the collection of reliable digital evidence.

The collected data from one inspection should not be changeable (without detection) after the inspection has been concluded. This is a necessary security requirement, as otherwise the inspectors could later change the data and claim that “everything was in order” when they performed the inspection or hide the fact that the inspection was performed poorly or incorrect. The result of an inspection should be a data package that contains all the data collected during the inspection and is unchangeable without detection, e.g., reliable digital evidence. This work is based on the threat model developed in (Sonntag & Schraml, 2023) for the inspection of critical infrastructures as part of the INFRASPEC project.

Generally, two different kinds of forensic analysis can be distinguished: “static analysis” and “live forensics”, as laid out in (Rafique et al., 2013). When performing static analysis, the system under examination is shut down and its logs, written files etc. are reviewed. During live forensics, the system is still running and information can be collected during its operation. This offers the possibility to record data that would otherwise not be present in static analysis, see (Adelstein, 2006). After the evidence has been obtained, it must be ensured that no changes to the data can be made and that a clear ‘chain of custody’ is maintained (Newman 2007).

2. System Design

The whole inspection system consists of at least two components: the robot and the ‘base-station’. The robot is the actual physical inspection device that navigates the supply shafts (either manually using a remote control or, in the future, autonomously) and collects the (forensic) data using its sensors (for example a Riegel VZ-400i 3D laser scanner). The base-station is the component that collects and securely stores all the recorded data from the robot. Additionally, it also collects and stores the network traffic between the robot and the human operator as forensic evidence. The system can also be extended by other components, for example, a second robot or other devices that collect forensic data.

2.1. Robot

The task of the robot is to survey the supply shafts and collect data for manual and automatic examination by the inspectors. For this purpose, the robot has multiple sensors installed, which are either connected to the ROS2 (Macenski et al., 2022) network or the CAN bus on the robot. ROS2 is at its core a message broker system used primarily in robotics. In the prototype, the robot is connected to the base-station using a Wi-Fi connection. This significantly simplifies the construction of the robot, as an antenna on the robot is enough to communicate with the base-station, even over a long distance. Cable-based systems could also be employed in the future and would offer a greater bandwidth; however, the robot would have to continually unroll a cable behind itself, wind up the cable when driving back, etc. The downside of the Wi-Fi connection is the lower bandwidth. To ensure a reliable and uninterrupted operation of the robot, steps have been taken to lower the bandwidth requirement of the software. For example, the forensic component on the robot only sends status information to the base-station during the inspection but no actual data. The upload of the forensic data to the base-station is executed after the inspection has concluded, when the robot is back at the base and doesn’t need to transmit any other data anymore.

Figure **Chyba! Nenalezen zdroj odkazů.** shows the flow of data on the robot. The robot collects the d ata from its CAN bus and from the ROS2 network. The CAN bus is used by the robot's motors and is recorded using a USB2CAN device attached to the bus. This way, all messages that are transferred on the bus are recorded. The other sensors (for example, the Vz-400i which performs 3D scans of the environment) are using ROS2. The ‘ROS2 subscriber’, seen in Figure **Chyba! Nenalezen zdroj o dkazů.**, subscribes to the relevant topics and records the data from the sensors. The robot also has a moving arm with an attached 3D camera that can be moved using ROS2. The recording node also captures this data, as the movement of the arm can be of interest later, for example, if the arm movement caused damage to the supply lines. The recorded data from both sources is then put into a messaging queue for processing.

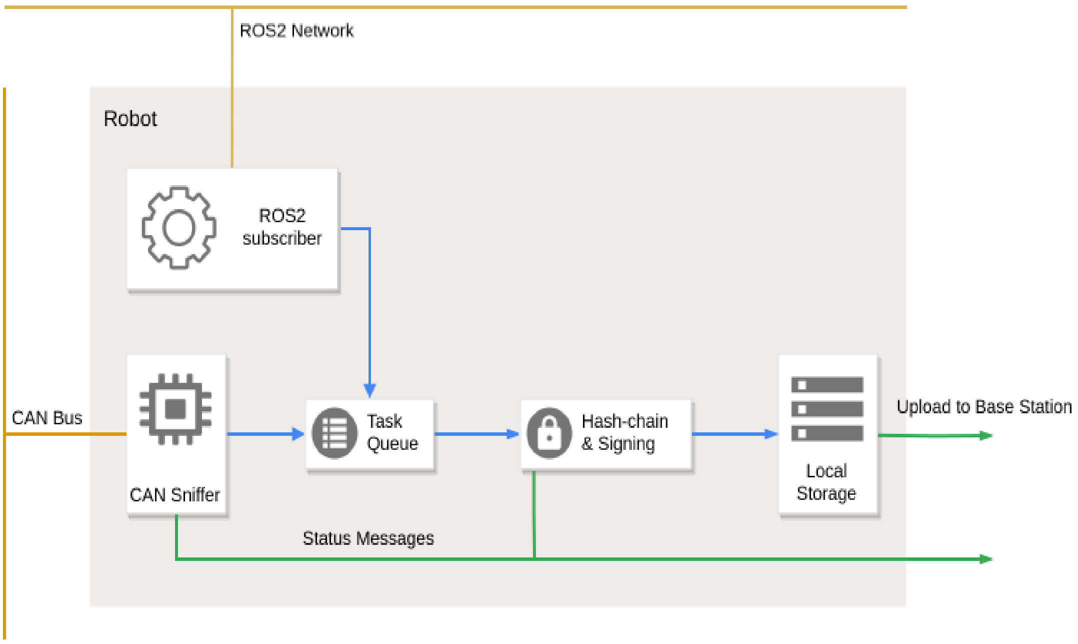


Figure 1. Data collection and temporary storage on the robot

2.2. Base-Station

The base-station component is responsible for collecting the data sent by the different data sources (for example, the robot). This component also has a web interface, in which the user can create new recordings (one recording per inspection), see the status messages of the individual components (health checks, amount of data collected, etc.), and initialize and finish the inspection process.

The incoming data is put into a task queue for processing. During the processing, the validity of the hash-chains, signatures, etc. will be verified. This is described in Section 3. After the inspection is complete, the user can export the project, which is a SQLite3 database containing the recorded data and the necessary signatures and public keys for verification.

3. Establishing Trust

In Section 2 an overview of the general system design was given. In essence, multiple forensic recording devices send data to the base-station, which collects and stores the data. The security requirements, as well as the threat model, are described in (Sonntag & Schraml, 2023). This section focuses on the technical implementation of the requirements, especially those concerning the secure transmission and storage of forensic data.

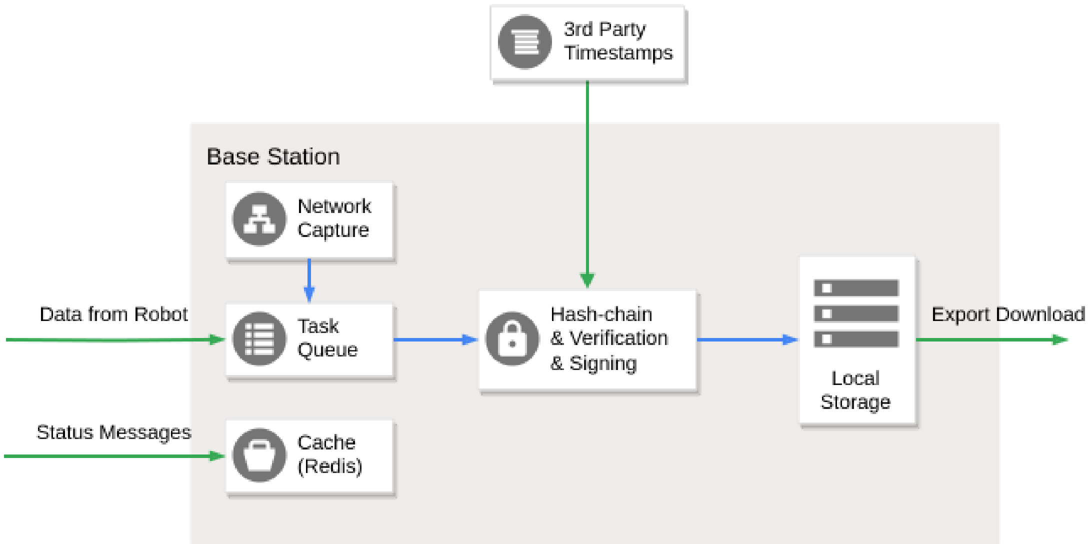


Figure 2. Data validation and signing on the base-station

3.1. Examining Requirements

The requirements listed below concern the secure transmission and storage of data and do not include other aspects that might be of relevance in other projects, including those related to what kind of data shoud be recorded or other additional security masures which are not within the scope of this paper.

- R1: A secure communication link needs to be established between the forensic components and the base-station.
- R2: The communication link must not be susceptible to MiT (man in the middle attacks).
- R3 *(R4* in (Sonntag & Schraml, 2023)*):* Inspections cannot simply be repeated or old inspections copied and tied to a new date.
- R4 (*R5*): When the application encounters an error, the error must be logged and included in the forensic recording (for example, temporary loss of communication).
- R5 *(R7)*: The recorded digital evidence must be unchangeable without detection.
- R6 *(R14)*: The recorded forensic data can be exported and act as 'digital evidence'.

When examining the requirements, we can map them to the five core security properties for information systems: "Confidentiality, Integrity, Availability, Authenticity and Non-repudiation", short *CIAAN*, and evaluate the current implementation according to the requirements and the *CIAAN* principle.

3.1.1. Availability (R4)

The availability property specifies that a bad actor should not be able to deny/interrupt the correct operation of the system. For example, they should be unable to prevent the robot from communicating with the base-station. While availability cannot be guaranteed, as a bad actor could use a jamming device to interrupt the communication between the components, such attacks can be detected and are attached to the forensic report, as demanded by (R4).

3.3.2. Authenticity (R6)

When data arrives at the base-station, we must be able to verify that it was recorded and sent by a trusted source. The approach that is implemented in the prototype for the INFRASPEC project uses YubiKeys (“Yubico Product Documentation”, 2024) to establish trust between the components. YubiKeys are Smart Cards that are able to generate, store and use public/private key pairs (in the implementation RSA 2048 keys are used). When starting a new inspection, the user navigates to the base-station’s web interface and the initialization process for the inspection begins. During that procedure, the user must plug in the YubiKey for the robot (and, if present, for the other data sources) as well as for the base-station. The software then generates a public/private key pair directly on each of the YubiKeys. The private key always remains directly on the YubiKey and cannot be extracted from there, while the public key and the name of the data source it belongs to (for example, ‘robot’) is saved to the database. Additionally, a third-party timestamp of the hash of all public keys is saved to the database as the first ‘data’ entry of the inspection. This hash is created by concatenating the base64 representation of the public keys and then calculating the hash of the resulting string. After the initialization, the YubiKeys are plugged into their respective components. The hash-chain value (see Section 3.5 Integrity) of the recorded forensic data is signed by the component’s YubiKey, and when the data is sent over to the base-station, the signature can be validated. This way, the base-station can assure that the data has been sent by a trusted source.

3.1.3. Confidentiality (R1, R2)

All transferred data over the network must be kept private and treated as a secret, as required by (R1) and (R2). A bad actor listening to the communication over the network should not be able to extract data gathered during the inspection, such as 3D scans or other sensor data. The transfer of data can be encrypted by storing the certificate of the base-station’s webserver on the YubiKey. The forensic components can use the certificate and mark it as trusted, which makes it possible to communicate securely via HTTPS.

3.1.4. Integrity (R5, R6)

The integrity of the collected evidence is ensured by a combination of multiple hash-chains and the signatures from the YubiKeys. Two different kinds of hash-chains exist, the ‘sub-hash-chains’ and the ‘main-hash-chain’. The sub-hash-chains exist on a per-component basis, and the main-hash-chain is created on the base-station and spans across all collected data.

When a new data point is recorded on a component, it stores the value in its local database, updates the sub-hash-chain value by including the hash of the new data, and then signs the resulting value using the attached YubiKey. Note that the hash-chain values are always updated in this fashion: *hash(hash_chain_value + hash(new_data)) = new_hash_chain_value* (where the ‘+’ operator is the string concatenation). The current implementation uses SHA256 hashes.

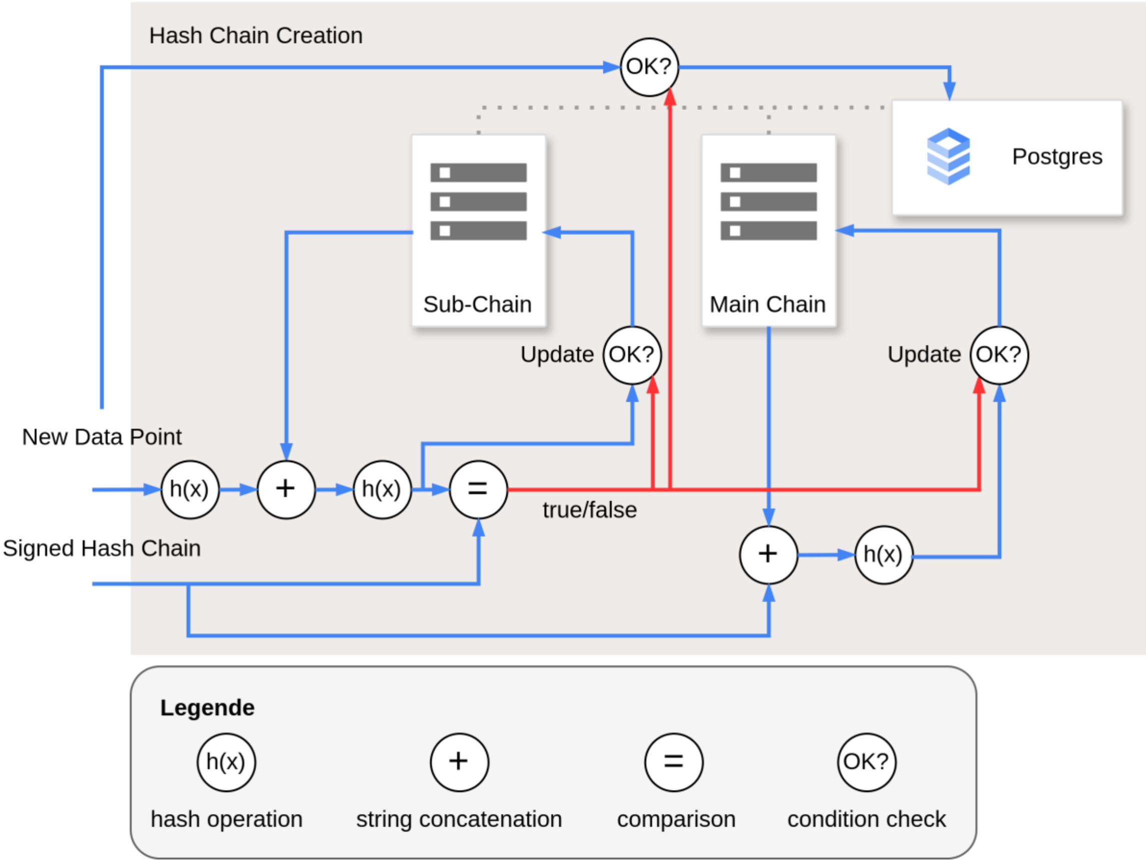


Figure 3. Validation of sub-hash-chain and creation of main-hash-chain upon upload of one data entry

After the inspection itself is finished, the components stop recording new data and start transferring their collected data to the base-station. For performance reasons, the data is always sent in batches, of a few MB. Along with the data, each batch contains a signature of the last sub-hash-chain value of the last entry in the batch. When the batch arrives at the base-station, it is processed as shown in Figure **Chyba! Nenalezen zdroj odkazů.**. The base-station also stores the sub-hash-chain value of e ach component and must 'catch up' when a new batch arrives. It does that by computing the sub-hash-chain for each data entry in the batch, until it arrives at the last data point in the batch. If the last computed sub-hash-chain value (the new tail of the sub-hash-chain) matches with the provided signed hash-chain value, the data has been successfully validated and can be trusted. The sub-hash-chain value is then attached to the main-hash-chain and the resulting new main-hash-chain value signed using the base-station's YubiKey. If the validation of either the signatures or the hash-chain is not successful. The error is logged in a special database table (which is also included in the final export) and is shown to the user.

3.1.5. Non-repudiation

To ensure non-repudiation, we must be able to detect if data is missing or has been intentionally denied by a bad actor. To some extent, this is already accomplished by the hash chain used to ensure integrity. If one batch is lost, the base-station can detect that as soon as the next batch arrives, as the sub-hash-chain values will not match up. However, if the last batch is lost, there is no next one for validation, and the absence of data would not be detected. To protect against this, the last batches must include a 'last_batch' value. This attribute is attached to the data upload of the last batch and is a random string of 64 characters. The value is also included in the sub-hash-chain, which in turn is signed by the YubiKey, as described in Section 3.5 (Integrity). This way, a bad actor cannot remove it from the last batch without detection. To finalize an inspection, the base-station demands that all forensic components send a last batch to indicate that all data has been successfully transferred.

After all components sent a last batch, the inspection can be finalized. An instruction is sent out to all components to overwrite their YubiKey. The base-station attaches one last value to the main-hash-chain that states ‘finalized’, then signs the resulting main-hash-chain value and also overwrites the contents of its YubiKey. The collected data is then added to one SQlite3 database (along with all the public keys, signatures, etc.) and made available to the user for download. The user is also provided with a Python script that can inspect the export and verify its integrity.

3.2. Key Deletion Problem

For the exported digital evidence to be truly unchangeable without detection, the private keys used to sign the hash-chain values must be deleted. The components and the base-station should overwrite the private keys on their YubiKey to prevent a bad actor from obtaining the physical devices and using it to rewrite parts of the digital evidence. However, securely deleting the private keys is a problem which remains in the current implementation. If the software correctly deletes the private key, it has no key anymore to sign a message that attests “that the key was deleted securely”. If the software first signs a message stating that its “next action will be to delete the private key”, a bad actor could still remove the physical YubiKey and stop the software from overwriting the key in between. Any messages sent after the key should have been deleted, would be unsigned either way and could not be trusted. This issue could be mitigated by using a third-party timestamp to sign the last main-hash-chain value. However, this solution would require access to the internet, which might not be available in the supply lines deep underground. In Section 4, we will present a different approach that uses SGX enclaves (Costan & Devadas, 2016), instead of YubiKeys, to sign the data and establish trust between the components.

3.1.6. Out of Scope Security Issue

Some security concerns remain but are considered out of scope. For example, the robot has a CAN-bus that is used to control its motors (e.g., movement). The forensic component on the robot has a USB2CAN device attached to it, which is plugged into the CAN-bus, and reads the messages transmitted on that bus. If the operator of the robot or the persons conducting the inspections have mal-intent, they could plug the USB2CAN device into a ‘false’ CAN-bus on which pre-recorded messages are published. There is no way in which the forensic component would notice that its sensor is plugged into the wrong CAN-bus and that it is reading spoofed data. These kinds of attacks are considered out of scope and cannot be prevented or detected by the current software. However, an analysis of the recorded data could give insights into how genuine the whole recording is. For example, if CAN-bus data is modified, this change could be detected by comparing the movement of the robot observed via the CAN-bus to the movement recorded by the other sensors, like the 3D camera mounted on the robot arm, etc. By combining that data and analyzing if the recorded values match overall, such attacks could be detected.

4. SGX Enclaves

Intel SGX enables the use of so called ‘enclaves’. Enclaves are special programs that can be executed in a hardware-protected manner. They are split into a ‘trusted’ and ‘untrusted’ part. The memory of the trusted part is protected by hardware on a CPU level, and SGX ensures that the memory is always encrypted and cannot be accessed by other applications. SGX also provides the functionality necessary to perform remote attestation, in which one program can verify the TCB (trusted computing base) of a running enclave remotely. Another useful feature is the ‘sealing’ process, in which secrets can be encrypted in such a way that they can only be accessed by a certain correctly initialized enclave.

4.1. Using Enclaves instead of YubiKeys

By replacing the YubiKeys with SGX enclaves, the signing process would happen inside the trusted part of the enclave. When starting the enclave, the trusted part generates a public and private key pair. The public key is relayed to the base-station and saved in the project just as before. The private key remains in the enclave's protected memory region and can be used to sign incoming data. By adding the last-batch value (generated inside the enclave) and singing the last sub-hash-chain tail, the enclave indicates that its next action is going to be deleting its private key. After the private key has been overwritten in memory, no new messages can be signed, and the enclave shuts down. Note that Intel SGX protects against cold boot attacks and similar attacks, and that a bad actor would only be able to extract the encrypted enclave memory. Even sniffing the bus between the CPU and the memory would not yield any protected data, as all data is already encrypted when it leaves the bounds of the CPU.

4.2. Ensuring Authenticity

The YubiKeys were not only used to simply have a signature, but also to attest the authenticity of the recorded data. Only a device with a registered YubiKey could create the appropriate signatures that the base-station would trust. Generating a public/private key pair, as mentioned in Section 4.1 is not enough to provide authenticity for the data. After all, a normal program can also generate a public/private key pair and start signing the data. It follows that the base-station must be able to verify that the public/private key pair was securely generated inside a trusted enclave and can’t be extracted by a bad actor. Usually, the remote attestation procedure is used to verify that a certain enclave is securely running on trusted SGX hardware. During that procedure, the Quoting Enclave (QE) is used to generate a verifiable ‘QUOTE’ from the ‘REPORTs’ of other enclaves. A report is a measurement of the whole code and data inside the enclave, as well as the other attributes of the running enclave instance. However, in order to perform remote attestation, internet access is required, as the Intel Attestation Service is used to verify quotes (Johnson et al., 2016).

Instead of the usual remote attestation procedure, we propose another technique of creating trust between the components without the use of Intel’s online services. SGX features a technique called ‘sealing’. Sealing can be used to persist enclave data in an encrypted format, such that only instances of the same enclave can access the sealed data. The data can either be sealed using a key derived with the *MRENCLAVE* value or one that depends on the *MRSIGNER* value. When the data is sealed to the *MRENCLAVE* value, only an enclave with the same code and security version number (SVN) can access the data. If the code of the enclave is changed, the enclave won’t be able to access the sealed data. When using the *MRSIGNER* value to derive the sealing key, enclaves signed by the same author will be able to access the sealed data.

A trusted third party (TTP) can be used to initialize the enclaves and generate the keys necessary for signing the message during an inspection. When the base-station enclave is started for the first time, it generates a public and private key pair inside the enclave. The TTP retrieves the public key from within the enclave and signs it using its own secret private key, which generates a certificate that is passed back into the enclave together with the public key of the TTP. The enclave then proceeds to seal its own private/public key pair, the public key of the TTP and the certificate received from the TTP.

The enclaves used to sign the recorded forensic data (for example, on the robot) also need to be initialized by the TTP. Similarly, the enclave generates a public/private key pair and passes the public key to the TTP in order to get the certificate. Additionally, the certificate for the base-station is passed into the enclave by the TTP. The enclave then seals its own public/private key pair and both certificates.

When the application is started during the inspection, the base-station starts by unsealing its secrets. This unsealing will only be successful, if the running enclave has the same *MRENCLAVE* value (meaning that the enclave consists of the exact same code as the enclave that was used to seal the data), or if the other technique is chosen, the same *MRSIGNER* value. The enclave then uses the certificate and private key to secure the webserver running inside the enclave (used by the forensic components to upload their recorded data, etc.). The forensic components also unseal their data, start adding incoming data to their sub-hash-chain and sign the values using their private keys. When the forensic component uploads its collected data to the base-station, it is able to establish a secure HTTPS connection from enclave to enclave by verifying the provided certificate by the base-station using the TTP's public key. On the other side, the base-station is able to verify the signatures of the forensic components by examining the certificate that contains the TTP's signature of the component's public key. After verifying the signatures (authenticity) and the sub-hash-chain values (integrity), the base-station can update the main-hash-chain and sign the resulting value with its own private key. When finishing the inspection, the base-station includes the necessary public keys and certificates in the export, then updates and signs the main-hash-chain one last time.

4.3. Replay Attacks

The requirement (R3) mentions the need to prevent replay attacks. If a serious problem is discovered during the inspection, the operator might want to cover it up (see threat model in Sonntag & Schraml, 2023). After quickly fixing the issue, the operator could redo the inspection to generate the digital evidence again, but this time without the problem. These attacks can be prevented by using the TTP to 'sign off' finished inspections. As the TTP signs the certificates passed to the enclaves, which are also included in the forensic data export of an inspection, the TTP could set a certain time for these certificates to expire. After the inspection, the exported digital evidence can be uploaded to the TTP, which then verifies the signatures, validates hash-chains etc. Finally, the TTP checks the expiry date of the certificates it created for all the enclaves public/private key pairs. If the certificates have not expired yet, it signs the last main-hash-chain value with its own private key. Note that the public key of the TTP must be signed by a trusted CA (certificate authority) for the digital evidence to be trustworthy.

4.4. Timestamps

As already mentioned throughout Section 2 (System Design), the robot and the base-station add 'timestamps' to the recorded data points. In the current prototype (using the YubiKeys), timestamps of a third party are used if internet is available (*openssl* is used to generate a *.tsq* file which is then signed by freetsa.org), otherwise the system time is used, and the signature comes from the YubiKey. Using SGX enclaves, this process can be made more secure by obtaining the time from the SGX SDK function 'sgx_get_trusted_time()' (Cen & Zhang, 2017). This function utilizes the Protected Real-Time Clock (PTRC) from the Intel Converged Security and Management Engine (CSME). Using this clock, it is possible to get the time relative to the first function call. This means, the timestamps indicate how much time has passed since the inspection was started. Combining this functionality with the mechanisms laid out in Section 4.3 (valid time of certificates), we are able to tie the forensic evidence to a certain global timeframe (using the certificates and the final signature of the TTP) and a relative time range (how long the inspection took).

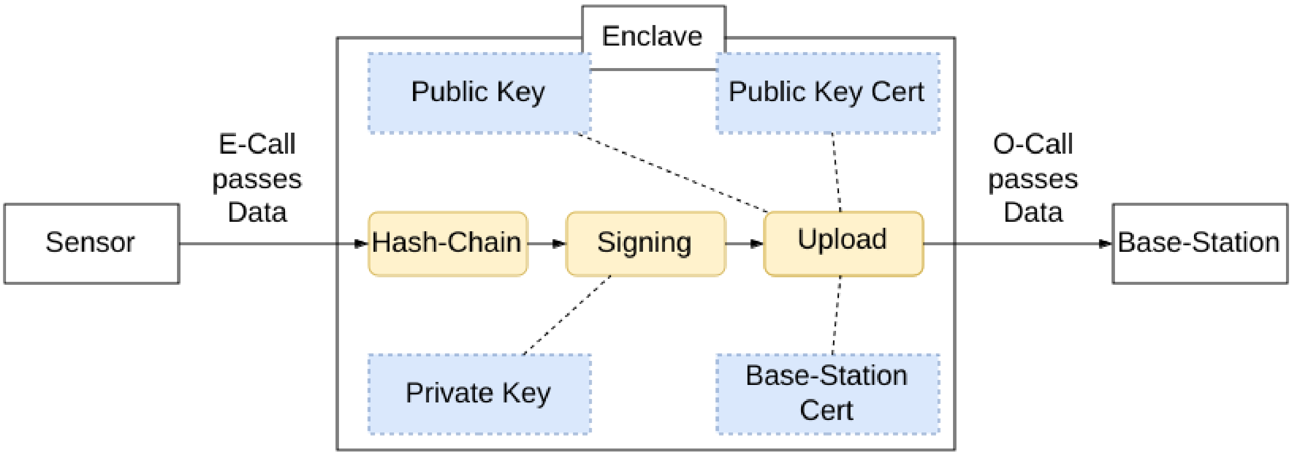


Figure 4. Enclave for the forensic components (for example the robot)

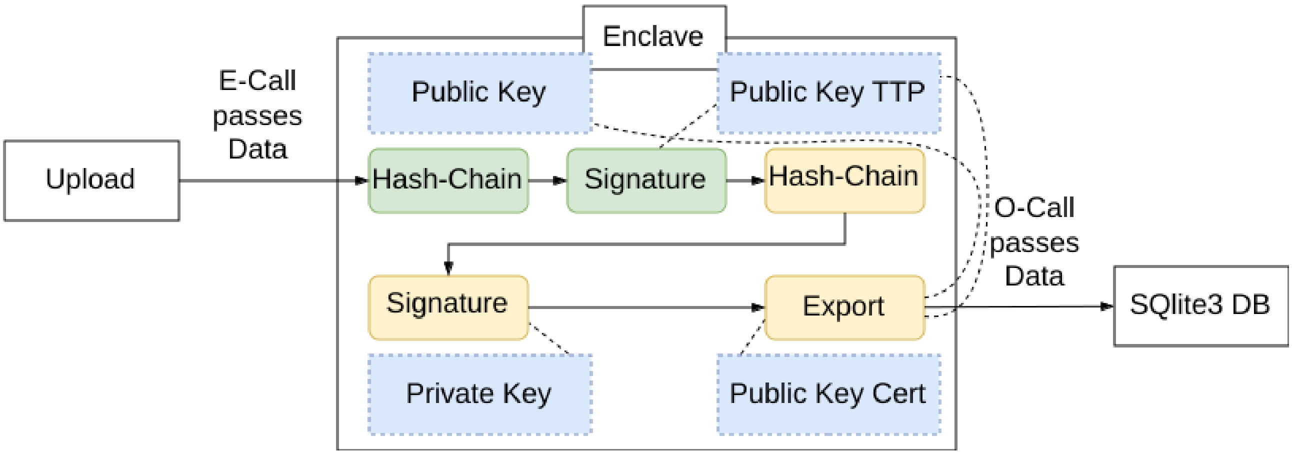


Figure 5. Enclave for the base-station

5. Summary

We have examined the current state of the INFRASPEC implementation with respect to the CIAAN (Confidentiality, Integrity, Availability, Authenticity and Non-repudiation) properties for information systems. The current prototype fulfils these properties and the laid out requirements only if the person performing the inspection is trusted to correctly setup the base-station, the robot, etc. This paper suggests an improvement to the current implementation. The YubiKeys, which are used to create the signatures for the forensic data, may be replaced by SGX enclaves provided by a TTP (trusted third party). Using these enclaves, it is possible to generate temper-proof, reliable digital evidence that cannot be altered without detection after its creation with only minimal trust placed in the inspectors.

Acknowledgement

This publication and the system described in it was developed within the scope of the project INFRASPEC, which is funded by the Austrian security research program KIRAS of the Federal Ministry of Finance (BMF).

References

Sonntag, M., & Schraml, S. (2023). An Evidence Collection System for Robot-Supported Inspection of Critical Infrastructure. In IDIMT-2023 New Challenges for ICT and Management: 31st Interdisciplinary Information Management Talks (pp. 51-62).

Rafique, M., & Khan, M. N. A. (2013). Exploring static and live digital forensics: Methods, practices and tools. International Journal of Scientific & Engineering Research, 4(10), 1048-1056.

Adelstein, F. (2006). Live forensics: diagnosing your system without killing it first. Communications of the ACM, 49(2), 63-66.

Johnson, S., Scarlata, V., Rozas, C., Brickell, E., & Mckeen, F. (2016). Intel software guard extensions: EPID provisioning and attestation services. White Paper, 1(1-10), 119.

Macenski, T. Foote, B. Gerkey, C. Lalancette, W. Woodall, "Robot Operating System 2: Design, architecture, and uses in the wild," Science Robotics vol. 7, May 2022.

Yubico Product Documentation (2024, June 9). Retrieved from https://docs.yubico.com/

RIEGL VZ-400i (2024, June 9). Retrieved from http://www.riegl.com/nc/products/terrestrial-scanning/produktdetail/product/scanner/48/

Wang, M., & Yin, X. (2022). Construction and maintenance of urban underground infrastructure with digital technologies. Automation in Construction, 141, 104464. doi:10.1016/j.autcon.2022.104464

Lee, A. J., Song, W., Yu, B., Choi, D., Tirtawardhana, C., & Myung, H. (2023). Survey of robotics technologies for civil infrastructure inspection. Journal of Infrastructure Intelligence and Resilience, 2(1), 100018. doi:10.1016/j.iintel.2022.100018

Jia, Z., Liu, H., Zheng, H., Fan, S., & Liu, Z. (2022). An intelligent inspection robot for underground cable trenches based on adaptive 2d-slam. Machines, 10(11), 1011.

Costan, V., & Devadas, S. (2016). Intel SGX explained. Cryptology ePrint Archive.

Cen, S., & Zhang, B. (2017). Trusted time and monotonic counters with intel software guard extensions platform services. Online at: https://software. intel. com/sites/default/files/managed/1b/a2/Intel-SGX-Platform-Services. Pdf.

Newman, R. C. (2007). Computer forensics: evidence collection and management. Auerbach Publications.

Myung, H., Jeon, H., Bang, Y. S., & Wang, Y. (2014). Robotic sensing for assessing and monitoring civil infrastructures. In Sensor Technologies for Civil Infrastructures (pp. 410-445). Woodhead Publishing.

ANNEX

STATEMENT OF THE PUBLICATION ETHICS AND PUBLICATION MALPRACTICE

IDIMT's Publication Ethics and Publication Malpractice Statement is based, in large part, on the guidelines and standards developed by the Committee on Publication Ethics (COPE).

We expect all parties commit to these publication ethics. We do not tolerate plagiarism or other unethical behaviour and will remove any manuscript that does not meet these standards.

The relevant duties and expectations of authors, reviewers, and editors are set out below:

1. Author Responsibilities

Authors must certify that their manuscripts are their original work, and they take full responsibility for it, including parts created/based on AI.

Authors must certify that the manuscript has not previously been published elsewhere.

Authors must certify that the manuscript is not currently being considered for publication elsewhere.

Authors must notify us of any conflicts of interest.

Authors must identify all sources used in the creation of their manuscript.

Authors must report any errors they discover in their manuscript.

Authors must list all persons who contributed to the manuscript (co-authors, collaborations, etc). AI systems cannot be named as (co-)authors

Authors must identify all parts created by or substantially based on AI tools (e.g., created/rewritten by LLMs, translated by software). The model or tool used and its manufacturer must be cited.

Authors remain fully responsible for the content of their manuscripts, including ensuring the accuracy of the information and maintaining the originality and integrity of their work.

2. Reviewer Responsibilities

Reviewers must notify us of any conflicts of interest.

Reviewers must keep information pertaining to the manuscript confidential.

Reviewers must bring to the attention of the Editor-in-Chief any information that may be reason to reject publication of a manuscript.

Reviewers must at any time evaluate manuscripts only for their intellectual content without regard to race, gender, sexual orientation, religious belief, ethnic origin, citizenship, or political philosophy of the authors.

Reviewer who feels unqualified to review the research reported in a manuscript or knows that its prompt review will be impossible should notify us and excuse himself from the review process.

3. Editorial Board Responsibilities

The Editorial Board must keep information pertaining to submitted manuscripts confidential.
The Editorial Board must disclose any conflicts of interest.
The Editorial Board must evaluate manuscripts only for their intellectual content.
The Editorial Board is responsible for making publication decisions for submitted manuscripts.

LIST OF AUTHORS