ALMA MATER

# TANULMÁNYOK AZ INFORMÁCIÓ- ÉS TUDÁSFOLYAMATOKRÓL 22. / STUDIES ON INFORMATION AND KNOWLEDGE PROCESSES 22


Információs Társadalomért Alapítvány

ALMA MATER

# TANULMÁNYOK AZ INFORMÁCIÓ- ÉS TUDÁSFOLYAMATOKRÓL 22. / STUDIES ON INFORMATION AND KNOWLEDGE PROCESSES 22

**Információs Társadalomért Alapítvány**

**ALMA- MATER könyvsorozat / ALMA-MATER Book Series**
https://infota.org/kiadvanyok/alma-mater/

**Sorozatszerkesztő / Series editor:**
Kiss Ferenc, a Szerkesztőbizottság vezetője / Head of Editorial Board

**A kötet szerkesztője / Volume Editor:** Kiss Ferenc

**Technikai szerkesztő / Technical editor:** Kiss Ferenc, Szanyi István

All research papers published in this volume underwent a double-blind peer review process conducted by independent reviewers.



**DOI of the book:** https://doi.org/10.60030/ALMAMS.k202501bT9BiX

**ISSN 1587-2386**

**ISBN 978-615-82082-9-1 (pdf)**

**Első kiadás / First edition**
Budapest, 2025. december / December, 2025



**Kiadó / Publisher:**

Információs Társadalomért Alapítvány / Foundation for Information Society
1507 Budapest, Pf. 213.
Web: https://www.infota.org
E-mail: info@infota.org
Tel.: +36-1-279-1510
Felelős kiadó / Responsible for the Publication:
az Alapítvány Kuratóriumának elnöke / Chairman of the Curatorium

# TARTALOMJEGYZÉK

https://doi.org/10.60030/ALMAMS.c202501GJxDvn

# Knowledge governance of artificial intelligence: the first steps

László Z. Karvalics

Doctor of the Hungarian Academy of Sciences
Institute of Advanced Studies (FTI-IASK), Kőszeg, Hungary
e-mail:laszlo.karvalics@iask.hu
https://orcid.org/0000-0002-3502-434X

## Abstract

Governance of Artificial intelligence (*AI governance*) and *knowledge management of artificial intelligence* are well-known, widely discussed, popular terms with extensive technical and methodological literature. However, existing knowledge governance practices do not yet have to contend with the opportunities offered by artificial intelligence (AI *for* knowledge governance).

Knowledge governance *of* AI, however, is an increasingly important yet undeservedly neglected issue. In this study, which is intended to stimulate discussion, I offer a definition of the concept of knowledge governance and place it in the context of fundamental issues related to artificial intelligence. I demonstrate that knowledge governance can be considered the strategic meta-level of all areas of AI development, while it also needs to support management and governance tasks by utilizing the results of knowledge sciences.

Finally, using a four-level model of knowledge governance, I will highlight one or two pressing issues, whether it be global AI knowledge governance (macro level), that of nation states (meso level), that of companies and large organizations (micro level), or even that of individuals (nano level). I do so in the hope that this will spark an exciting dialogue on the subject.

**Keywords:** *AI for Social Good, Humanistic AI, knowledge governance, knowledge landscape, design thinking*

## Knowledge governance and artificial intelligence?

When it comes to artificial intelligence (hereinafter: AI) beyond its technical and engineering aspects, legal, regulatory and ethical issues have gained overwhelming prominence—both in communication statistics and in the world of government regulatory efforts and international organizations' position statements.

Meanwhile, *governance of artificial intelligence* (AI governance, AI strategy and leadership) and knowledge management of AI (knowledge management of organizations engaged in basic research, experimental development or application Régi of artificial intelligence) are well-known and increasingly discussed terms, with extensive discourse. However, it is important not to confuse the latter *with artificial intelligence solutions used in knowledge management practices* (AI in knowledge management, AI-ready knowledge management), which are one of the most important areas of application for, for example, new-generation applications based on large language models (Rezaei, 2025), promising to renew, transform and complete the traditional knowledge management world of companies and large organizations of any profile.

The scope of *knowledge management of artificial intelligence* can also be understood by identifying the components of an *artificial intelligence ecosystem* (AI ecosystem) of any size where knowledge processes, knowledge actors and knowledge sources can be identified, and, after thoroughly understanding and analyzing them, ensuring that artificial intelligence ecosystems and their individual elements can be more sustainable, more capable of development and more innovative. This is not an easy task, as everyone from end users to developers, traders and consulting firms, as well as the researchers and policymakers involved, are affected. It is no wonder that the need for oversight and data is growing enormously at the most comprehensive levels: when the *Global AI Ecosystem* created its decentralized and open community-building platform on a non-profit basis to provide a democratized and universally accessible environment for knowledge building, knowledge sharing, analysis, knowledge exchange and knowledge creation for the "global AI community" (https://www.ai-ecosystem.org/), its mission went far beyond

management tasks: in fact, it can be seen as a preparatory step towards global knowledge governance of artificial intelligence—even if its operators modestly refer to themselves as a Knowledge Hub.

However, existing knowledge governance practices do not yet need to contend with the possibilities offered by artificial intelligence (*AI for knowledge governance*), as there are few high-level knowledge processes in knowledge governance that can be partially or wholly transferred to AI. Fernandes (2024) also believes that the governance of artificial intelligence, the combination of AI and knowledge governance, is only just emerging. It is no coincidence that only one systematic publication on the subject of knowledge governance of AI has appeared to date, in April 2025, written by Daniel W. Rasmus, a leading expert at an analytical boutique firm (Rasmus, 2025). This excellent overview links previous typologies of knowledge types to different areas of AI application in such a thorough and innovative way that we will discuss its findings in detail later on. However practical Rasmus' model may be, as a knowledge management approach it can be considered highly simplified: its starting point is that artificial intelligence systems do not necessarily consume, produce, transform and maintain knowledge in the way we would expect. *"If organizations are not aware of the types of knowledge that functioning artificial intelligence systems come into contact with, they risk a large gap between intent and performance, and between expectations and models"* (Rasmus, 2025).

The term "organization" should be understood primarily to mean corporate players in the business world, although the lessons learned are, of course, valid and applicable in many ways to other (large) organizations (public administration, higher education, etc.). Among the additional issues not addressed by Rasmus but requiring review and analysis, and which are sure to be discussed more and more in the near future, I will focus on two particularly important areas in this study:

- the relationship between knowledge governance of artificial intelligence and the various strategic and practical professions of artificial intelligence
- outside the corporate-organizational world, I present and review each of the knowledge governance microcosms that have expanded from one level to four (macro-meso-micro-nano). Although many knowledge management considerations apply to

and are true for all system levels (each has its own rules, laws, characteristics and functions that differ from the others), which is why a separate, brief discussion of each is strongly justified.

First and foremost, however, it is worth defining knowledge governance itself, as it is such an under discussed and neglected concept and practice. According to an earlier definition (Z. Karvalics, 2013), the knowledge governance paradigm in its most general sense *refers to techniques for influencing interconnected knowledge processes, the design of structures and mechanisms that support the production and sharing of knowledge, selection, construction and control of structures and mechanisms that support the production and sharing of knowledge, at the highest level of intervention with transformative power in a given system, using a holistic approach.*

Today, I would rephrase this definition as follows: *"if any leader or representative of any action community who is sensitive to the knowledge ecosystem in which the community operates and, in order to act more effectively in the future, approaches any element of this ecosystem with a designer's mindset to shape, improve, maintain, develop, recreate, or perform knowledge governance, even if they view all this as a mere management task."*

This knowledge ecosystem is extremely rich and complex: it includes *knowledge processes* themselves, with their channels and media of knowledge flow, *knowledge carriers (educated human beings)*, *knowledge repositories*, *knowledge platforms*, *knowledge markets* and *supporting knowledge technologies*. And since the most advanced artificial intelligence systems are now capable of supporting every element of this ecosystem, and even replacing more and more parts of traditional solutions, it is high time to start a serious dialogue about high-level knowledge governance approaches.

The suitability of the knowledge ecosystem chosen as a conceptual framework can also be clearly understood from the perspective of long-established path dependency research.

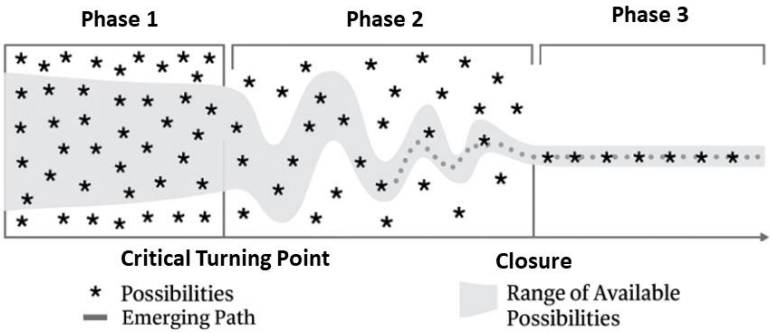



**Figure 1: Knowledge governance can also be interpreted as a path dependency challenge. Three-phase model of path dependency based on Sydow (2009) Kovács-Weber (2022)**

With an eye to the future, in the extensive, point-like range of possibilities, the number of elements to be considered can be reduced along the lines of exclusions and preferences, and where choice becomes action, it is possible to gradually approach a desired range. However, early steps can become hugely significant, even a little later: in many cases, it is the small, random, seemingly insignificant moments that exist in the space of alternatives at the beginning of the journey that can be decisive in shaping the final outcomes. And although even professional knowledge management practices cannot guarantee that it will be possible to steer artificial intelligence, which is permeating more and more corners of society, in the desired direction (especially given that the desired direction itself is determined in a discursive space), the risks can be reduced, the dangers can be better avoided, and there is a greater chance of steering reality closer to the desired objectives.

## Knowledge governance of artificial intelligence as a strategic meta-level

The ultimate meaning of the "promotion" of the knowledge governance perspective is that there is no basic discourse related to artificial intelligence that cannot form its "meta-level". In fact, the key to success, effectiveness and authenticity of every choice, decision, situation assessment, planning and action is the quality of the processing of the knowledge elements involved, based on what background knowledge. Behind every "input" there is some kind of live knowledge management

solution, routine or workflow element, and their current state and capabilities are the consolidated result of previous knowledge management moments (instants, events, practices). Their further development, modernization, modification, recalibration or replacement opens new knowledge management cycles.

Another important aspect of the "meta-level nature" is the special situation in which a properly tuned knowledge governance perspective must simultaneously take into account all mutually interacting strategic areas, regardless of where the thinking and pathfinding starts. From this point of view, the most important are decisions and actions, regulations and institutions related to the governance of artificial intelligence, because these fundamentally determine what each area will be capable of. They provide resources and set directions for the research community, orient the educational arena, and regulate economic actors. Governance control is perhaps least prevalent in the cultural dimension (arts, education, media), and it is very interesting that it is precisely these "soft" areas that sometimes exert a decisive influence on how politics or science think about artificial intelligence issues. The above correlations are presented in Table 1, illustrated with ten selected key areas of artificial intelligence.

**Table 1: Knowledge governance as a strategic meta-level of AI**

| **Knowledge governance of Artificial Intelligence (AI)** | | **→ ←** | **Knowledge sciences Management sciences** | |
|---|---|---|---|---|
| AI governance | AI development | Sciences of AI | AI Economy | AI literacy |
| AI policies | AI law | AI Education | AI Art | AI (in the) Media |

It is striking that knowledge management also has a meta-level: this includes the knowledge that determines the quality of knowledge management. This is based in part on the mobilization of decades of management science knowledge, and in part on precedents dating back several millennia, drawing on specialized disciplines dealing with the phenomenon of knowledge (Wierzbicki, 2007—epistemology, sociology of knowledge, cognitive science, knowledge engineering), but to a certain extent also creatology, talent science, linguistics, and even information and knowledge history. These background sciences are partly theoretical in nature but can also be of great practical importance (which is why applied knowledge sciences have become a separate category). It is particularly interesting that they overlap with the sciences of artificial intelligence:

cognitive science, epistemology and linguistics are of particular importance in both fields.

Rasmus (2025) also mentions in his article on knowledge governance that if we have a well-developed typology of primary knowledge types, we should not hesitate to use it and build on it: it is enough to compare these with the types of tasks related to artificial intelligence systems to find out what management and governance tasks belong to them. The "inner circle" of knowledge types (explicit, implicit, tacit) is surrounded by four basic categories that are distinct based on the nature of knowledge. Declarative knowledge is the world of facts and statements, embedded knowledge is found in tools and artefacts, procedural knowledge answers the question "how to know", and contextual knowledge adapts current challenges to the situation and interpretative framework.

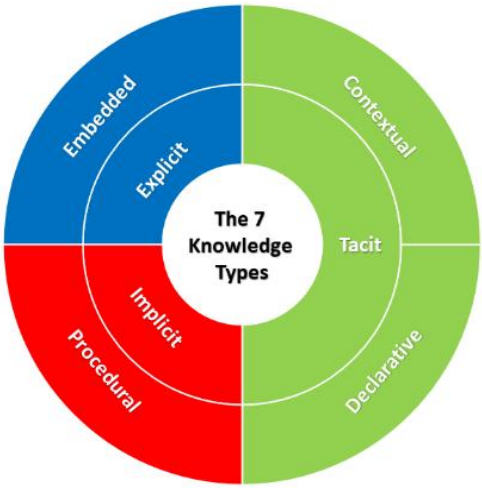



**Figure 2 Types of knowledge: inner and outer circles**

**From knowledge types to governance tasks (Rasmus, 2025)**

For those coming from the world of AI development, it may be convincing to see how types of knowledge can be identified in the familiar subtasks of system construction and operation—and from there, it is easy to see what this means for governance.

**Table 2: From knowledge types to governance tasks (Rasmus, 2025)**

| Type of knowledge | AI system examples | Type of governance action |
|---|---|---|
| Explicit | Prompts, model metadata, configuration files | Versioning, documentation, change control |
| Implicit | Compliance (guardrails), agent behaviour patterns | Testing, benchmarking, drift detection |
| Implicit (Tacit) | Prompting skills, agent coordination | knowledge capture, communities of practice |
| Declarative | Retrieval-supported language model, knowledge graphs | Source authentication, reliability ranking |
| Embedded | Agent architectures, immutable settings | Transparent design, bias testing |
| Procedural | Sequential (multi-step) agents, learning pattern recognition | Process mapping, risk reduction, "override" mechanisms |
| Contextual | Dynamic content management, script-based rules | Context and trigger logging, audit |

So what is the point and benefit of this kind of systematization?

A simple conclusion can be drawn from this: what tasks arise for a knowledge management leader from the above correlations?

- Developing the knowledge management practices of AI product teams
- Knowledge-based audit of current AI systems
- Building a common vocabulary to facilitate collaboration
- Urging governance measures that resonate with real knowledge dynamics
- Extracting the valuable tacit knowledge of AI system builders and users
- Raising awareness among managers that *all* AI is knowledge work
- Planning practices aligned with the knowledge asset lifecycle within the AI workflow

Finally, as an excellent example, I am pleased to present Mihály Nyáry's (2024) summary of the latest professional material from one of the world's leading think tanks, the legendary American RAND (Vermeer, 2024), which examines historical analogies that can be used in the management of artificial intelligence.

Classic knowledge management approach: if, in practice, our question is "*how can we maximize the benefits* of artificial intelligence with increasing

capabilities *while minimising the potential risks",* what source of knowledge can we draw on for the right answers?

This leads the author to four categories of technology where the questions of governance have been very similar in recent decades: nuclear technology, the Internet, encryption products, and genetic engineering. A historical overview of these four areas clearly shows that *"the history of their governance illustrates three themes: the need for consensus on technological standards, the important differences between the governance of physical and non-physical goods, and the role of public-private partnerships in governance.*

Based on these theoretical starting points, three recommendations that can be used for governance can be formulated:

*(1) Artificial intelligence systems that pose a serious risk of widespread harm, require significant resources to acquire and use, and have physical resources that can be tracked and controlled can be governed using the structure created for the international governance of nuclear technologies.*
*(2) The Internet governance model may be suitable for governing artificial intelligence systems that pose minimal risk.*
*(3) For accessible but high-risk AI systems, the genetic engineering governance model could be applied, but stakeholders should exercise caution when applying governance models similar to those used for encryption.*

**The four levels of knowledge governance—application to the world of Artificial Intelligence**

In conclusion, I believe it is necessary to reflect on the constantly changing and expanding knowledge governance scene in relation to the knowledge governance of artificial intelligence. The initially homogeneous interpretative framework, which was considered to apply exclusively to corporate knowledge processes, has gradually expanded over the past quarter of a century, and now the knowledge phenomena that require governance (and, of course, management) can be interpreted at four system levels.

**Table 3 Models of knowledge governance—expansion of the interpretative framework**

| *Original model (Grandori, 1997)* | *Enhanced model (Whitley, 2000)* | *Dalal-Z. Karvalics three-level model (2009)* | *Z. Karvalics four-level model (2012)* |
|---|---|---|---|
| company (company/ corporate/ enterpreneurial) | company (*micro level*) | company (*micro level*) | personal (individual) (*nano level*) |
| | | | company/organisation (*micro level*) |
| | nation-state (national) (*macro level*) | nation-state (*meso level*) | nation-state (*meso level*) |
| | | global (*macro level*) | global (*macro level*) |

Since attention is typically focused on the corporate-organizational world, it is worth considering some more powerful ideas in the case of less frequently explored system levels.

### *Macro level: Towards global knowledge governance of artificial intelligence*

A separate study could be devoted to the fundamental global/planetary issues related to artificial intelligence that affect everyone. Discussions on this topic are not yet taking place in the context of knowledge governance; the most concrete ideas are seeking *an institutional framework* that could be suitable for building a kind of international knowledge bridgehead around critical coordination tasks. However, the interpretation of these tasks is not insignificant.

In an interview, Gary Marcus and Anka Reuel (2025) propose the establishment of an international AI agency, while Josh Simons argues in his book for a new AI Platforms Agency that would not entrust the task to politicians but would instead set itself the mission of creating mechanisms for participatory governance. Shared experiences and mutual learning could be a means of preserving equality, self-determination and democratic foundations in the face of large corporations that are built on machine learning, sensitive only to profit considerations and striving for monopoly (Simons 2023: 218-220).

In the summer of 2025, the Chinese president presented his country's action plan for global artificial intelligence governance (Goh, 2025), in which he called on governments, international organizations, businesses and research institutes to work together and promote international

knowledge exchange, including through a cross-border open source community. However, whether we think in terms of institutions or communities of values, only continuous exchange of ideas and dialogue can bring us closer to either goal.

### *Meso level: nation states and knowledge governance*

So far, we see a single strategic "filter" emerging not only in the minds of rival superpowers, but also in those of political leaders of smaller countries: not to fall behind in the technological race represented by AI, to somehow promote the international business success of companies in the national AI sector—while at the same time reassuringly conveying the awareness that the political class is capable of averting the dangers posed by AI and robotisation.

From a knowledge management perspective, an audit would find serious shortcomings in this single, fragmented filter. Discussions of risk are disproportionately exaggerated, shaped not by state-of-the-art science but by panic narratives in the media. For many smaller countries, it makes no sense to seek a place in endlessly competitive and resource-intensive development processes, given that even the most powerful players burn through vast amounts of money on poorly calculated development directions. Instead, they should turn to applying and disseminating existing results and solutions. But above all, government plans typically serve the ambitions of the dominant companies in the AI industry, while these companies also dominate research institutes in the "academic" world (Jacobides, Brusoni and Candelon, 2021). It is no coincidence that the People's AI Action Plan movement (https://peoplesaiaction.com/), which brings together many players and aims to steer the development of AI towards common well-being, a sustainable future and fairer distribution, when it became clear that the Trump administration's AI action plan was "top-heavy" for a well-defined business circle. From a knowledge governance perspective, it is particularly concerning that both the position prophesying the imminent explosion of artificial intelligence (AI boosterism) and the panic narratives that exaggerate the dangers come from both AI industry leaders and background players, who are also eagerly recruited by international organizations to sit on committees compiling technical materials on the future of AI. (Suffice it to mention the names of Geoffrey Hinton, Joshua Bengio or Sam Altman.) Schneier and Sanders (2025) have devoted an entire monograph to the question of

what key actions governments can take to steer AI developments in the direction of the public good. In fact, the raw material for this challenge has been taking shape for a long time.

Thirty years ago, Star (1989:41) proposed introducing the Durkheim test for AI: where the result of a community's collaboration is the real-time design, acceptance, use and modification of an AI system, the "intelligence" of the system could be directly measured by its "beneficial effects on community activity or the community's ability to change and adapt".

In any case, it is the states that bear the greatest responsibility for how deeply they understand the information and knowledge components of the national AI ecosystem and how they are able to create and use the necessary knowledge to shape this ecosystem.

I think it is worth presenting what I consider to be the best practice, the Dutch Artificial Intelligence Innovation Centre *( Innovation Centre for Artificial Intelligence, ICAI ).* The Dutch government wanted to give a boost to social, community and human applications, which were sorely lacking in research and development, by combining many resources to create the ROBUST network of 51 partners. Development work is carried out in 17 newly established research centers (laboratories), in line with the UN's Strategic Development Goals: each center is organized to investigate a specific social issue defined here. And not just any old way: right from the start, 85 PhD-positions were created so that, with a young research staff of 170 within five years, they could move forward with a work plan that, instead of long thesis maturation, aims to influence key players through shared results and continuously displayed application possibilities with a turnaround time of a few months. This speed is particularly beneficial to small and medium-sized enterprises, which, lacking sufficient capital and human resources, are unlikely to consider long-term projects. In addition, ROBUST is attracting this group of entrepreneurs to partnership by creating data laboratories for them, based on a previous successful practice. In these laboratories, skilled students with expertise in various data technologies solve sensitive problems for companies in cycles of up to twelve weeks for a nominal fee.

The network's activities focus on transparency, reliability and measurability, and the key word in its development philosophy is "context sensitivity" (it matters what the environment and scope of a technology is).

This provides a good basis for fulfilling the role that ROBUST leaders refer to as "orchestrating" the Dutch AI ecosystem: this concept includes both the coordination of stakeholders and the task of connecting to EU joint initiatives and networks.

### *Micro level: companies and large organizations*

There is no doubt that the application of knowledge governance considerations is a necessity for companies developing artificial intelligence systems. However, it is just as important in the world of *knowledge companies*, whose entire value chain is being challenged by the latest AI solutions. Numerous studies have already been conducted on knowledge management in knowledge-based companies. And in fact, a hospital, hospital operations and healthcare systems as a whole can be considered knowledge companies, just like the world of universities and colleges, or even public education itself.

In all cases, the key is having the right basic knowledge about what existing systems are worth using and how. If someone at a university conducts a poorly designed and methodologically questionable experiment with students using ChatGPT, and then joins the chorus of those who say that "artificial intelligence makes you stupid," then their institution will be at a disadvantage compared to those who develop a vision and strategy for the future of AI in their institution based on valid knowledge. With appropriate knowledge management routines, this type of danger can be easily avoided—but it is clear how important it is, for example, to formulate effective counter-messages to the messages of the media world, which influence the thinking of decision-makers and greatly discount their preparedness.

AI knowledge management is progressing slowly at this level because knowledge management practices themselves are slow to gain ground in the corporate world. Although the first Chief Knowledge Officers (CKOs) took up their posts before the turn of the millennium, the culture and organizational solutions of knowledge management have not yet found their way into many companies—and they face even greater difficulties in the world of AI, because the time requirements of certain cycles of knowledge production and use are seriously challenged by the astonishing pace of change in the AI market and technology.

### *Nano level: individual knowledge management strategies*

It is clear that if even properly "trained" decision-makers can be misled by clickbait tabloid media, then this danger is even greater for the average, "ordinary" citizen who is looking for guidance, support and direction in the world of AI. Their goal can be nothing other than to increase their own life and career opportunities, keep up with the rapidly changing world of tools and services, and develop personalized adaptation and learning strategies. The relevant knowledge is spread through family and friend channels, and the development of general "artificial intelligence literacy" (AI literacy) is still pending. However, the European Commission, for example, already recognized the importance of this in 2018 as a skill and civic competence (Hirvonen, 2024) that enables people to act as responsible citizens, participate fully in the life of their communities, understand social, economic and political structures, know key concepts, follow global developments and be sufficiently sensitive to sustainability issues.

And at the individual level, there is not just one but several layers of problems, depending on the social role each person finds themselves in. As an employee, the goal is job security and meaning in work, love for one's job, and higher income, which is why it is necessary to constantly monitor developments in the world of AI. As a parent, the challenge is to create and supervise an AI-supported digital environment for your child (games, health monitoring, learning support, skill development, communication). As a social being and network citizen, the challenge is how to bring existing AI practices closer to what you would expect on a normative basis. Meanwhile, the time is approaching when young people will develop increasingly complex relationships with their personal digital companions (Artificial Lifetime Companions, ALCs) from the moment they first encounter the digital ecosystem. Such a cohabitation relationship goes far beyond the current concepts of AI literacy, as it is built on continuous, human-assisted self-education, self-development and learning (Z. Karvalics, 2024), and can play a key role in the success of one's life path.

**Summary**

The study had a single objective: to demonstrate that the time has come for discourses on knowledge governance in the world of artificial intelligence. The review is admittedly not systematic and does not even strive for partial completeness: it aims to stimulate interest in knowledge governance issues and presents examples that demonstrate the importance and viability of this approach and discourse. All this in the hope that a lively dialogue on the subject might ensue.

**Bibliography**

Fernandes, A. A., et al. (2024). *Implications of artificial intelligence governance on organizational knowledge governance in the era of generative artificial intelligence* [Conference presentation]. 19º Congresso Brasileiro de Gestão do Conhecimento (KM Brasil LATAM 2024). https://www.researchgate.net/publication/384113936_Implications_of_Artificial_Intelligence_Governance_on_Organizational_Knowledge_Governance_in_the_Era_of_Generative_Artificial_Intelligence

Goh, B. (2025, July 26). China proposes new global AI cooperation organisation. *Reuters*. https://www.reuters.com/world/china/china-proposes-new-global-ai-cooperation-organisation-2025-07-26/

Grandori, A. (1997). Governance structures, coordination mechanisms and cognitive models. *Journal of Management and Governance, 1*, 29–42. https://doi.org/10.1023/A:1009977627870

Hirvonen, N., et al. (2024). Artificial intelligence in the information ecosystem: Affordances for everyday information seeking. *Journal of the Association for Information Science and Technology, 75*, 1152–1165. https://doi.org/10.1002/asi.24860

Jacobides, M. G., Brusoni, S., & Candelon, F. (2021). The evolutionary dynamics of the artificial intelligence ecosystem. *Strategy Science, 6*(4), 265–445. https://doi.org/10.1287/stsc.2021.0148

Klikauer, T. (2025). Review of Josh Simons' book *Algorithms for the people: Democracy in the age of AI*. *tripleC, 23*(1), 152–155. https://doi.org/10.31269/triplec.v23i1.1617

Kovács, S., & Weber, E. (2022). The impact of road dependency on spatial economic development. *Danubius Noster, 4*, 277–289. https://doi.org/10.55072/DN.2022.4.277

Marcus, G., & Reuel, A. (2023, April 18). The world needs an international agency for artificial intelligence, say two AI experts.

*The Economist*. https://www.economist.com/by-invitation/2023/04/18/the-world-needs-an-international-agency-for-artificial-intelligence-say-two-ai-experts

Nyáry, M. (Ed.). (2024, August 5). Historical analogies that can help in the governance of AI. *E-Gov Newsletter*. https://hirlevel.egov.hu/2024/08/25/tortenelmi-analogiak-amelyek-segithetnek-az-mi-iranyitasaban/

Rasmus, D. W. (2025, April 15). A practical AI knowledge governance framework: Mapping AI approaches to knowledge types. *Serious Insights*. https://www.seriousinsights.net/ai-knowledge-governance-framework/

Rezaei, M. (2025). Artificial intelligence in knowledge management: Identifying and addressing the key implementation challenges. *Technological Forecasting and Social Change, 217*, Article 124183. https://doi.org/10.1016/j.techfore.2025.124183

Schneier, B., & Sanders, N. E. (2025). *Rewiring democracy: How AI will transform our politics, government, and citizenship*. MIT Press. https://doi.org/10.7551/mitpress/15845.001.0001

Simons, J. (2023). *Algorithms for the people: Democracy in the age of AI*. Princeton University Press. https://doi.org/10.1353/book.109841

Star, S. L. (1989). The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem solving. In L. Gasser & M. N. Huhns (Eds.), *Distributed artificial intelligence* (pp. 37–54). Morgan Kaufmann.

Vermeer, M. J. D. (2024). *Historical analogues that can inform AI governance* (Research report, pp. 1–40). RAND Corporation. https://www.rand.org/pubs/research_reports/RRA3408-1.html

Whitley, R. D. (2000). The institutional structuring of innovation strategies: Business systems, firm types and patterns of technical change in different market economies. *Organization Studies, 21*, 855–886. https://doi.org/10.1177/0170840600215002

Wierzbicki, A. P., & Nakamori, Y. (2007). Knowledge sciences: Some new developments. *Zeitschrift für Betriebswirtschaft, 77*, 271–296. https://doi.org/10.1007/s11573-007-0021-8

Z. Karvalics, L., & Dalal, N. (2009). An extended model of knowledge governance. In M. D. Lytras, P. Ordóñez de Pablos, E. Damiani, D. Avison, A. Naeve, & D. G. Horner (Eds.), *Best practices for the knowledge society: Knowledge, learning, development and technology for all* (Communications in Computer and Information Science, Vol. 49). Springer.

https://doi.org/10.60030/ALMAMS.c202501uqxnyj

# Trust and Adoption of AI in Business: Perceptions, Risks, and Solutions

Levente Szabados

associate professor
Frankfurt School of Finance and Management
e-mail: l.szabados@int.fs.de
https://orcid.org/0009-0003-6046-9361

Ádám Buza

AI consultant
Neuron Solutions Ltd.
e-mail: buza.adam@neuronsolutions.hu
https://orcid.org/0009-0007-8161-0968

## Abstract

Artificial intelligence (AI) is transforming business processes; however, trust remains a key factor in its adoption. Drawing on a survey of Central European professionals, we examine the relationship between AI use and trust, approaches to assessing AI reliability, and the role of different strategies in responsible deployment. The results highlight the importance of trust-enhancing tools and regulatory frameworks, which can support the responsible uptake of AI in the business environment.

***Kulcsszavak:*** *artificial intelligence, AI, business application, trust, RAG, evaluation*

## Introduction and context

The development of Artificial Intelligence (AI) systems has gained significant momentum over the past fifteen years, particularly with the latest wave of deep neural network technology, known as *Deep Learning* (for the origin of the term, see also: (Dechter 1986), (Fradkov 2020). This progress has led to the emergence of so-called *foundational models*

(Bommasani et al. 2022), and more specifically, *Large Language Models* (LLMs; (Brown et al. 2020)), whose widespread and low-cost accessibility has driven their adoption (regarding the decreasing consumer cost of LLM usage, see also: (Appenzeller 2024)).

However, the concept of Artificial Intelligence is not new: AI development can be divided into several phases, during which technology has attempted to simulate different aspects of human thinking using various approaches. The AI systems that emerged in the 1950s were primarily based on symbolic logic, fitting within the paradigm of *expert systems*, which operated using strict rule-based frameworks and manually encoded knowledge bases. While these systems were effective in certain domains such as medical diagnostics and financial analysis, they lacked the ability to dynamically adapt to new situations.

In the early 2000s, the *Machine Learning* (ML) paradigm took the lead, enabling systems to learn from data without explicitly programmed rules. ML methods, particularly supervised learning, were successfully applied in fields such as image recognition and market predictions. The evolution of this technology reached a breakthrough point in 2012 when deep neural networks (*Deep Learning*) demonstrated their dominance in computer vision and other AI applications. Nevertheless, AI at this stage remained strictly *narrow AI*, meaning that models were designed for specific tasks and optimized for particular applications.

After 2020, AI research underwent another paradigm shift, driven primarily by the emergence of *Large Language Models* (LLMs). Unlike earlier models trained for single tasks, LLMs exhibited general problem-solving capabilities, demonstrating the ability to synthesize human language and solve complex problems through linguistic interaction. The release of OpenAI's *ChatGPT* model in November 2022 widely showcased these capabilities, generating global media attention and unprecedented industrial and economic interest.

Given this context, it is not surprising that AI's application in the world of work has also undergone rapid development (Anthropic 2025), and also the economics community took note (see (Szabados 2024)). A McKinsey survey (Singla et al. 2024) from early 2024 found that 65% of respondents' organizations use generative AI, mainly LLMs in some capacity. With this in mind, we set out to investigate the adoption patterns and surrounding attitudes about AI usage in the Central-European region to get a better

insight into the possibilities as well as impediments with respect to AI usage in business relevant contexts.

## Emerging trust problems with AI systems

Parallel with the drastic increase in capabilities of foundational models in general and LLMs in particular some problems seemed to rise quite naturally from the fact that these models are in a sense "too good" at playing their roles, so even in their early work (Radford et al. 2019) mentioned the capability of models to produce convincing but factually incorrect answers (later on popularly called "hallucinations"), that undermine the trust of their users in their abilities, and especially their usefulness in business relevant applications.

Current broad consensus from the technical community seems to have settled (for now) on the solution framework proposed by (Lewis et al. 2020), namely "Retrieval Augmented Generation" (or RAG for short), that is basically constituting a hybrid solution, where the LLM is paired with a knowledge retrieval / search capability that provides relevant (often company specific or even internal) sources to "ground" the behaviour of LLM models, thus enabling them to substantially mitigate the problems caused by hallucination.

Thus said, even with these efforts in place, pretty impactful - and reputationally as well as monetarily damaging - cases appeared in the news that report failure cases of LLM based solutions in business applications, like the case (Ea 2024) when Air Canada was ordered by the British Columbia Civil Resolution Tribunal to honour a policy created by its chatbot, which incorrectly promised a bereavement fare refund, after the airline argued the chatbot was a separate legal entity, or the case when a ChatGPT-powered AI chatbot at a Chevrolet dealership was tricked into agreeing to sell a 2024 Chevy Tahoe for $1 after a buyer manipulated it into accepting any customer request as a legally binding offer (AIAAIC 2024).

These - and multiple similar - events were quickly disseminated by the media, and contributed to a substantial degree of (quite justified, one might argue) mistrust towards the application of AI in the business community. With our survey we specifically set out to investigate this trust relationship, and the possible venues for mitigation of distrust.

## Research question

With the previous context in mind, we set out to investigate the following research questions:

- What is the current state of adoption of AI systems in the Central European region?
- What is the perceived level of trust in AI systems?
- How does this influence the decisions to apply AI in business contexts?
- What can be done methodically to increase trust in AI systems?

## Methodology

This study employs a quantitative approach to assess perceptions of trust in AI systems and their influence on business adoption. The methodology primarily revolves around a structured survey distributed to professionals across various industries and organization sizes.

## Survey Design and Distribution

The survey was designed to capture respondents' attitudes toward AI adoption, trust levels, and factors influencing AI integration. Key components of the questionnaire included:

- Demographic information (region, industry, organizational role, organization size)
- AI usage within the organization (frequency, scope, and strategic importance)
- Trust in AI systems (perceived risks, confidence in AI safety measures)
- Evaluation of AI trust-establishing tools

The survey was distributed using a **snowball sampling** technique, leveraging social media platforms (e.g., LinkedIn, industry forums) and business networks. This approach ensured the recruitment of professionals actively engaged in AI-related discussions.

## Sample Characteristics

- The survey received responses from **150 participants**.

- The regional distribution was predominantly from Hungary (58%) and Germany (30%), ensuring a strong representation from Central Europe.
- Respondents represented various industries, with a strong presence from the technology (50%) and financial sectors (12.7%).
- Participants came from organizations of various sizes, ensuring a balanced view of AI adoption across different business scales.
- Respondents held diverse organizational roles, including IT, executive management, and research and development.

**Data Collection Period**

The survey was conducted over a defined period, concluding on **August 25, 2024**. Responses were collected, anonymized, and analysed to ensure data integrity and respondent privacy.

**Limitations**

While the survey methodology provides valuable insights, several limitations must be acknowledged:

- **Self-selection bias:** Participants who are already engaged with AI topics may be overrepresented.
- **Geographic limitation:** The majority of respondents are from Hungary and Germany, which may limit global generalizability.
- **Industry concentration:** The predominance of technology sector respondents could skew results toward AI-intensive industries.

Despite these limitations, the methodology provides a robust foundation for understanding AI trust dynamics within business contexts.

**General description of survey participants**

The survey collected responses from a **total of 150 participants**. The geographical distribution of respondents can be studied on Figure 1, which illustrates the country-wise representation, with the majority of respondents originating from Hungary (58%). This indicates that our insights will be heavily weighted toward Hungarian perspectives on AI adoption and related issues.

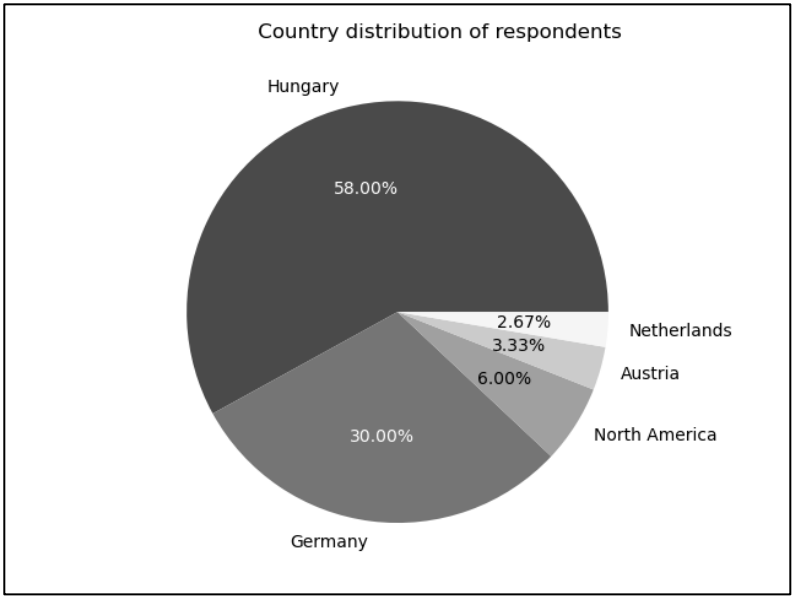



**Figure 1: Country distribution of respondents**

Germany accounts for the second-largest share of respondents (30%), followed by Austria (3.33%), North America (6%), and the Netherlands (2.67%). The strong representation from Hungary and Germany ensures that the findings are most reflective of AI adoption trends in Central Europe. The inclusion of Austria adds further context within the DACH region, though Switzerland is not represented.

The smaller number of responses from North America and the Netherlands suggests that the survey results may not fully capture perspectives from Western Europe or the broader international AI landscape. As a result, while the findings are relevant for understanding AI deployment in Hungary and Germany, they may not be as generalizable to other regions with different economic, regulatory and technological environments.

Figure 2 provides an overview of the industry distribution of respondents. The majority of participants (50%) are from the technology sector, reflecting the strong interest of professionals in AI-related developments. The financial sector represents the second-largest group (12.7%), followed by manufacturing (8%), education (6.7%), and retail (5.3%). Other industries, such as consultancy (2.7%), government (2.0%), research (1.3%),

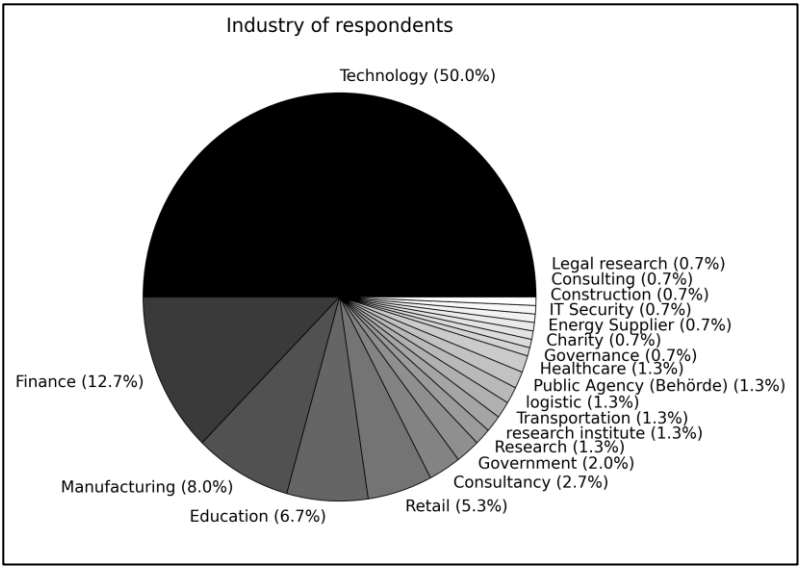



**Figure 2: Industry of respondents**

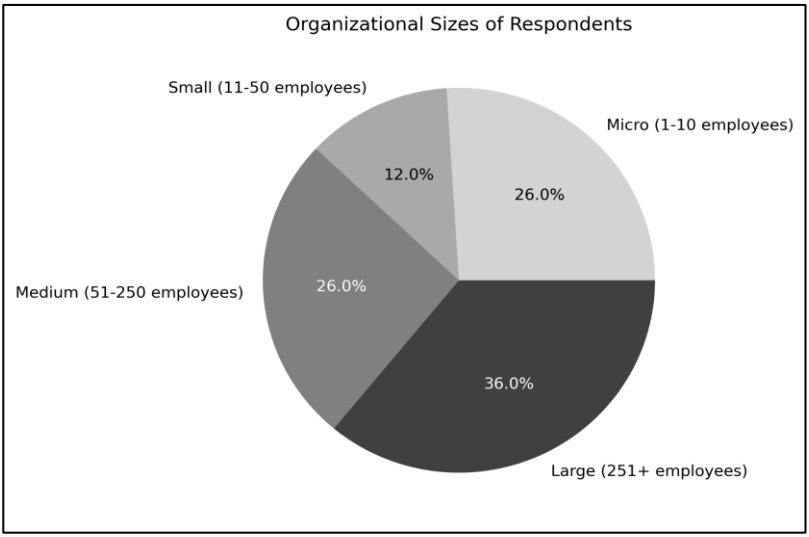



**Figure 3: Organizational Sizes of Respondents**

healthcare (1.3%), and public agencies (1.3%), are represented to a lesser extent. The remaining respondents are distributed across a variety of sectors, including logistics, transportation, construction, and legal research, each contributing a marginal share. This industry distribution suggests that the survey findings primarily reflect the perspectives of professionals working in AI-intensive sectors, particularly technology and finance, while industries with lower AI penetration, such as healthcare and public administration, are less represented.

Figure 3 presents the distribution of organizational sizes among respondents. The largest share of participants (36%) work in large organizations with more than 251 employees, while medium-sized organizations (51-250 employees) and micro-enterprises (1-10 employees) each account for 26% of respondents. Small organizations (11-50 employees) make up 12% of the sample. The dataset primarily represents companies with a sufficient scale to consider AI adoption as a strategic initiative, avoiding a strong bias toward very small businesses that may lack the resources or incentive to engage in AI-driven transformation. The relatively even distribution of organizations across different size categories ensures that the survey captures insights from both well-established enterprises and smaller, more agile businesses, offering a balanced perspective on AI adoption across various operational scales.

Figure 4 presents the detailed breakdown of organizational roles among survey respondents. The data highlights a diverse representation across multiple functions, with the largest proportion belonging to IT (26.7%), followed by Research and Development (24.0%) and Executive Management (14.7%). Other roles such as Finance, Marketing, Sales, Procurement, and Operations are represented to a lesser extent, each accounting for a few percentage points of the total. The long-tail of roles, including Data Science, Organizational Development, and Management Consultancy, suggests that AI adoption is relevant across a wide spectrum of business functions, albeit with varying degrees of involvement.

To facilitate the analysis of AI usage an views across broad functional categories, the roles were semantically grouped into key domains, as depicted in Figure 5. The classification consolidates various functions into four primary categories: IT (26.67%), Research and Development (27.33%), Executive Management (14.67%), and Other roles (31.33%).

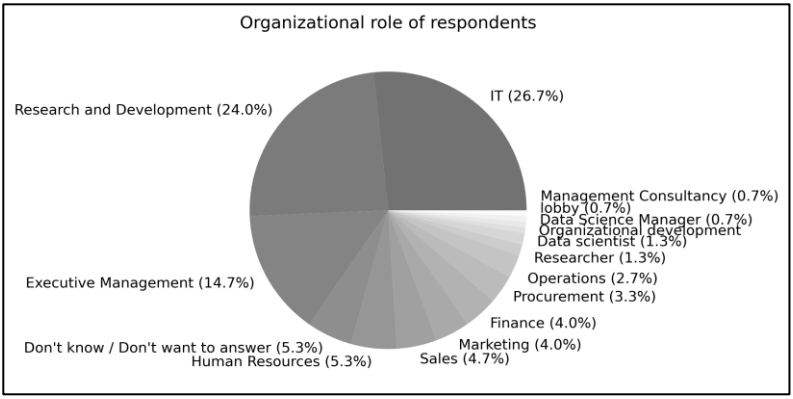



**Figure 4: Organizational role of respondents**

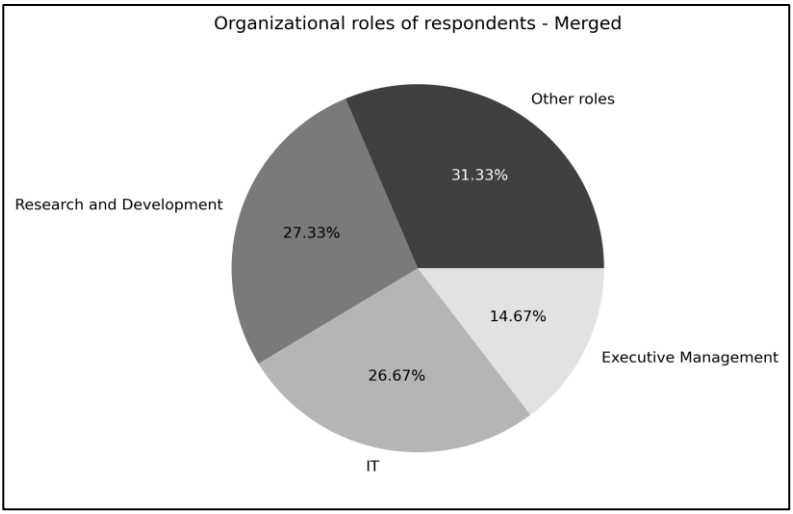



**Figure 5: Organizational role of respondents - merged**

This grouping helps to better understand how different functional areas engage with AI adoption and trust-building efforts. The high representation of IT and Research and Development indicates a strong involvement of technical and innovation-driven teams in AI deployment, while the presence of executive management underscores the strategic importance of AI initiatives. The “Other roles” category captures the diversity of respondents whose roles may not fit neatly into the primary groupings but still contribute to AI-related decision-making and implementation.

## Survey results

### AI usage in the organizations

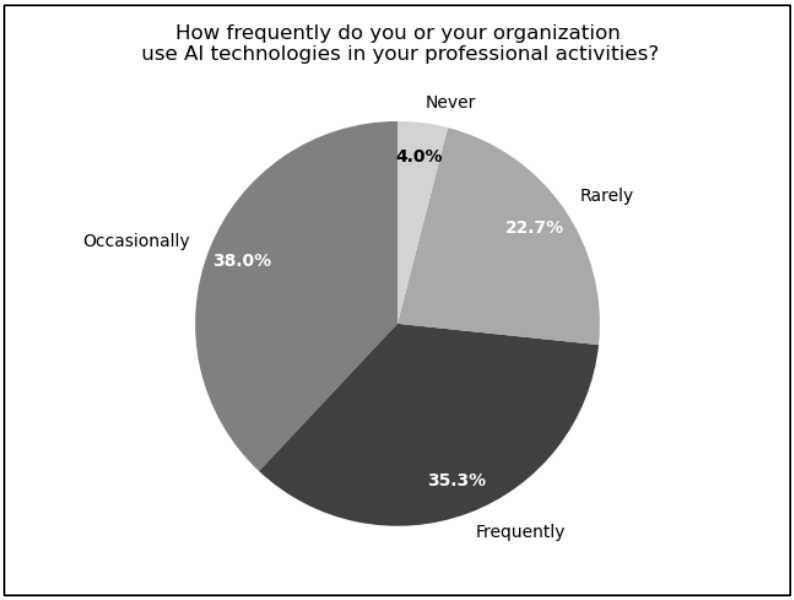



**Figure 6: How frequently do you or your organization use AI technologies in your professional activities?**

Our survey results (Figure 6) indicate a high level of AI adoption among respondents:

- **Frequently:** 35.3% of respondents use AI technologies regularly in their professional activities.
- **Occasionally:** 38.0% use AI from time to time.
- **Rarely:** 22.7% report minimal AI usage.
- **Never:** Only 4.0% of respondents state that they do not use AI at all.

These figures suggest that AI is widely integrated into the professional activities of our survey participants, with more than **73% using AI at least occasionally**.

### Comparison with Eurostat AI Adoption Data

To put these results into perspective, we compare them with the latest AI adoption statistics from (Eurostat 2024), which reports the percentage of enterprises using at least one AI technology:

- **Germany:** 19.75% of enterprises used AI in 2024, up from 11.55% in 2023.
- **Hungary:** 7.41% of enterprises used AI in 2024, up from 3.68% in 2023.
- **EU Average:** 13.48% of enterprises used AI in 2024, up from 8.03% in 2023.

Compared to these figures, our survey shows a much higher rate of AI adoption among respondents. The gap between our results and the Eurostat data suggests that our sample is **not representative of the general enterprise landscape in Germany, Hungary, or the EU**. Instead, our dataset is likely biased toward professionals who are already engaged in AI discussions and implementation.

### Interpretation of the Results

This response bias is expected and does not diminish the relevance of our findings. Since our research focuses on **AI trust and adoption concerns**, we are primarily interested in the perspectives of those who use AI, rather than those who do not. The high level of AI adoption in our dataset ensures that our analysis captures the opinions of those actively working with AI technologies, making our insights more relevant to understanding the challenges, risks, and trust factors associated with AI adoption in business contexts.

Figure 7 aims to examine how organizational size correlates with the strategic importance attributed to AI by our respondents. A general observation is that across all organization sizes, a considerable proportion of respondents regard AI as either important or critical, indicating its growing relevance in business strategy. Large enterprises tend to recognize AI as highly significant, with 25.9% of respondents considering it critical and an additional 29.6% rating it as important. This trend is expected, as larger organizations typically have more resources, structured innovation strategies, and established AI-driven initiatives, making AI integration a key component of their operations.

## AI's Importance and Impact on the Organization

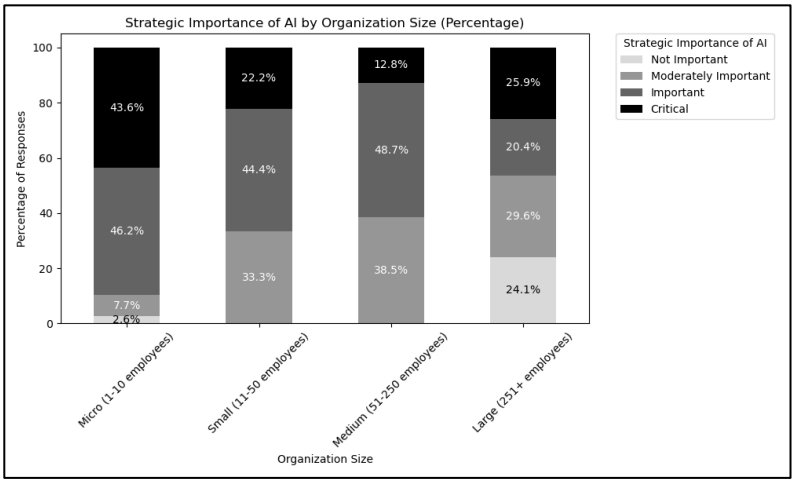



**Figure 7: Strategic Importance of AI by Organization Size**

An intriguing aspect of the data is that micro-enterprises report the highest percentage of respondents perceiving AI as critical, at 43.6%. This finding is somewhat unexpected, as smaller organizations often lack the same level of financial and technological resources as larger enterprises. However, it is likely influenced by the composition of survey respondents, many of whom may come from technology-centric micro enterprises, consulting firms, or other AI-driven sectors where the adoption and reliance on AI are fundamental to business operations. This suggests that in certain industries, particularly in technology and professional services, even very small organizations may consider AI to be a crucial element of their strategic approach.

In contrast, small and medium-sized enterprises exhibit a more balanced distribution of responses. For small enterprises, 22.2% of respondents classify AI as critical, while 44.4% consider it important, indicating that while AI holds value, it is not yet universally seen as an indispensable strategic asset at this scale. Similarly, medium-sized enterprises have a strong presence in the important category (48.7%), with a lower percentage (12.8%) identifying AI as critical. This suggests that while AI is widely recognized as an important tool for competitiveness and

innovation, its absolute necessity may be less pronounced compared to micro and large enterprises.

Overall, the data highlights that AI is increasingly viewed as a strategic priority across businesses of all sizes, though the degree to which it is considered critical varies. Large enterprises predictably show a strong inclination toward integrating AI into their strategic framework, but the particularly high importance placed on AI by micro-enterprises points to industry-specific factors at play. This finding underscores the fact that while AI adoption is often associated with large organizations, smaller firms, especially those in AI-intensive industries, may be just as, if not more, reliant on it for their survival and growth.

The above mentioned effects are also reinforced by the respondents' views about the impact of AI on their general industries, so as Figure 8 shows, they feel, that not only for their specific company setups is the adoption of AI a necessary component, but also estimate the impacts of it for their whole industries to be considerable. As visible on the chart, the effect of organizational size is also very similar on the respondent's views, so large and micro organizations (with sectoral caveats) might be more prone to feel the need for AI adoption.

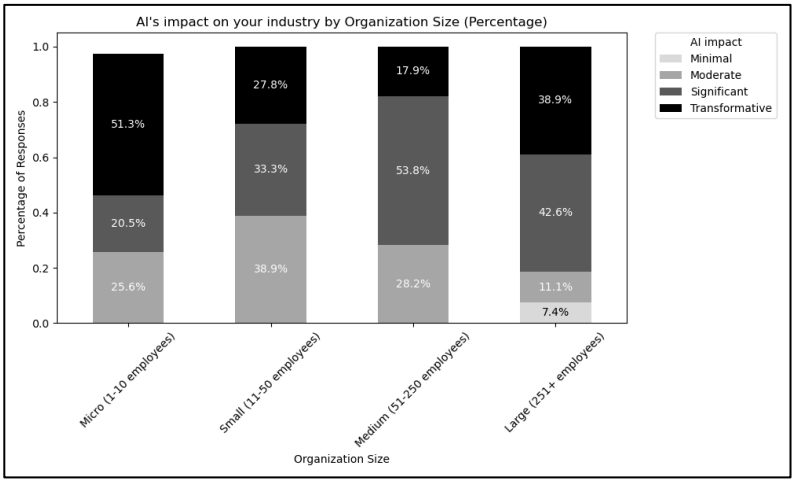



**Figure 8: AI's impact on your industry by Organization Size**

## Most frequent use cases of AI

As a deeper layer of analysis for the actual usage of AI (with emphasis on most recent AI capabilities) we included questions to assess the areas and use cases that respondents encounter the most frequently in their organizational contexts.

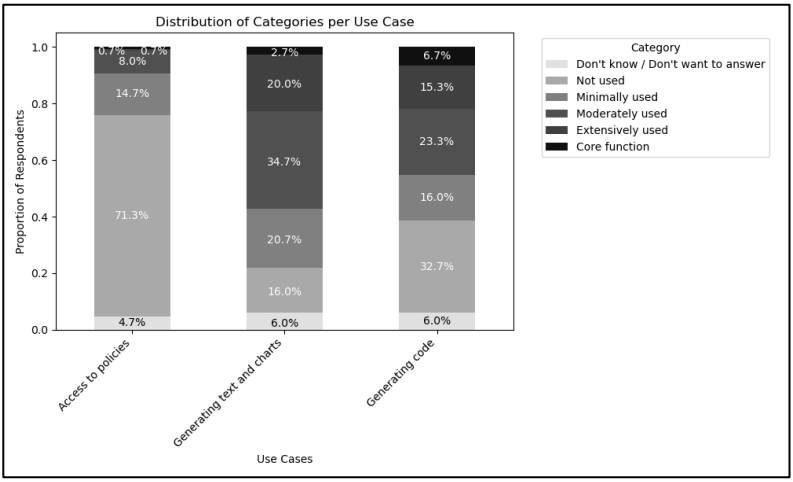



**Figure 9: Distribution of Categories per Use Case - Internal Use Cases**

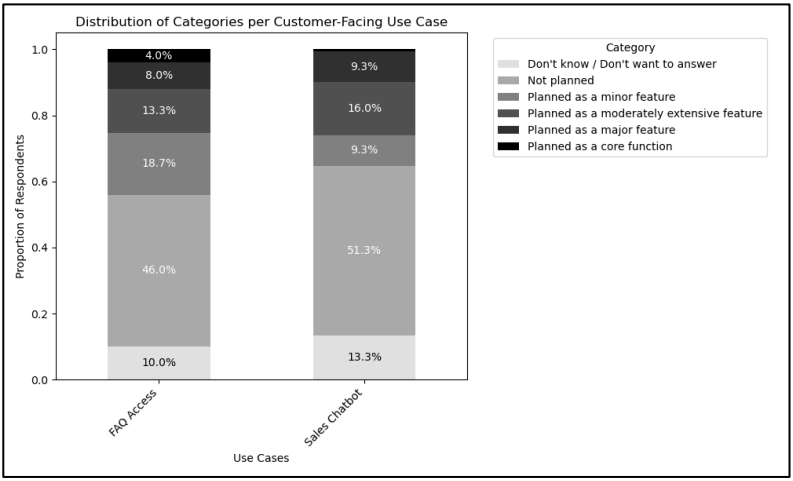



**Figure 10: Distribution of Categories per Use Case - Customer-Facing Use Cases**

Figures 9 and 10 illustrate the most frequent AI use cases in organizations, categorized into internal and customer-facing applications. The data highlights that AI adoption is relatively higher for internal use cases, with an average of **54.5%** of respondents indicating at least minimal usage. In contrast, external AI adoption stands at **43.95%**, reflecting both interest and caution in deploying AI in customer-facing roles.

For internal applications, AI is most frequently used for accessing policies (71.3% indicate it is not used, while 28.7% report some level of adoption), generating text and charts (54.7% adoption, with 20.0% extensively using it and 2.7% marking it as a core function), and generating code (54.0% adoption, including 15.3% extensively using it and 6.7% considering it a core function). These results suggest that AI plays a significant role in knowledge retrieval and automation of documentation, as well as in content generation where Large Language Models (LLMs) are often implemented using Retrieval-Augmented Generation (RAG) techniques. However, challenges such as hallucinations—incorrect AI-generated outputs—remain an issue, which we will explore in further detail later in the article.

Externally, AI is primarily applied in **FAQ access** and **sales chatbots**. The survey data indicates that FAQ access is actively planned by **44% of organizations**, while sales chatbots are even more widely adopted, with **51.3% of respondents planning for at least minimal integration**. Notably, sales chatbots have a **higher proportion of respondents (9.3%) identifying them as a core function**, indicating that businesses see AI-driven customer interactions as a crucial part of their digital strategy.

A key observation is that **external AI use cases carry a higher risk compared to internal applications**. Since customer-facing AI directly interacts with consumers, errors in chatbot responses, bias in FAQ generation, or misinterpretations could lead to reputational damage, diminished brand trust, and even financial losses. This is particularly relevant for organizations where AI-generated interactions influence customer decision-making and support experiences.

The comparison between internal and external AI adoption suggests a **more cautious approach toward customer-facing AI applications** due to their potential risks. However, the fact that nearly **44% of organizations are already integrating AI externally** reflects the growing confidence in AI's capabilities and its increasing role in business

automation. In the following sections, we will further analyse how organizations manage AI-related risks and trust-building mechanisms, especially in high-stakes external use cases.

**Trust vs. usage**

The core subject of our investigation is the trust (and the associated concerns / possible solutions) in AI systems applied in the professional business context. To analyse the attitudes and perceived trust in AI systems we decided to conduct a detailed evaluation of the usage patterns of said technologies with respect to the given organizational roles of the respondents (in a "merged" manner, as defined already above).

But before we go deeper into the levels of trust and it's relationship to organizational role, as a prerequisite (and in a sense a possible confounder) we have to ponder the general usage patterns of AI for the given roles.

As Figure 11 demonstrates, though in our "tech-savvy" respondent pool, the utilization of AI tools is generally high (especially in comparison to the Eurostat data (Eurostat 2024) quoted before), so on average over the roles 35.0% report frequent usage, 37.15% report occasional usage, or in sum a total 72.15% report noteworthy usage, the levels vary considerably with respect to their individual roles.

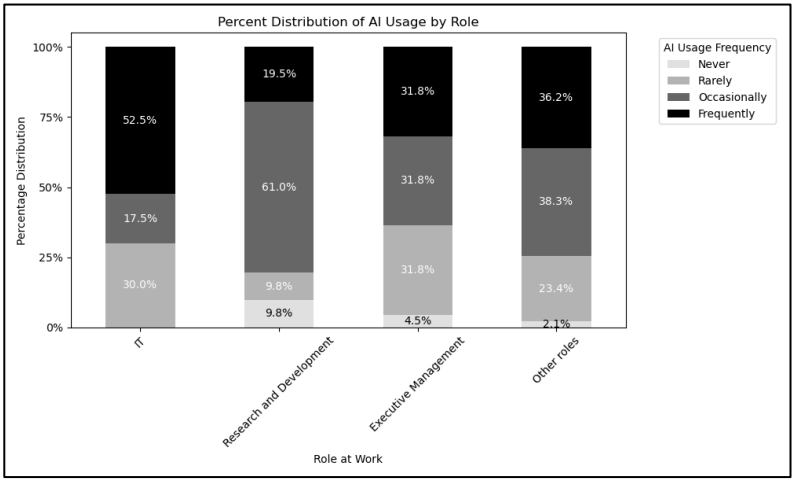



**Figure 11: Percent Distribution of AI Usage by Role**

Beside the maybe more expected result that IT related roles are reporting the most dominant frequent usage of AI systems, research and development staff seems to rely in total (so frequent and occasional together) the most (at 80.5%) on AI, significantly higher than the total of "other roles" standing at 74.5% (which is in itself impressive). Also noteworthy is the observation, that though the AI usage of executive and management personnel does not reach the average, but with its 63.% total still can be considered highly relevant.

With that in mind, looking at the data with respect to the connection between the roles at work and the respective trust in AI systems (illustrated by Figure 12) an interesting picture emerges.

As a starting point, we have to observe, that if we average across roles the fully, strongly and limitedly trusting individuals, and sum these average proportions, we can state, that 53.4% of respondents place some level of trust in AI systems, which on the one hand can be considered promising (it is a slim majority after all), but prominently points to the fact, that **establishing and enhancing trust in AI systems is a crucial challenge in their adoption**, and most likely some **systematic approaches, on the policy and toolset level have to be implemented**.

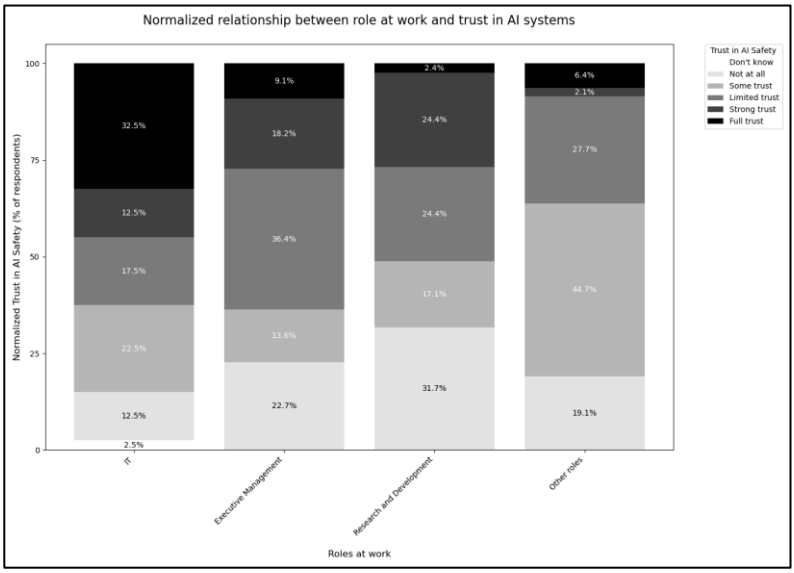



**Figure 12: Normalized relationship between role at work and trust in AI systems**

But going deeper into the details reveals more interesting insights: generally, IT roles are the very strongly trusting 62.5% if we sum fully, strongly and limitedly (or with the highest proportion of fully, 32.5%), which is in a sense as expected. We could theorize, that beyond the general openness towards technology in these roles the possibly deeper knowledge / understanding of the way of operation of AI systems can play an important role here.

It is also noteworthy, that management related roles exhibit a remarkably strong trust, 63.7% (summed), so the highest of all groups! This is especially important, since it means that we do not just see a kind of "grassroots" level of naive trust emerging, but IT (as the "topic owners") and management (as the ones giving directions to the organization) jointly exhibit positive attitudes about AI systems, that can foster faster adoption that if only the "general" employees would turn to AI in their work (as represented by either the 53% average or the 36.2% of the "other" roles).

Further unpacking the notion of "more knowledge about AI's workings leads to higher trust" - as we theorized in case of IT, we can endeavour to directly quantify this relationship by examining the connection between usage frequency and trust directly, irrespective of roles.

Figure 13 tries to make this connection tangible by visualizing the dependence of the level of subjective trust and the frequency patterns of usage. Clearly, we can see a considerable relationship emerging, or to quantify specifically: a Spearman correlation of 0.371 with a P-value of <0.00001, that we can definitely consider statistically significant. We can thus conclude, that one of the essential aspects of trust towards AI systems can be derived from specific knowledge, since **the more people use AI solutions, the more experience they have with them, hence the more they trust them**, which is in a sense extremely encouraging with respect to the further perspectives of adoption, but on the other hand points us towards the questions: what can we do, to enhance this trust even further?

On one hand: early and frequent exposure to AI technologies, so a culture and business investment that **enables employees to gather first hand, real life experience** is one of the most important factors we can think of. On the other hand, we can investigate some policy actions and practical technological solutions that can potentially boost trust, thus the level of adoption. In the remaining sections we focus our attention on these topics.

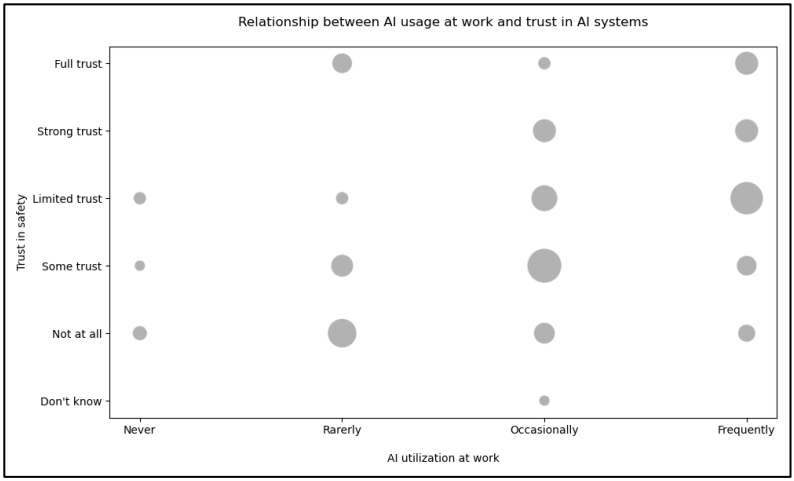



**Figure 13: Relationship between AI usage at work and trust in AI systems**

## Organizational readiness and fears

If we want to make systematic progress in deploying AI systems in the broadest set of business applications possible, we have to also take a sharp look at the flip side of the above detailed phenomena, namely the topic of fears.

The first topic that we tried to quantify is what the areas exactly are - phrased in quite broad strokes so as to allow more respondent opinions to converge - that professionals are concerned about with respect to the broad scale adoption of AI technologies throughout business life.

Figure 14 summarizes the fears, and makes it evident, that there is a clear hierarchy in the perceived level of risk with respect to topics. Most interestingly, the direct labour market impacts, so the potential of AI for causing job displacement lands squarely at the last place, therefore a 58.7% majority of respondents consider it as "non-issue", strongly contradicting the general economic narratives of the current day (though one could argue, the approx. 40% agreeing is by no means negligible), followed only closely by the topic of AI's inherent bias with 46.7% agreement, which might point to the fact that this topic has been already worked upon extensively in the last decade of AI progress. (It is well worth mentioning as a contrast, though, that the (Eurostat 2024) dataset reports

for the proportion of "Enterprises which have measures to check the results generated by AI technologies for possible biases towards individuals" in the few EU member states which have at al this data sub 1% frequency, so it is quite questionable, if the problem is solved or simply neglected...)

The question of erosion of human skills is the first issue where the respondents seem to be ever so slightly (50.7%) more concerned than not, which also points towards the perceived potential of the technology: if the general consensus would point towards AI as being "incompetent" or "useless", the issue with over-reliance and thus, the diminishing of human capabilities would not even emerge in the first place.

The topic of security vulnerabilities and of data privacy issues move hand in hand, a 60% majority consider them as noteworthy concerns, thus pointing to the fact, that clear policy guidelines and internal company regulations must be established to mitigate risks, which might in turn help establish and enhance trust.

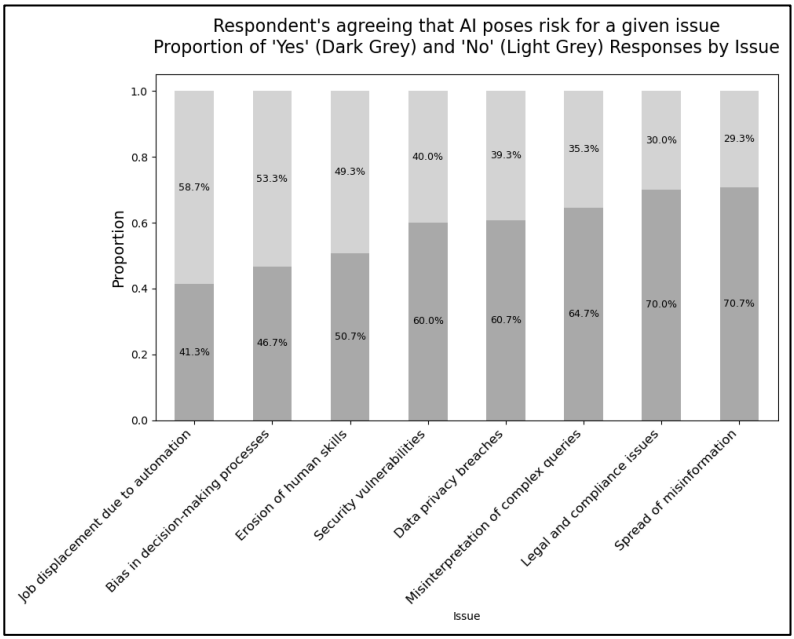



**Figure 14: Respondent's agreeing that AI poses risk for a given issue**

Despite the overall trend of increasing trust in AI systems, and our reasoning above, that there does exist a high level of baseline trust in AI capabilities, concerns about performance persist, particularly regarding the misinterpretation of user queries. **64.7%** of respondents highlighted this issue, indicating that AI systems, while powerful, still struggle to always accurately understand and respond to complex input. This suggests that improving AI comprehension and contextual reasoning remains a key challenge for trust-building in professional environments.

Legal and compliance concerns are even more pronounced, with **70.0%** of respondents identifying them as significant barriers to AI adoption. Notably, security risks themselves were not perceived as a primary issue; instead, regulatory compliance—especially with GDPR—seems to be the dominant concern. Though we do not have direct evidence, we can hypothesize, many respondents view data related regulations not necessarily as a safeguard but as an external burden that complicates AI deployment, suggesting that organizations may benefit from clearer regulatory guidance and standardized compliance frameworks to mitigate these concerns.

It is also notable for this point to ponder, that a general trend seems to exist in AI adoption representing the growing concern regarding data protection, privacy, and legal uncertainty, which has acted as a cooling factor on business enthusiasm for AI integration. Recent Eurostat data (Eurostat 2024) demonstrates a notable increase in enterprises citing privacy and legal concerns as barriers to AI adoption. Specifically, the percentage of EU enterprises refraining from AI use due to data protection and privacy concerns rose from **2.74% in 2023 to 4.89% in 2024**, reflecting a heightened sensitivity to regulatory compliance. This trend is even more pronounced in Germany, where **7.6% of enterprises** in 2024 reported privacy concerns as a primary deterrent to AI implementation. Similarly, the proportion of enterprises avoiding AI due to a lack of clarity about legal consequences increased from **2.94% to 5.26%** across the EU, with Germany again exhibiting a significantly higher share at **8.25%**. Hungary, while demonstrating a lower overall prevalence of these concerns, still experienced a measurable increase, with privacy-related worries growing from **1.54% in 2023 to 2.56% in 2024** and legal uncertainty concerns rising from **1.84% to 2.92%** in the same period. These statistics suggest that businesses are not only struggling with the technical implementation of AI but are also facing escalating regulatory and compliance-related

hesitations. The increased apprehension surrounding data protection laws, particularly the GDPR, underscores the necessity for clearer regulatory guidance, standardized compliance frameworks, and organizational policies that facilitate responsible AI deployment while maintaining legal certainty. Without targeted interventions to address these concerns, such as government-backed support programs, streamlined compliance mechanisms, or industry-wide best practices for AI governance, regulatory uncertainty may continue to hinder AI adoption at a critical juncture of technological advancement.

Perhaps the most striking result is the widespread apprehension regarding AI's role in spreading misinformation. **70.7%** of respondents regarded this as a critical issue, reflecting not only concerns about AI-generated errors but also a broader awareness of the risks associated with deliberate misuse of AI technologies. This highlights the necessity for stringent verification mechanisms, transparency in AI-driven content generation, and policies to prevent the amplification of misinformation at both the corporate and societal levels.

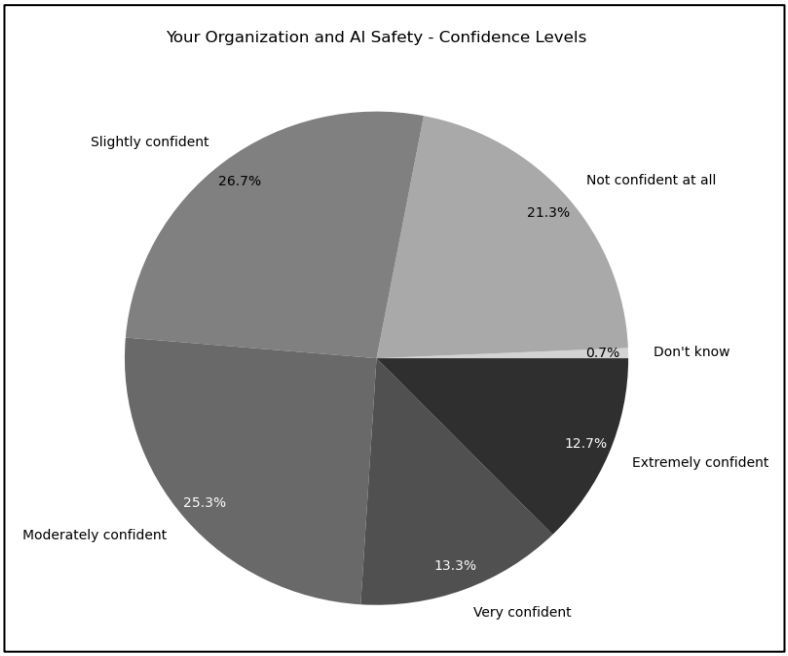



**Figure 15: Respondents perception of safety measures in their organizations**

To round out the picture, we have to mention the question of confidence by the respondents, that the safety measures utilized by their respective organizations are sufficient in counteracting the possible dangers analysed above. As Figure 15, the opinions are quite split: in total 48% of the respondents is slightly or not at all confident that the policies and measures are adequate, so we can see a pretty obvious gap and need for serious work on the internal policy and trust-enabling technology area. These topics should merit considerable investment!

### Tools for Establishing Trust

As discussed above, there are two main avenues for enabling trust in AI systems: organizational policies and technical evaluation tools. Here we focus on the latter—the recent tools and methodologies that allow systematic auditing and error correction of AI solutions, thereby establishing a technical basis for trust. These solutions have advanced rapidly and now make the abstract concept of "AI trust" measurable and actionable in practice. By deploying the right evaluation frameworks, organizations can obtain objective evidence of their AI systems' trustworthiness, which supports informed decisions about AI use and provides assurance to stakeholders.

### What kind of tools exist for establishing trust?

Several evaluation frameworks have emerged to quantify the performance and trustworthiness of AI systems. Notably, RAGAS, ARES and TrustMyAI are three recent methodologies developed for this purpose. All three target retrieval-augmented generation (RAG) scenarios—for example, Large Language Model (LLM) chatbots that retrieve and cite external documents—but their principles apply broadly to AI assistants based on large language models. Each framework shares a common goal: to assess how well an AI model uses retrieved context to produce relevant, accurate, and trustworthy answers. They typically evaluate along key dimensions such as **context relevance (does the model appropriately use the provided information?)**, **answer faithfulness (is the answer grounded in the source data?)**, **answer relevance (does it address the user's query fully and correctly?)**, and sometimes **transparency (does the system indicate where its information comes from?)**. These tools thus move beyond ad-hoc, human gut-checks toward formalized scoring

of AI outputs. In effect, frameworks like RAGAS and ARES (Mehrotra et al. 2023; Chen et al. 2023) enable a quantifiable trust score for AI responses, providing a structured way to judge whether an AI's output can be trusted in a business context. This represents a significant recent development in AI evaluation: trust, which was once hard to measure, can now be tracked with metrics.

## State-of-the-art methods for AI trust

### RAGAS

RAGAS (Retrieval-Augmented Generation Assessment) is an open-source framework introduced in 2024 to streamline RAG evaluation (Mehrotra et al. 2023). It provides a suite of automatic metrics that jointly analyse the retrieval and generation components of an AI system. Rather than relying on manual annotation or predefined answers, RAGAS uses reference-free checks—for instance, scoring how well the retrieved context matches the query and how closely the AI's answer stays true to that context. It measures attributes like relevance and factual consistency using language model scoring and similarity measures. In practical terms, RAGAS allows developers to quickly identify if a chatbot or QA system is hallucinating or diverging from its source material. Businesses have found this useful for rapid iteration: by running RAGAS metrics on test queries, teams can pinpoint weaknesses (e.g. irrelevant citations or unsupported statements) and improve the system before deploying it to real users. This improves the system's reliability and reduces the chance that end-users encounter incorrect information, directly boosting trust in the AI.

### ARES

ARES (Automated RAG Evaluation System) is another state-of-the-art methodology, emerging from the latest research to further automate trust evaluations (Chen et al. 2023). ARES introduces AI-based judges into the evaluation loop: it generates synthetic Q&A pairs and fine-tunes lightweight language models to act as specialized reviewers for each component of a RAG pipeline. These AI "judges" score the context relevance of retrieved documents, the faithfulness of the answer to those documents, and the overall answer quality. By training on synthetic data

(with a small set of human-verified examples for calibration), ARES minimizes the labour required for evaluation while still aligning with human judgment on correctness. The result is a largely automated scoring system that can be run continuously as an AI application evolves. In a business setting, ARES can serve as an ongoing monitor of an AI assistant's trustworthiness—flagging when the model's answers might be straying from verifiable facts or when retrieval quality drops. This proactive feedback loop helps organizations ensure their AI systems remain reliable over time, catching issues early (before users do) and maintaining consistency even as content or user queries change. Such capabilities are invaluable in high-stakes industries (finance, healthcare, etc.), where continuous validation of AI outputs is required for compliance and risk management.

**TrustMyAI**

TrustMyAI is Neuron Solutions' proposed evaluation framework, which builds on concepts from RAGAS and ARES while tailoring them to practical enterprise needs. Like the others, TrustMyAI combines rule-based metrics with AI-driven assessment to produce an overall "trust score" for model outputs. In essence, it uses RAGAS-style scoring rubrics (checking context usage, factual alignment, relevance) and augments them with an LLM-based evaluator that judges answer quality in context—similar to ARES's approach. The aim is to provide a comprehensive yet easy-to-use toolset for organizations to audit their AI systems. For example, TrustMyAI can be integrated into a company's chatbot platform to automatically rate each answer given to customers, highlighting low-trust responses for human review or retraining. By doing so, it operationalizes trust metrics: rather than trust being a vague notion, it becomes a dashboard item that product managers and AI developers can monitor and improve. The business impact is significant—TrustMyAI offers a repeatable process to ensure AI deployments meet predetermined trust thresholds. This means companies can set standards (e.g. "the answer must be 90% faithful to provided documents") and regularly verify compliance, which is especially important for meeting internal quality benchmarks or external regulatory requirements.

### Overview of RAG evaluation solutions

In sum, frameworks like TrustMyAI, RAGAS, and ARES exemplify the cutting edge of AI evaluation. They turn qualitative concerns (like honesty or relevance) into quantitative scores, making it feasible to validate AI systems at scale and iterate on them much like one would refine any other critical business process. Organizations adopting these tools gain a measure of control and confidence over AI behaviour that previously was difficult to attain.

### Platform integration of AI trust solutions

Importantly, the concepts behind these AI trust evaluation tools are being embraced and built into major AI platforms, reflecting an industry-wide move towards "responsible AI" by design.

### Microsoft Azure

For instance, Microsoft Azure now provides an integrated RAG evaluation framework as part of its AI offerings. Azure's architecture guidelines for LLM applications recommend measuring metrics such as groundedness, relevancy, and correctness of responses—essentially similar criteria to the frameworks above—to ensure that generated answers remain faithful to the retrieved data (Microsoft 2024). Microsoft has even introduced GPT-4-based evaluation metrics within Azure Machine Learning Studio, allowing developers to automatically grade their chatbot responses on these trust dimensions. This means that for example, if you would build a RAG-based Q&A system on Azure, you can leverage built-in tools to calculate how well your model is using its knowledge sources and where it might be making unfounded claims. By embedding such evaluation at the platform level, Azure simplifies the adoption of trust measures: even non-technical teams can get feedback on AI output quality in real time, without needing to assemble a custom evaluation pipeline from scratch. This streamlined evaluation process helps businesses quickly identify issues (like a wrong citation or an incomplete answer) and improve their models, ultimately leading to more trustworthy AI services deployed on Azure.

### Amazon Web Services

Similarly, Amazon Web Services (AWS) has invested in AI trust and monitoring tools through its SageMaker platform. Amazon SageMaker Clarify, for example, is a tool that helps developers detect bias in models and understand the reasoning behind predictions, providing transparency into how AI decisions are made. (Amazon Web Services 2025a) Clarify produces detailed reports on potential biases in datasets and model outputs (e.g. checking if a model's answers favour certain groups), and it offers feature importance analysis to explain why the model arrived at a given answer. These capabilities are crucial for building fair and transparent AI systems, which are key pillars of user trust. In addition, AWS offers SageMaker Model Monitor (Amazon Web Services 2025b), a managed service that continuously watches deployed models for quality issues. Model Monitor can automatically detect data drift, concept drift, and performance degradation in real time—for instance, alerting if users start asking questions outside the AI's training distribution or if the accuracy of responses declines over time. By receiving such alerts, businesses can take corrective action (retraining models, adjusting data inputs, etc.) before small issues snowball into major failures.

### Other key players in RAG Evaluation

Beyond Azure and AWS, other leading platforms are integrating trust and explainability features into their AI ecosystems. Google's Vertex AI platform offers tools for evaluating model fairness and explainability. For instance, Vertex Explainable AI provides feature-based and example-based explanations to enhance understanding of model decision-making processes.(Google Cloud 2025b)

IBM's Watsonx platform supports the development of Retrieval-Augmented Generation (RAG) applications, enabling enterprises to build AI solutions that generate factually accurate outputs grounded in information from knowledge bases. In addition, IBM provides some of the aforementioned frameworks for evaluating and monitoring RAG pipelines, such as the Ragas framework, assessing the performance of RAG systems.(Gutowska and Lukashov 2024)

## Overview of solutions for AI trust

The convergence of these platform-level integrations from the aforementioned global software companies amplify the impact of frameworks like RAGAS, ARES, and TrustMyAI. By embedding evaluation and monitoring features directly into their products, cloud providers operationalize key concepts such as measuring groundedness, faithful-ness, and bias at scale. This alignment with business demands for trusted AI facilitates the deployment of reliable and transparent AI solutions across various industries. For enterprises, this means adopting AI trust tools is becoming easier and more cost-effective. A team using Microsoft or AWS can leverage built-in evaluators and dashboards as a starting point, then extend with specialized frameworks for additional rigor if needed. The result is a robust end-to-end approach to AI trust—from development to deployment—that aligns with both technical best practices and emerging regulatory standards. In practice, organizations that embrace these integrated trust solutions are better positioned to deploy AI innovations with confidence. They can demonstrate to clients and regulators that measurable safeguards are in place, and they gain a competitive advantage by offering AI products and services distinguished by their reliability and accountability. In short, the convergence of new AI trust evaluation frameworks and their adoption by major platforms is making trustworthy AI a tangible, achievable goal for businesses today.

## Value of trust tools

After his short survey of truest enabling technologies, the question naturally arose, what is the perceived value of these approaches in the eyes of the practitioners, the professionals in the field who responded to our survey?

We directly stated this question, with some short explanation about what we think such a trust enabling technology would look like, trying to estimate the attitudes of respondents towards them.

As Figure 16 illustrates, there is broad consensus about the value of trust enabling tools: in total **72% of respondents agree, that a tool that for example assesses the output of Large Language Model based solutions is at least moderately, very or extremely valuable**, pointing to the very specific felt need for some more systematic evaluations than just purely "manual" evaluation by humans.

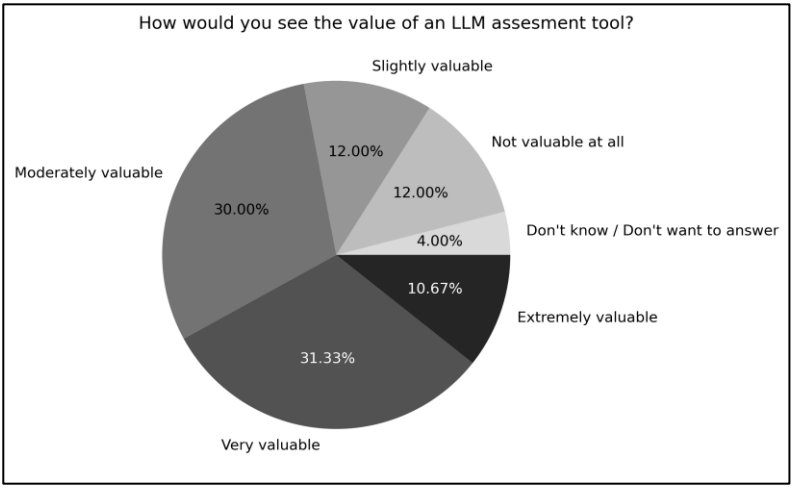



**Figure 16: Participants perception of the value of an LLM assessment tool**

As Figure 16 illustrates, there is a broad consensus about the value of trust enabling tools: in total **72% of respondents agree, that a tool that for example assesses the output of Large Language Model based solutions is at least moderately, very or extremely valuable**, pointing to the very specific felt need for some more systematic evaluations than just purely "manual" evaluation by humans.

**We consider this finding very significant, and would thus highly encourage business decision makers to adopt such technologies, methodologies.**

As for a more detailed, role based analysis of this value perception we can see by studying Figure 17 that generally the more "hands on" roles are laying emphasis on the usage of LLM evaluation methods, but even in case of management and executive roles, more than 66% of the respondent agree, that such solutions carry at minimum moderately value for the organizations, making this an unanimously perceived need on the market.

As a further important remark, the organizational size distribution of this value judgement (illustrated by Figure 18) reveals that especially large organizations exhibit a particularly strong perceived need for AI evaluation methodologies. Statistical analysis supports this observation: the relationship between organizational size and the perceived value of an LLM assessment tool is statistically significant $\chi^2 = 55.82,\ p < 0.001,$

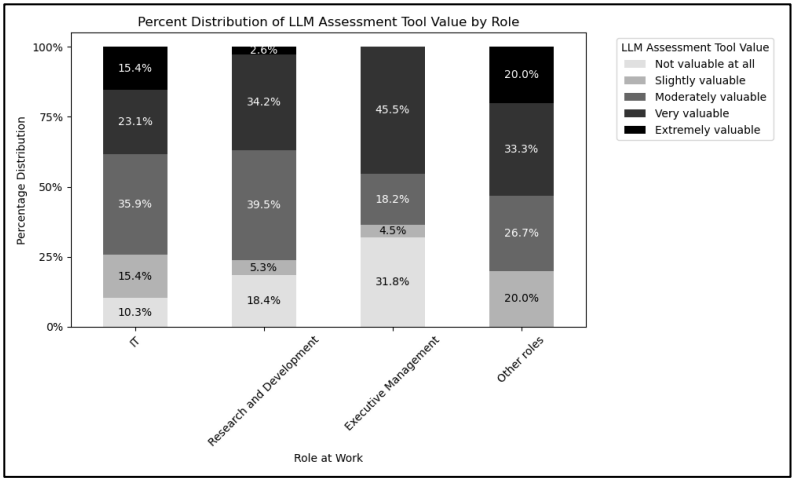



**Figure 17: Percent Distribution of LLM Assessment Tool Value by Role**

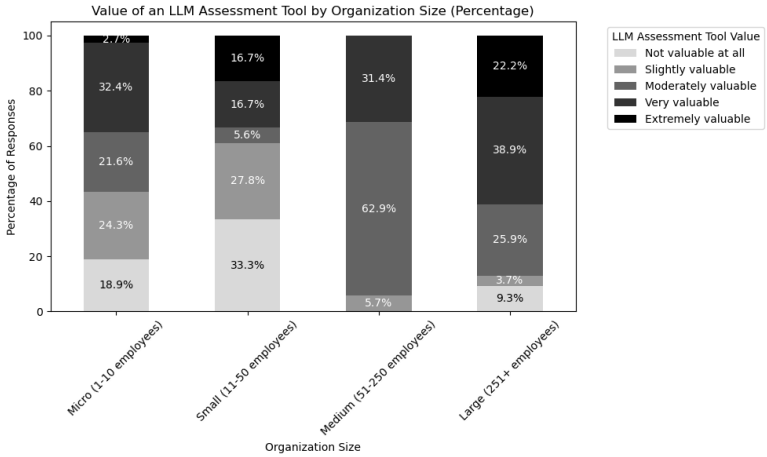



**Figure 18: Value of an LLM Assessment Tool by Organization Size**

with a Cramér's V of 0.3595. According to the interpretative scale proposed by (Rea and Parker 1992), this represents a moderate association, indicating a meaningful link between company size and attitudes toward AI evaluation solutions.

## Discussion of Results

Our findings underscore that **trustworthiness is a cornerstone of AI adoption** in business. Even in our tech-savvy sample, only about **53.4% of respondents expressed at least some trust in AI systems**, a slim majority that highlights a significant trust gap. This gap matters because trust (or the lack thereof) directly influences the degree to which organizations embrace AI. Notably, we observed a virtuous cycle between **AI experience and trust**: respondents **who use AI more frequently tend to trust it more** (Spearman's rho = 0.37, $p < 0.00001$). This correlation suggests that familiarity breeds confidence—an encouraging sign that **hands-on exposure can build trust**.

However, experience alone may not bridge the trust gap for everyone. The key question that emerges is **what can organizations do** to enhance trust further and ensure that scepticism about AI does not hinder its potential benefits? To answer this, we place our results in the context of the evolving AI governance landscape and derive practical strategies for businesses.

## Policy Implications

The AI governance landscape is rapidly maturing, setting clear expectations for organizations to prioritize trustworthiness. Four major international governance or regulatory frameworks shape this environment, as we will discuss below.

## EU AI Act

The European Union's Artificial Intelligence Act is the first comprehensive legal framework for AI, adopting a risk-based approach to ensure trustworthy AI.(European Commission 2021). The purpose of the AI Act was to build a comprehensive risk-based regulatory framework to ensure that AI systems are safe, transparent, and accountable.

It categorizes AI systems into four risk levels: unacceptable, high, limited, and minimal. High-risk systems (such as those used in critical domains like healthcare or finance) must satisfy stringent criteria including thorough risk assessments, data quality control, human oversight, and transparency measures. The Act also mandates clear documentation and establishes mechanisms for auditing and tracing AI-driven decisions,

thereby aligning AI development with public trust and safety requirements.

The AI Act entered into force on 1 August 2024, and will be fully applicable 2 years later on 2 August 2026, with some exceptions:

- Prohibitions and AI literacy obligations entered into application from 2 February 2025
- The governance rules and the obligations for general-purpose AI models become applicable on 2 August 2025
- The rules for high-risk AI systems - embedded into regulated products - have an extended transition period until 2 August 2027 (European Commission 2024).

In practice, the AI Act means for boards and executives that **they must treat AI risks much like financial or legal risks**, with formal processes to audit and trace AI decisions so that systems remain aligned with public safety and trust. Non-compliance could **not only invite regulatory penalties but also erode public trust**, a risk few businesses can afford in an era where nearly 44% of organizations are already integrating AI into customer-facing functions, based on our survey data.

The EU's regulatory push thus serves as a wake-up call: organizations must be proactive in building trustworthy AI before mishaps lead to reputational damage or legal liability. As our respondents noted, errors in consumer-facing AI (e.g. chatbots giving wrong answers or exhibiting bias) can diminish brand trust and even cause financial losses.

As we see in this context, the high-level message of the EU AI Act for business leaders is that **AI trust is now a governance imperative**. The Act requires companies to institute robust oversight for higher-risk AI applications, for example, by requiring thorough risk assessments, documentation, and human oversight for AI used in sensitive domains.

### OECD AI Principles and NIST AI Risk Management Framework

Importantly, this European approach aligns with global trends. The OECD AI Principles provide international, consensus-based guidelines emphasizing human-centric and ethical use of AI. The principles include promoting inclusive growth, human-centred fairness, transparency,

robustness, security, and accountability. They act as foundational guidelines influencing national and industry-level standards, contributing significantly to global alignment around trustworthy AI (OECD 2019).

Likewise to the OECD AI Principles, the **U.S. NIST AI RMF**, while voluntary, gives a practical blueprint to identify, assess, and mitigate AI-related risks through its core functions: **Govern, Map, Measure and Manage**. By operationalizing AI principles into actionable strategies, the RMF helps businesses achieve compliance and cultivate trust in AI deployment (NIST 2023).

These aforementioned frameworks translate abstract principles into a risk management process that companies can adopt internally, bridging the gap between lofty ideals and day-to-day practice. In essence, there is an emerging global alignment on what "trustworthy AI" entails—**transparency, fairness, accountability, and security are recurring pillars**. Businesses operating across borders can take note that adhering to these principles will not only ease compliance with any one regulation but also bolster their credibility in the eyes of customers and partners.

### ISO/IEC 38507:2022 (AI Governance Standard)

Another key piece of the governance puzzle is corporate-level standards, such as **ISO/IEC 38507:2022 on AI governance**. ISO/IEC 38507:2022 provides detailed guidance for organizational governance of AI technologies. It reinforces corporate accountability for AI systems and stresses the importance of governance oversight at the board level. According to the standard, boards must set policies and oversee AI deployment processes to ensure ethical use and minimize risks associated with AI deployment (ISO/IEC 2022).

Boards are expected to set clear AI policies and oversee their implementation, ensuring ethical use and risk mitigation as part of corporate governance. This also corresponds to our previous findings that trust in AI is now a strategic issue, demanding executive attention. Our survey results show that trust in AI is not just a concern of the technical staff. In fact, we found that management respondents exhibited among the highest trust levels in AI (nearly 64% reporting some degree of trust).

This suggests that leaders are increasingly optimistic about AI's potential—an optimism that needs to be backed by solid governance. By

formally integrating AI oversight into corporate governance, organizations signal both internally and externally that they are serious about managing AI's risks and rewards. Such signals can enhance stakeholder confidence, aligning with what the emerging regulations and standards seek to achieve.

**Building AI Trust in practice: Organizational strategies**

Translating these policy insights into practice, our study points to several strategies that organizations can pursue to foster trust in AI.

**Trust Evaluation in AI Processes**

First, businesses should embed systematic trust evaluation into their AI development and deployment processes. Rather than relying on ad-hoc or purely manual assessments of whether an AI system is trustworthy, organizations can adopt structured evaluation methodologies as a routine "quality control" for AI.

Our respondents clearly see the merit in this approach: an overwhelming **72% indicated that tools for assessing AI outputs (e.g. evaluations of LLM responses) would be at least moderately valuable**. This consensus signals a felt need in the industry for more **rigorous, repeatable validation of AI systems** beyond what human intuition can provide. In line with this, we recommend that companies establish regular AI performance audits focused on trust metrics like accuracy, fairness, transparency, and robustness.

Emerging evaluation frameworks make this feasible. For example, methodologies such as **RAGAS** and **ARES**—originally developed for **retrieval-augmented generation (RAG) systems**—use a combination of automated metrics and AI-based judges to rate how well a model's answers align with provided context and factual correctness. Our own proposed framework, **TrustMyAI**, builds on similar concepts, combining RAGAS-style scoring with LLM-based judging to systematically measure an AI system's context relevance, answer faithfulness, and overall quality.

The specifics of these tools aside, the broader point is that **quantitative trust assessments are becoming accessible**. By adopting such tools (or others like them), organizations can obtain objective evidence of their AI systems' trustworthiness. This has two major business implications:

internally, it supports more informed decision-making about where AI can be safely deployed; externally, it provides **assurance to regulators and clients**. In fact, as AI regulations tighten, having **quantifiable trust metrics** can demonstrate compliance with requirements for transparency and risk management, as envisioned by frameworks like the EU AI Act and NIST RMF.

In short, systematic evaluation methodologies offer a practical bridge between high-level governance mandates and on-the-ground AI operations, making the abstract concept of "AI trust" measurable and actionable.

### Internal AI Policies and Culture

Second, organizations should **align their internal policies and culture with the emerging best practices on AI ethics and governance**. This involves training and awareness to ensure that everyone from developers to executives understands AI's limitations and risks. Our results show that **security and data privacy are top-of-mind concerns**, with about 60% of respondents citing these as noteworthy issues with AI.

To address such concerns, companies need clear internal guidelines—for example, data handling rules for AI, bias mitigation checklists, and incident response plans—that translate external principles into daily workflow checkpoints. By instituting these policies, firms create an environment where compliance and ethical considerations are second nature, not afterthoughts.

Moreover, **transparency with users and stakeholders is paramount**. Organizations should openly communicate what their AI systems can and cannot do, including known limitations or error rates. This kind of honesty was echoed in our recommendations for enhancing AI literacy and transparency.

**When users are educated** about how an AI-driven decision is made (and what safeguards are in place), **they are more likely to trust the outcome**. Such openness is not only a good practice but may soon be a legal requirement for high-impact AI systems (as transparency is a common thread in AI regulations).

## AI Certification and Labelling

Third, there is value in exploring **certification and labelling mechanisms for AI trustworthiness**. Just as industries have certifications for quality management or cybersecurity, AI is likely to move in this direction. Policymakers and industry groups are considering **trust seals** or standardized **trust scores** for AI.

Our study's context suggests that tools like TrustMyAI, RAGAS, ARES and similar evaluation frameworks could contribute to these certification programs by providing consistent criteria to score AI systems. For organizations, participating in such certification schemes could offer a **competitive advantage**—a certified AI product may inspire greater confidence among customers and partners. It's a way to **signal trustworthiness** in a crowded marketplace.

Furthermore, establishing industry-wide benchmarks for trust can help avoid a scenario where each company invents its own metrics (leading to confusion and "trust washing"). Instead, with common standards, businesses can compete on delivering genuinely trustworthy AI, not just marketing claims. Given that **larger organizations in our survey expressed a particularly high need for AI evaluation methodologies**, we anticipate that leading firms will push for these sorts of standardized trust evaluations, which smaller enterprises might then adopt as they become more accessible.

## AI Literacy

Finally, fostering trust also means addressing the **human element** within organizations. Trust in technology often hinges on how well the people using or affected by that technology understand and accept it. Our results hinted that **when management and IT teams jointly trust AI** (as we saw with their relatively high trust levels), **AI initiatives can gain momentum**.

To replicate this, companies should invest in **AI literacy programs** and cross-functional dialogues about AI. By educating employees on AI's capabilities and pitfalls, and by engaging stakeholders in setting AI ethics guidelines, organizations create a culture where building trustworthy AI is a shared mission. This cultural alignment complements the technical and

policy measures discussed above, ensuring that trustworthiness isn't just a box to tick, but a core value driving AI adoption.

### Key takeaways: AI Trust for competitive advantage

Stepping back, **why do these findings matter?** In practical terms, they illuminate a path for organizations to navigate the **dual challenge of leveraging AI and managing its risks**. Our research confirms that businesses see immense potential in AI—for instance, a majority of large enterprises in our survey regard AI as important or even critical to their strategy.

However, realizing this potential hinges on trust: without stakeholder trust, AI projects may face resistance, underutilization, or backlash when things go wrong. The discussion above outlines how organizations can respond. By **anticipating regulations** like the EU AI Act and embracing the spirit of those rules proactively, companies can avoid playing catch-up and instead position themselves as **trust leaders**. By adopting **rigorous evaluation tools and best practices**, they can directly address the concerns that make people wary of AI, whether it's an employee unsure about an algorithmic recommendation or a customer hesitant about an AI-driven service. The payoff is twofold: internally, smoother AI adoption and innovation, and externally, enhanced reputation and **regulatory compliance**.

In essence, our results and analysis can serve as a **note for organizations to convert AI trustworthiness into a strategic asset**. Firms that invest in AI governance and trust-building not only **mitigate risks**—such as biases, errors, or security breaches—but also create an environment where AI can be deployed with confidence and creativity. Trust in AI is not just a moral or compliance issue, but a **business enabler**. It determines which companies will successfully scale their AI deployments and win public trust, and which might stumble due to unseen pitfalls. As AI technologies continue to evolve, those organizations that have woven trust into their AI strategy will be best placed to harness AI's transformative power responsibly and sustainably.

Our study contributes to this forward-looking perspective, showing that while the challenges of AI governance are significant, the tools, frameworks, and insights to tackle them are increasingly within reach. By

combining high-level governance foresight with on-the-ground practical measures, organizations can confidently navigate the emerging era of AI with both **innovation and integrity**.

**Acknowledgement**

This research was conducted within the framework of the Infota project "Development of artificial intelligence-based methods to support the innovation capabilities of SMEs" supported by the Batthyány Lajos Foundation.

## Appendix: Specific questions of the survey

### Section 1: AI and Your Organization

1. **Country of Operations**

   Please write the name of the country where your organization operates.

2. **Your Role in the Organization**

   Please select the role that best describes your position within the organization:
   - IT
   - Finance
   - Human Resources
   - Operations
   - Sales
   - Marketing
   - Research and Development
   - Executive Management
   - Research
   - Don't Know / Don't Want to Answer
   - Other

3. **Organization Size**

   Please select the size of your organization based on employee count:
   - Micro (1-10 employees)
   - Small (11-50 employees)
   - Medium (51-250 employees)
   - Large (251+ employees)
   - Don't know / Don't want to answer

4. **Organization Industry**

   Please select the industry your organization operates within:
   - Technology
   - Finance
   - Healthcare
   - Manufacturing

- Education
- Retail
- Don’t know / Don’t want to answer
- Other

**Section 2: Awareness of Artificial Intelligence (AI)**

5. **Experience in interacting with AI in professional life**

   How experienced are you in interacting with AI in your professional life?
   - Not experienced at all
   - Slightly experienced
   - Moderately experienced
   - Very experienced
   - Extremely experienced

6. **Frequency of AI Use**

   How frequently do you or your organization use AI technologies in your professional activities?
   - Never
   - Rarely
   - Occasionally
   - Frequently
   - Always

7. **Experiences with Inappropriate AI Behaviors**

   Have you experienced or heard about inappropriate behaviours from AI models in a professional context?
   - Never
   - Rarely
   - Occasionally
   - Frequently
   - Always

8. **Current AI Use Cases**

   What are the top 5 use cases for AI within your organization? Please list them.

9. **Business Use Cases for AI**

For each scenario below, indicate if your organization currently uses AI and to what extent:

– Internal Use Case: Accessing internal policies
- Not used
- Minimally used
- Moderately used
- Extensively used
- Core function

– Customer-Facing Use Case: Interactive access to FAQs
- Not used
- Minimally used
- Moderately used
- Extensively used
- Core function

– Customer-Facing Use Case: Sales chatbot
- Not used
- Minimally used
- Moderately used
- Extensively used
- Core function

10. **Planned Business Use Cases for AI**

For each scenario below, indicate your organization's intended level of AI integration:

– Internal Use Case: Enhancing internal policy access
- Not planned
- Planned as a minor feature
- Planned as a moderate feature
- Planned as a major feature
- Planned as a core function

– Customer-Facing Use Case: Enhancing interactive FAQ access
- Not planned
- Planned as a minor feature

- Planned as a moderate feature
- Planned as a major feature
- Planned as a core function

– Customer-Facing Use Case: Enhancing sales chatbot functionality
- Not planned
- Planned as a minor feature
- Planned as a moderate feature
- Planned as a major feature
- Planned as a core function

**Section 3: Risks and Challenges of LLMs in Business**

11. **Potential harm from LLMs in business**

In your view, what are the potential risks or ways that Large Language Models (LLMs) like GPT could cause harm or pose challenges in a business setting? Please select all that apply, and feel free to add any additional concerns.

– Data privacy breaches
– Spread of misinformation
– Bias in decision-making processes
– Job displacement due to automation
– Security vulnerabilities
– Misinterpretation of complex queries
– Legal and compliance issues
– Erosion of human skills
– Other (please specify)

**Section 4: Assessment of AI Trust**

12. **AI Initiative Budget**

If your company has AI initiatives, how much do you plan to spend on these initiatives annually? (Specify in Euros)

13. **Company Competence in AI Safety**

How confident are you in your company's competence in ensuring the safety and ethical use of AI technologies?

- Not confident
- Slightly confident
- Moderately confident
- Very confident
- Extremely confident

14. **Value of an LLM Assessment Tool**

    How valuable would you find a third-party assessment tool for ensuring the appropriate behavior of your AI systems, including LLMs?
    - Not valuable
    - Slightly valuable
    - Moderately valuable
    - Very valuable
    - Extremely valuable

15. **Willingness to Pay for AI Assessment**

    How much would you be willing to pay annually for a third-party service that assesses and ensures the appropriate behavior of your AI systems? (Please specify in Euros. Write 0 if you are not willing to pay.)

## References

AIAAIC. (2024). *Driver persuades chatbot to sell car for USD 1.* AI, Algorithmic, and Automation Incidents Repository. https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/driver-persuades-chatbot-to-sell-car-for-usd-1

Amazon Web Services. (2025a). *Detect bias and explain predictions with Amazon SageMaker Clarify*. https://aws.amazon.com/sagemaker/clarify/

Amazon Web Services. (2025b). *Monitor models for quality with Amazon SageMaker Model Monitor.* https://aws.amazon.com/sagemaker/model-monitor/

Anthropic. (2025). *The Anthropic Economic Index.* https://www.anthropic.com/news/the-anthropic-economic-index

Appenzeller, G. (2024, November 12). *Welcome to LLMflation—LLM inference cost is going down fast.* https://a16z.com/llmflation-llm-inference-cost/

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … Liang, P. (2022). *On the opportunities and risks of foundation models.* arXiv preprint. https://arxiv.org/abs/2108.07258

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Amodei, D. (2020). *Language models are few-shot learners. Advances in Neural Information Processing Systems, 33*, 1877–1901.

Chen, Y., Yang, J., Yu, Y., & Luo, J. (2023). *ARES: An automated evaluation framework for retrieval-augmented generation systems.* arXiv preprint. https://arxiv.org/abs/2311.09476

Dechter, R. (1986). Learning while searching in constraint-satisfaction problems. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 178–185).

Ea, H. (2024). *Air Canada forced to honour chatbot offer. Law Society Journal.* https://lsj.com.au/articles/air-canada-forced-to-honour-chatbot-offer/

Eurostat. (2024). *Artificial intelligence use in enterprises.* https://doi.org/10.2908/ISOC_EB_AI

European Commission. (2021). *Proposal for a regulation laying down harmonised rules on artificial intelligence (AI Act).* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

European Commission. (n.d.). *AI Act—Shaping Europe's Digital Future.* https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Fradkov, A. L. (2020). *Early history of machine learning. IFAC-PapersOnLine, 53*(2), 1385–1390. https://doi.org/10.1016/j.ifacol.2020.12.1888

Google Cloud. (2025a). *Introduction to model evaluation for fairness.* https://cloud.google.com/vertex-ai/docs/evaluation/intro-evaluation-fairness

Google Cloud. (2025b). *Introduction to Vertex Explainable AI.* https://cloud.google.com/vertex-ai/docs/explainable-ai/overview

Gutowska, A., & Lukashov, V. (2024). *Evaluate RAG pipeline using Ragas in Python with watsonx.* IBM. https://www.ibm.com/think/tutorials/ragas-rag-evaluation-python-watsonx

ISO/IEC. (2020). *ISO/IEC TR 24028:2020—Overview of trustworthiness in artificial intelligence.* https://www.iso.org/standard/77608.html

ISO/IEC. (2022). *ISO/IEC 38507:2022—Governance implications of the use of artificial intelligence by organizations.* https://www.iso.org/standard/56641.html

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … Riedel, S. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks.* arXiv preprint. https://arxiv.org/abs/2005.11401

Mehrotra, S., Bhatia, K., Sharma, R. A., & Gupta, P. (2023). *RAGAS: Retrieval-augmented generation assessment for open-domain question answering.* arXiv preprint. https://arxiv.org/abs/2309.15217

Microsoft. (2024). *Assess AI systems by using the Responsible AI dashboard.* https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai-dashboard?view=azureml-api-2

NIST. (2023). *AI Risk Management Framework (AI RMF 1.0).* https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf https://doi.org/10.6028/NIST.AI.100-1.jpn

OECD. (2019). *Recommendation of the Council on Artificial Intelligence.* https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

OECD. (2024). *Artificial Intelligence—OECD Policy Issues.* https://www.oecd.org/en/topics/policy-issues/artificial-intelligence

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Amodei, C. (2019). *Language models are unsupervised multitask learners.* arXiv preprint. https://arxiv.org/abs/1901.11504

Singla, A., Sukharevsky, A., Yee, L., & Chui, M. (2024, May). *The state of AI in early 2024.* McKinsey & Company. https://www.mckinsey.com/~/media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai/2024/the-state-of-ai-in-early-2024-final.pdf

Szabados, L. (2024). *AI's effect on labour: What does economic literature say? The International Journal of Engineering and Science, 13*(5), 328–350.

https://doi.org/10.60030/ALMAMS.c202501PNAAaF

# From Perceived Disadvantage to Competitive Advantage: Generative AI and the SME Opportunity[1]

Ottó Werschitz

founder
Cognivate AI
e-mail: otto.werschitz@cognivate-ai.com
https://orcid.org/0009-0008-0937-0768

## Abstract

Small and medium-sized enterprises (SMEs) across Europe are widely perceived as lagging behind large firms in the adoption of artificial intelligence (AI). This article argues that generative AI fundamentally reshapes this narrative. Drawing on practitioner experience and recent empirical insights, it shows that generative AI lowers long-standing barriers related to cost, expertise and access, while at the same time placing new emphasis on organisational readiness, human factors and strategic intent. The article reframes AI not as a technological silver bullet, but as an enabler whose value depends on industry context, digital intensity and deliberate use. By examining SME heterogeneity, knowledge-intensive use cases and practical considerations for safe and productive adoption, the article highlights how perceived disadvantages can, under the right conditions, become sources of competitive advantage.

***Keywords:*** *Generative AI, Small and medium-sized enterprises (SMEs); AI adoption, Digital transformation, Organisational readiness, Competitive advantage, Practitioner perspective*

[1] This article builds on and further develops ideas originally presented at the "Us and AI" — Collaboration with Artificial Intelligence in the World of Work conference, held at the University of Pannonia in Veszprém, Hungary, in October 2025.

## Introduction - A persistent gap behind the headlines

Despite the global visibility of generative AI and the intense public discourse around its transformative potential, small and medium-sized enterprises (SMEs) in Hungary remain at an early stage of meaningful adoption. Recent empirical evidence confirms what many practitioners observe in daily work with business owners: openness toward AI is relatively high, but actual usage is limited and fragmented. A nationwide survey among Hungarian SMEs shows that while more than four-fifths of entrepreneurs express no fundamental fear of AI tools, adoption is constrained by very practical barriers—most notably lack of expertise, insufficient IT preparedness, financial limitations and uncertainty around data protection (VOSZPORT, 2024).

Importantly, this is not a uniquely Hungarian phenomenon. Across the European Union, SMEs consistently lag behind their larger counterparts in adopting and scaling AI technologies. OECD analysis shows that European SMEs are significantly less likely than large enterprises to adopt AI technologies across core business functions, despite increasing availability and declining entry costs (OECD, 2024). This persistent adoption gap underscores a broader structural challenge: while the tools are technically accessible, the organizational capacity to integrate and leverage them remains uneven.

From a practitioner's perspective, the core issue is not simply technology access, but capability, confidence and strategic navigation. The paradox is stark—generative AI has lowered barriers to entry that once favoured scale, yet many SMEs are not translating this accessibility into competitive advantage. Understanding why this potential remains largely untapped is a necessary starting point for any realistic discussion about how SMEs can turn perceived disadvantage into competitive advantage.

## Why this article does not treat AI as a silver bullet

Against this backdrop, it is tempting to search for quick fixes—tools, platforms, or turnkey "AI solutions" that promise immediate productivity gains or competitive advantage. In practice, this mindset often leads to disappointment. Generative AI is frequently approached either with inflated expectations or with cautious scepticism, neither of which helps SMEs make meaningful progress.

In this article, we deliberately take a different stance. We argue that artificial intelligence—generative AI in particular—is not a silver bullet

that automatically compensates for structural disadvantages in scale, capital, or market power. At the same time, we contend that it is far more than a marginal efficiency tool. When approached with clear intent and realistic expectations, AI can act as a powerful enabler: lowering entry barriers, amplifying existing capabilities and allowing smaller organizations to operate with a level of sophistication that previously required far greater resources.

From a practitioner's perspective, the critical question is therefore not *whether* SMEs should adopt AI, but *how* and *why*. This article explores how SMEs can navigate AI adoption in a way that is strategically grounded rather than technology-driven; why mindset, capability development and use-case selection matter more than tool choice; and how generative AI can be integrated into everyday business practices without requiring large-scale transformation programmes.

*By focusing on real patterns observed in SME contexts—rather than abstract maturity models or idealized case studies—we aim to show how perceived disadvantages can, under the right conditions, be reframed into sources of competitive advantage.*

## What we mean by AI—and why generative AI changes the equation for SMEs

Before discussing adoption paths and practical benefits, it is useful to clarify what we mean by *artificial intelligence* in this article. In business practice, AI primarily refers to systems based on machine learning (ML) and deep learning (DL): data-driven approaches that learn patterns from historical data and use them to make predictions, classifications, or recommendations. These systems do not follow fixed rules; their behaviour is shaped by data and training. Within this broad category, generative AI represents a distinct—and strategically important—subset. While often described by its ability to "generate" text, images, or code, generative AI is better understood as a general-purpose interface to intelligence: capable of reasoning, summarising, translating, analysing and interacting across a wide range of business contexts, rather than solving one narrowly defined task.

In this article, we therefore focus primarily on the benefits and implications of generative AI–based tools for SMEs. This is a deliberate choice. Traditional ML solutions—such as demand forecasting models, recommendation engines, or custom optimisation algorithms—can certainly deliver value and in some cases they are indispensable. However,

they typically require higher levels of data readiness, technical expertise, integration effort and ongoing maintenance. For many SMEs, these requirements represent a significant entry barrier. By contrast, generative AI tools are already embedded in widely used business software and can be applied immediately to everyday knowledge work, making them a far more accessible starting point.

This accessibility is reinforced by the rapidly expanding ecosystem of cloud-based, subscription-driven generative AI services. For organisations already working within the Microsoft ecosystem, **Microsoft 365 Copilot** provides native AI capabilities across email, documents, spreadsheets and collaboration tools (Microsoft, 2025a; Microsoft, 2025b). Similarly, companies using Google Workspace can build on generative AI services delivered through **Google Vertex AI**, which integrates advanced models into familiar productivity and analytics environments (Google Cloud, n.d.-a; Google Cloud, 2025). Beyond these platforms, most major generative AI providers now offer enterprise-grade subscriptions with governance, security and compliance features. In parallel, a growing number of specialised tools have emerged—for example **Perplexity** for research and information synthesis (Perplexity, n.d.), alongside solutions focused on presentations, visual design, marketing content, or software development (CB Insights, 2024).

Taken together, the current generation of generative AI tools demonstrates why the entry barrier for SMEs has fundamentally shifted. In practice, this low threshold is driven by a combination of structural and operational characteristics:

- **No upfront capital investment (CapEx):** Most generative AI solutions operate on a subscription-based or pay-as-you-grow model, allowing SMEs to start small and scale usage in line with actual value creation rather than speculative investment.
- **No requirement for IT or data science expertise:** These tools are designed for business users, not technology specialists, enabling meaningful use without in-house AI, software development, or advanced analytics capabilities.
- **Rapid onboarding at a basic level:** Core functionalities can be learned quickly and simple, high-impact use cases can be applied almost immediately, lowering the cognitive and organisational cost of first adoption.
- **Integration into (small) enterprise environments:** Generative AI tools are increasingly embedded in familiar business platforms and

workflows, making them usable within existing organisational contexts rather than as standalone technologies.
- Enterprise-grade information security commitments: Leading providers offer formal data protection, compliance and security assurances, addressing one of the most common concerns among SME decision-makers.

Together, these characteristics explain why generative AI represents not merely a new technology, but a structurally different adoption opportunity—one that allows SMEs to engage with advanced AI capabilities earlier, faster and with significantly lower risk than in previous AI waves.

*From a practitioner's perspective, this combination—general-purpose capability, widespread availability and low initial commitment—is what makes generative AI uniquely relevant for SMEs today. It does not eliminate strategic choices or capability challenges, but it fundamentally reshapes where and how organisations can begin.*

**Why SMEs Can Move Faster Than They Think**

Beyond technology availability, SMEs also possess several structural characteristics that can actively support faster and more effective AI adoption—often in ways that large organisations struggle to replicate. In practice, we repeatedly see three enabling factors:

- **Greater speed in decision-making and implementation:** Compared to large enterprises, SMEs typically face fewer layers of approval, allowing them to experiment, decide and deploy AI-supported solutions much more rapidly.
- **Looser internal rules and governance structures:** While not without risks, less rigid internal regulations often make it easier to try new tools, adapt workflows and iterate without lengthy compliance or procurement processes.
- **More flexible job roles:** Broader and less strictly defined roles enable employees to incorporate AI into their daily work more organically, without waiting for formal role redesign or process reengineering.
- Taken together, these characteristics mean that SMEs are not merely passive recipients of AI innovation. Under the right conditions, they can act faster, learn earlier and adapt more fluidly—turning organisational agility into a genuine advantage in the age of generative AI (see Figure 1).

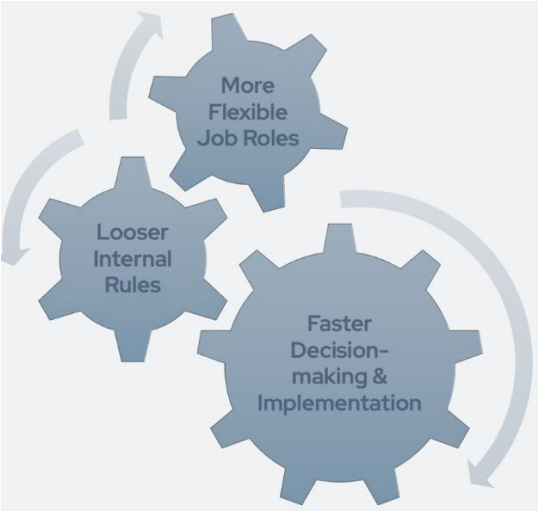



***Figure 1. Structural characteristics enabling faster AI adoption in SMEs (Werschitz, 2025)***

Taken together, these characteristics mean that SMEs are not merely passive recipients of AI innovation. Under the right conditions, they can act faster, learn earlier and adapt more fluidly—turning organisational agility into a genuine advantage in the age of generative AI (see Figure 1).

## Beyond Firm Size: Rethinking AI Adoption in SMEs

In policy, statistics and funding frameworks, SMEs are often treated as a single category. In Hungary, this label spans an extremely wide range—from sole proprietors and micro-enterprises to firms with up to more than one hundred employees and annual revenues in the range of 10 billion HUF. While this classification is useful for measurement and regulation, it is far less helpful when discussing AI adoption. In practice, firm size alone explains very little about whether, how and where AI can create value.

What matters far more is the nature of the business itself—particularly the industry context and the role that digital activities play in value creation. Different types of SMEs encounter very different AI opportunity spaces, even when they are similar in size.

Consider a manufacturing company producing physical goods. In such cases, currently available, off-the-shelf generative AI tools typically have limited relevance for the core production process itself. However, this does not mean AI is irrelevant. Generative AI can still create tangible value by supporting administrative work, documentation, reporting, planning, internal communication, or quality-related paperwork—effectively offloading both managers and shop-floor staff from routine cognitive tasks and freeing up attention for operational priorities.
By contrast, a brick-and-mortar retail business with online sales channels or active digital marketing operates in a hybrid model. Here, generative AI can support content creation, customer communication, campaign planning, product descriptions, or basic analytics—activities that are not the physical core of the business, but increasingly decisive for competitiveness.

The picture changes most dramatically in knowledge-intensive firms such as consulting, design, marketing, or professional services. In these cases, the core output of the company is digital knowledge—documents, analyses, presentations, concepts, or creative artefacts. Here, generative AI is not merely a support function but can directly augment core value creation, reshaping how work is done, how quickly it is delivered and how expertise is scaled.

Viewed through this lens, AI adoption is best understood along the dimension of digital intensity: how central digital processes and knowledge work are to the firm's business model. This perspective explains far more than firm size why some SMEs can derive immediate, visible benefits from generative AI, while others experience more indirect or incremental gains.

At the same time, industry and digital maturity are not the only factors at play. Human and organisational aspects—such as perceived risk, openness to experimentation, leadership attitude and readiness to change—also strongly influence adoption trajectories. These softer dimensions often determine whether the technical potential of AI translates into actual business impact.

From a practitioner's perspective, these differences (i.e. linked to human and organisational aspects) are not only visible in day-to-day work with SMEs but are increasingly being examined in a more structured way. At the time of writing, an online, questionnaire-based survey is underway to

explore how organisational readiness and human factors shape AI adoption decisions among Hungarian SMEs. The study examines organisational conditions along the Technology–Organisation–Environment (TOE) framework and captures individual perceptions using key dimensions of the Technology Acceptance Model (TAM). While this article does not draw on the survey's results, the research reflects a growing recognition that technology availability is rarely the primary constraint. Once the study is completed, its findings will be published and made publicly available to contribute to a more evidence-informed discussion on SME AI adoption.

## From data to judgement: how generative AI reshapes value creation in consulting

Among the different SME archetypes discussed earlier, consulting and other knowledge-intensive firms represent a particularly powerful use case for generative AI. Their core output is not a physical product but insight: recommendations, analyses, interpretations and structured judgement delivered in digital form. As a result, AI does not merely support peripheral activities but can intervene directly along the entire value-creation process.

From a practitioner's perspective, understanding this impact requires a conceptual lens that connects everyday (consulting) work with different levels of cognitive effort. Rather than focusing on specific tools or features, it is more useful to consider **how generative AI interacts with the progression from raw inputs to informed judgement** and where human expertise remains indispensable.

While the Data–Information–Knowledge–Wisdom (DIKW) model, which is illustrated in Figure 2, is often regarded as an oversimplified and, in some respects, outdated representation of knowledge creation (Frické, 2009), it continues to serve as a useful conceptual shorthand for practitioners. In this article, DIKW is employed explicitly as a heuristic rather than as a descriptive or normative theory of cognition. Its purpose here is to structure how generative AI supports different layers of consulting work, while making clear that judgement, responsibility and accountability remain human.

Viewed through this heuristic lens, generative AI can play a role at each level of the progression—from data to information, from information to

knowledge, and ultimately to judgement—but its value increases as it moves closer to higher-order cognitive tasks.

At the most basic level, AI supports work with **data**. Generative AI tools can summarise raw datasets, perform exploratory analysis, create visualisations and generate structured reports. Tasks that previously required significant analyst time—such as preparing first-pass insights or recurring status summaries—can now be accelerated, allowing consultants to spend less time on mechanical preparation and more time on interpretation.

Moving upward, AI helps transform data into **information** by synthesising inputs from multiple sources and placing them into context. It can compare findings, identify patterns, highlight inconsistencies and tailor outputs to specific client situations. In this role, AI acts as a powerful assistant for sense-making, supporting practical usability rather than abstract analysis.

At the level of **knowledge**, generative AI becomes a genuine thinking partner. It can engage in structured dialogue during ideation, challenge established assumptions, suggest alternative frameworks, or help refine emerging concepts. Used well, AI does not provide "answers" but stimulates better questions—helping consultants stress-test their thinking and explore solution spaces more systematically.

Crucially, however, the ultimate focus in consulting remains **judgement and wisdom**. Strategic decisions, value judgements, prioritisation and responsibility for outcomes cannot be delegated to AI. Instead, AI's role is to amplify human judgement by reducing cognitive load, expanding perspective and accelerating exploration. The consultant remains accountable for choices, ethics and strategic direction; AI serves as an enabler of clearer, better-informed decisions.


***Figure 2. From data to judgement: the role of generative AI in knowledge-intensive work (Werschitz, 2025)***

From a practitioner's perspective, this layered use of AI illustrates why consulting firms stand to benefit disproportionately from generative AI adoption. When deliberately aligned with the progression from data to judgement, AI

enhances—not replaces—the human expertise at the heart of consulting value creation.

### Introducing AI into the enterprise: beyond experimentation

As generative AI moves from individual experimentation to broader organisational use, the challenge for SMEs is no longer whether to engage with AI, but how to do so in a way that is productive, safe and sustainable. In practice, successful introduction rarely hinges on a single decision or tool; instead, it emerges from a small number of interconnected considerations that shape how AI is perceived, adopted and governed inside the organisation.

- **Tool selection:** As discussed earlier, many publicly available generative AI tools already exist; choosing among them sets implicit boundaries around accessibility, integration and control, often influencing adoption behaviour more strongly than formal strategy statements.
- **Use cases:** Value creation depends on identifying where AI meaningfully supports existing work, rather than forcing technology into activities where it adds little practical benefit.
- **Adoption and learning:** Organisational impact is shaped by how quickly and confidently employees learn to use AI and whether experimentation is encouraged or implicitly discouraged.
- **Information security and legal compliance:** Trust in AI use is closely tied to clarity around data handling, confidentiality and regulatory obligations, particularly when external platforms are involved.
- **Responsible and ethical use:** Beyond formal compliance, organisations must define how AI aligns with their values, professional standards and accountability toward clients, employees and society.

These interconnected considerations reflect recurring patterns of more advanced forms of generative AI use within organisations, as outlined in Figure 3.

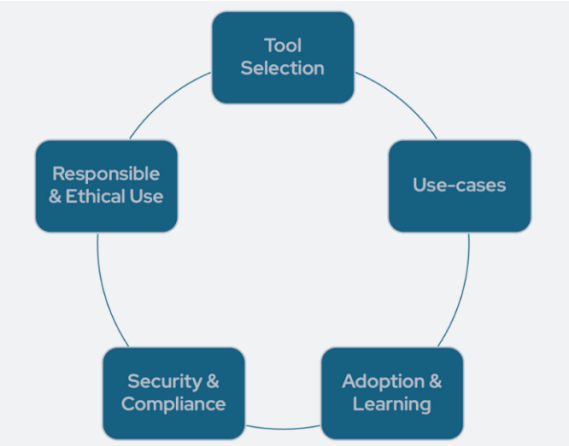



***Figure 3. Practical patterns of advanced generative AI use (Werschitz, 2025)***

In combination, these areas highlight that introducing AI is less a technical rollout and more an organisational learning process. For SMEs, approaching this transition with deliberate attention—rather than rigid control or unchecked enthusiasm—can make the difference between short-lived experimentation and lasting value creation.

## From Use Cases to Business Impact

Before concluding, it is worth briefly broadening the lens. Practical experience increasingly aligns with frameworks such as those proposed by **Gartner (2023)**, which emphasise that generative AI use cases differ not only by function, but by the *type of business impact they create* (Figure 4). AI can drive efficiency and productivity improvements, enhance decision-making, enable new ways of working, support innovation and in some cases reshape value propositions altogether. For SMEs, this distinction matters: the real opportunity is not simply to apply AI everywhere, but to understand *what kind of impact is realistically achievable* in their specific context.

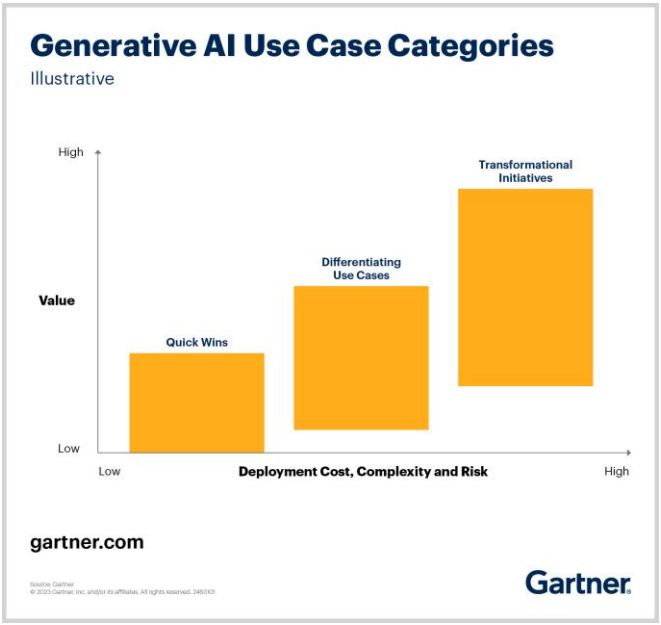



***Figure 4. Categories of generative AI use cases and their business impact (Gartner, 2023)***

## From accessibility to advantage: navigating AI with intent

This article has argued that the current wave of generative AI represents a structurally different opportunity for small and medium-sized enterprises. Not because AI is a universal solution, but because it fundamentally lowers barriers that once restricted advanced capabilities to large organisations. Accessibility, however, is only the starting point.

From a practitioner's perspective, the decisive factors lie elsewhere. SMEs are not a homogeneous group and firm size alone explains little about AI potential. Industry context, the role of digital work in value creation, organisational readiness and human attitudes toward experimentation all shape what AI can realistically deliver. In some cases, generative AI supports peripheral activities; in others—particularly in knowledge-intensive firms—it directly augments core value creation.

Equally important is recognising that AI's impact is multidimensional. Generative AI can improve efficiency, enhance insight, support learning

and enable new forms of collaboration. Yet value does not emerge automatically from tool adoption. It emerges when organisations make deliberate choices about where AI fits, how it is used and what remains a human responsibility—particularly in areas of judgement, accountability and ethics.

For SMEs, the strategic challenge is therefore not catching up with large enterprises, but navigating AI adoption with clarity and intent. Those that treat AI neither as a silver bullet nor as a distant threat—but as an enabler embedded in everyday work—are best positioned to turn perceived disadvantage into sustainable competitive edge.

## Acknowledgement

This research was conducted within the framework of the Infota project "Development of artificial intelligence-based methods to support the innovation capabilities of SMEs" supported by the Batthyány Lajos Foundation.

## References

CB Insights. (2024, May 24). *The generative AI market map*. CB Insights.

Frické, M. (2009). The knowledge pyramid: A critique of the DIKW hierarchy. *Journal of Information Science, 35*(2), 131–142. https://doi.org/10.1177/0165551508094050

Gartner. (2023). *Generative AI use case categories and business impact*. Gartner Research.

Google Cloud. (n.d.-a). *Vertex AI platform*. Google Cloud.

Google Cloud. (2025, June 27). *Generative AI on Vertex AI: Overview*. Google Cloud Documentation.

Microsoft. (2025a, November 7). *What is Microsoft 365 Copilot?* Microsoft Learn.

Microsoft. (2025b). *Microsoft 365 Copilot | AI productivity tools for work*. Microsoft.

OECD. (2024). *AI adoption by small and medium-sized enterprises: Challenges and opportunities*. Organisation for Economic Co-operation and Development. https://www.oecd.org https://doi.org/10.1787/357b63f7-en

Perplexity. (n.d.). *Perplexity Enterprise*. Perplexity AI.

VOSZPORT. (2024). *KKV-k és a mesterséges intelligencia: Felmérési eredmények*. https://voszport.com/kkv-mesterseges-intelligencia

Werschitz, O. (2025, October 7). *Hátrányból előny: Hogyan kovácsolhatnak a KKV-k versenyelőnyt a generatív MI-ből?* [Conference presentation]. "Us and AI" — Collaboration with Artificial Intelligence in the World of Work (Conference of the MTA GTB Knowledge Management Working Committee), University of Pannonia, Veszprém, Hungary.

https://doi.org/10.60030/ALMAMS.c202501z6rCG9

# Understanding Frontier AI's Implications for Business Model Innovation

Werschitz Ottó

founder
Cognivate AI
e-mail: otto.werschitz@cognivate-ai.com
https://orcid.org/0009-0008-0937-0768

## Abstract

Recent advances in large-scale, multimodal, reasoning-capable and agentic artificial intelligence—often described as *frontier AI*—are expanding the scope of what AI systems can autonomously generate, reason about and execute. While these developments create unprecedented possibilities for internal organisational and external market-facing innovation, current research shows that the adoption of AI technologies does not automatically lead to economic value creation. In addition, even if existing business model frameworks and AI-specific business model approaches provide useful foundations, they remain insufficient for capturing the characteristics of frontier AI systems from economic and business value creation points of view.

This introductory article seeks to clarify the conceptual gap by examining how frontier AI differs from classical AI in its technical properties, functional capabilities, organisational implications and market-facing innovation opportunities. It reviews the limitations of traditional business model approaches in this context and builds on theoretical insights from business model innovation, technological opportunity framing and organisational readiness research. The paper argues that frontier AI challenges core assumptions of existing business model tools and highlights the need for developing a new methodological approach to designing and evaluating business models enabled by frontier AI. The article

concludes by outlining a research agenda intended to guide such future methodological development and provides two **illustrative preview lenses**—a working typology of frontier AI configurations and value-chain business model archetypes—to motivate subsequent papers in the series.

***Keywords:*** *Frontier AI, Business Model Innovation, Foundation Models, AI Value Creation; General-Purpose Technology, Cognitive Task Automation*

## Introduction

The last few years have seen the emergence of a new family of artificial intelligence systems and applications based on large-scale foundation models. These systems differ from earlier generations of AI in their scale, generality and functional scope, enabling capabilities such as multimodal input processing, reasoning, content generation, planning and tool-mediated action across a wide range of domains (Bommasani et al., 2021; UK Government, 2025). In policy and governance discourse, these developments are commonly referred to as *frontier AI*, denoting highly capable general-purpose models operating at or beyond the current state of the art (UK Government, 2025).

Despite the rapid diffusion of frontier-AI-based tools—particularly generative AI applications—empirical evidence suggests that their adoption has not yet translated into widespread and measurable enterprise value creation. Recent survey-based research indicates that only a small fraction of organisations deploying generative AI at scale are able to demonstrate significant financial impact, while the majority of initiatives remain confined to pilots or use cases without clear profit-and-loss effects (Weill et al., 2025). This pattern echoes earlier findings in the broader AI literature, which show that while some firms successfully capture value from AI, many fail to do so despite substantial investment and technological sophistication (Climent, Haftor, & Staniewski, 2024; Kiss et al., 2015).

From an economic perspective, artificial intelligence exhibits several characteristics commonly associated with general-purpose technologies, including broad applicability across sectors, rapid performance improvements and strong dependence on complementary organisational and institutional innovations (Bresnahan & Trajtenberg, 1995). Prior

research on such technologies suggests that their economic impact does not arise automatically from technological advancement alone, but rather depends on significant changes in organisational structures, workflows and business models (Bresnahan & Trajtenberg, 1995; Teece, 2010).

In response to these challenges, a growing body of research and practice has sought to develop AI-specific business model frameworks and methodologies. These include economic interpretations of AI as a prediction technology (Agrawal, Gans, & Goldfarb, 2018), systematic reviews of AI-driven business model innovation (Jorzik et al., 2024), studies of AI-enabled product–service innovation (Naeem, Kohtamäki, & Parida, 2025) and practitioner-oriented tools such as AI canvases designed to bridge business and machine-learning perspectives (Kerzel, 2020). While these approaches provide valuable insights, they are largely grounded in assumptions derived from classical, task-specific AI systems.

At the same time, there is limited research—and virtually no validated, practice-oriented methodologies—that explicitly address the business model implications of frontier AI as a distinct class of systems. Moreover, the concept of frontier AI itself remains under-defined from an economic and organisational perspective, as existing definitions focus primarily on capability thresholds and risk categories rather than on differentiated system types relevant for value creation and capture (UK Government, 2025). This conceptual ambiguity complicates efforts to design business models and innovation strategies tailored to frontier AI.

To support this agenda without pre-empting later articles, the paper also introduces two non-exhaustive preview constructs: (1) illustrative business model archetypes along the AI value chain and (2) a working typology of business-model-relevant frontier AI configurations.

Against this background, this research is initiated to better understand how frontier AI differs from classical AI (i.e., task-specific/narrow ML systems), particularly with respect to value creation mechanisms and business model innovation options. The paper aims to clarify the conceptual foundations necessary for analyzing frontier-AI-enabled business models and to outline a research agenda that can support the development of more systematic and practically relevant methodologies in this emerging domain.

As an introductory contribution, the paper does not propose a complete business model innovation methodology for frontier AI. Instead, it establishes conceptual groundwork and motivating preview lenses for subsequent methodological development to operationalise and empirically examine business model frameworks tailored to frontier AI systems.

## Business Models as the Locus of Value Creation and Capture

Research on business models has consistently emphasised that economic value is not created by technology alone, even when the underlying technology is highly innovative. Instead, value creation and value capture are realised through the business model, which mediates between technological capabilities and economic outcomes (Baden-Fuller & Haefliger, 2013).

Baden-Fuller and Haefliger (2013) conceptualise the business model as the central mechanism through which technological innovation affects firm performance. In this view, technology provides the potential for value creation, but it is the business model that determines how this potential is translated into customer value and how that value is subsequently captured by the firm. The authors argue that technological innovation influences performance indirectly, with business model choice acting as a key moderator between technology and economic outcomes. Value creation and capture thus emerge from the interaction between technological capabilities and business model design choices, including customer identification, value delivery and monetisation mechanisms. As a result, even superior technologies may fail commercially under poorly designed business models, while more modest technologies can generate substantial returns when paired with effective business model configurations.

This perspective is reinforced by Teece (2010), who defines the essence of a business model as specifying how an enterprise delivers value to customers, induces them to pay for that value and converts those payments into profit (value capture). Teece emphasises that without a well-developed business model, innovators are unlikely to either deliver or capture value from technological advances. Although his analysis draws primarily on internet-based service firms, the underlying logic applies more broadly to digital and AI-enabled innovations, where intangible assets, complementarities and organisational design play a central role in value creation.

Alongside these theoretical approaches, a widely adopted practical framework for articulating business models is the Business Model Canvas (BMC) introduced by Osterwalder and Pigneur (2010). The BMC provides a structured representation of key business model elements, including value propositions, customer segments, key resources, activities and revenue streams, and is extensively used in both academic and practitioner contexts as a tool for business model design and communication. Its popularity reflects the broader recognition that business models function as integrative constructs linking technology, organisation and market interaction.

Taken together, these perspectives establish the business model as the primary locus of value creation and value capture at the firm level. Technology—regardless of its novelty or performance—remains an essential but insufficient input whose economic effects are realised only through appropriate business model choices. This framing is particularly relevant for digital and AI-based innovations, where changes in the cost, scope and structure of economic activities must be translated into organisational processes, workflows and market-facing value propositions through deliberate business model design.

Critique of the Business Model Canvas in the Context of AI

While the Business Model Canvas (BMC) is a widely used tool for articulating value creation and capture, it offers limited support for the design of AI-enabled business models. Unlike the preceding theoretical perspectives, which operate at a conceptual level, the BMC functions as an operational design instrument; its limitations therefore have direct implications for practice.

In particular, the BMC leaves implicit several issues that are central to AI-enabled value creation, including data availability and quality, model constraints, the translation of algorithmic outputs into operational decisions and the redistribution of decision rights and workflows when decision-making is partially delegated to algorithms. These omissions become especially striking in AI contexts as learning dynamics, feedback loops and capability accumulation play a strategic role.

As a result, while the BMC remains useful for high-level business model articulation, it provides insufficient guidance for capturing the distinctive economic and organisational implications of AI. This limitation has

motivated the development of AI-specific canvases and frameworks intended to bridge business and machine-learning perspectives (Kerzel, 2020; Climent, Haftor, & Staniewski, 2024). In this sense, the critique of the BMC closes the general discussion of business models and motivates a more focused examination of AI-centric business model research in the following section.

## AI-Centric Business Model Research: Insights and Limitations

A growing body of literature has examined how artificial intelligence enables new forms of value creation and business model innovation. While these contributions differ in emphasis and level of analysis, they collectively provide a coherent understanding of AI-enabled business models under assumptions associated with classical, task-specific AI systems.

One of the most influential economic perspectives on AI-enabled business models is offered by Agrawal, Gans and Goldfarb. In their work on *Prediction machines,* AI is conceptualised primarily as a technology that dramatically reduces the cost of prediction (Agrawal, Gans, & Goldfarb, 2018). From this perspective, value creation arises not from prediction accuracy alone, but from the redesign of decisions, workflows and organisational processes in which prediction represents a bottleneck. The authors emphasise that successful AI-enabled business models must secure access to data across the lifecycle—both for model training and for continuous improvement (Agrawal et al., 2018). This framework provides a clear and internally consistent economic logic for AI-enabled value creation in settings where AI systems are designed to optimise performance on well-defined tasks within bounded workflows.

Research on AI-enabled product–service innovation further extends this understanding by examining how AI contributes to value creation across multiple organisational levels. Naeem, Kohtamäki and Parida synthesise prior studies to show that AI enables not only internal efficiency gains through enhanced decision-making and process optimisation, but also innovation-oriented value creation through smart products, smart services and integrated product–service systems (Naeem, Kohtamäki, & Parida, 2025). Beyond incremental improvements, this literature positions AI as a driver of business model innovation, particularly in servitisation and digital servitisation contexts where firms redesign value propositions,

revenue mechanisms and ecosystem roles (Naeem et al., 2025). Importantly, these contributions emphasise that AI-enabled value creation is contingent on complementary organisational and business model transformations, rather than on technology adoption alone. At the same time, the analyzed cases predominantly assume task-bounded AI systems embedded within relatively stable organisational contexts.

Systematic reviews of AI-driven business model innovation further highlight both the promise and the fragmentation of the field. Jorzik and colleagues identify multiple roles that AI can play in business models—ranging from supportive and enabling functions to more central and disruptive roles—and distinguish different types of business model innovation depending on the degree of AI centrality (Jorzik et al., 2024). While this framework brings analytical clarity to the heterogeneity of AI-related business model change, it largely treats business model innovation as a consequence of AI implementation rather than as a primary design challenge. As a result, the review consolidates insights about AI-enabled business model variation without fully addressing how fundamentally new AI capabilities might reshape underlying business model design logics.

Complementary insights are provided by research on AI-enabled business models for competitive advantage. Climent, Haftor and Staniewski emphasise that value creation in AI-enabled business models depends on activating established business model themes such as efficiency, novelty, lock-in and complementarity, while also accounting for data ownership and data network effects (Climent, Haftor, & Staniewski, 2024). Their situated AI theory explains competitive advantage as emerging from organisational activities that ground, bound and recast AI capabilities within specific contexts (Climent et al., 2024). This perspective foregrounds the importance of organisational alignment and environmental fit in AI-enabled value capture. However, it similarly assumes AI systems that are tailored, task-specific and closely embedded within organisational boundaries, leaving open questions about how these mechanisms operate when AI capabilities become more general-purpose and reusable.

Across these research streams, practitioner-oriented frameworks such as AI canvases have been developed to support organisational AI adoption. These tools aim to operationalise AI-related business model considerations by bridging business and machine-learning perspectives

(Kerzel, 2020). While valuable as planning and communication aids, they largely reflect the same underlying assumptions as the academic literature, treating AI as a bounded capability integrated into a specific business model configuration rather than as a shared or cross-context capability platform.

Positioning AI Business Models within the AI Value Chain

While the AI-centric business model literature has significantly advanced our understanding of how AI enables new forms of value creation, it predominantly adopts a firm-level perspective, focusing on how organisations integrate AI capabilities into products, services or internal processes. Across these approaches, AI is typically treated as a bounded technological resource embedded within a specific business model configuration (Agrawal et al., 2018; Naeem et al., 2025; Jorzik et al., 2024; Climent et al., 2024).

A complementary perspective is offered by the concept of the AI value chain, which conceptualises AI-enabled value creation as a multi-layered system extending beyond individual firms. In this view, value creation spans upstream enabling layers (such as data resources, computational infrastructure and governance mechanisms), core AI capability development (including model training, adaptation and operationalisation) and downstream applications and services that interface with end users and markets (Heeks & Spiesberger, 2024). This layered structure emphasises that AI-based offerings are embedded in broader ecosystems of technological, organisational and institutional interdependencies.

Notably, this value-chain perspective is largely absent or only implicitly addressed in existing AI-centric business model frameworks. As a result, current approaches offer limited guidance on how business model design choices relate to different positions within the AI value creation system, or how shifts in underlying AI technologies may reconfigure these positions. This limitation becomes particularly striking when considering advanced forms of AI, where value creation, control and capture increasingly span multiple layers of the AI value chain. For this reason, the AI value-chain perspective provides an important conceptual backdrop for the subsequent discussion of frontier AI.

To anticipate the later frontier-AI-focused parts of this article series without pre-empting their full development, we introduce below an

**illustrative, frontier-biased positioning lens**: a small set of **provisional business model archetypes** mapped to the AI value chain. The purpose is not to provide an exhaustive classification, but to make explicit how frontier AI can shift value creation and capture across layers (e.g., compute, data/governance, model platforms, orchestration, and downstream applications), thereby motivating the more detailed frontier AI conceptualisation and methodological work developed in subsequent articles.

Business model archetypes along the AI value chain

This perspective becomes particularly important for frontier AI because value creation, control, and value capture increasingly span multiple layers rather than remaining contained within a single firm or a single workflow-level application.

To make this actionable for business model analysis, Table 1 proposes seven archetypal frontier-AI business models positioned along the AI value chain. Each archetype is defined by (i) its value proposition and locus of value capture, (ii) typical revenue architecture, (iii) dominant sources of defensibility (moats), and (iv) primary governance and risk constraints. This archetypal view complements firm-level AI adoption frameworks by explicitly modelling ecosystem positioning and cross-layer dependencies as first-class business model variables.

These archetypes are presented as a preview lens, whereas the formal conceptual definition of frontier AI is developed in the subsequent section.

How these archetypes connect to frontier AI's "regime shift".

The archetypes make explicit why classical, firm-contained AI logics do not transfer straightforwardly to frontier AI. When frontier AI behaves as a reusable cognitive capability platform—frequently accessed via APIs and embedded into multi-actor ecosystems—value capture often depends less on the model itself and more on positioning within the value chain and controlling complementary assets (data, distribution, orchestration layers, governance, and workflow integration; Lupeikiene et al., 2014).

This view also clarifies why governance intensity and failure propagation become business-model-central: reuse and homogenisation can propagate defects across downstream systems, increasing the strategic importance of evaluation, guardrails, auditability, and accountability mechanisms as part

of the business model rather than merely as technical afterthoughts. (Caplinskas et al., 2012; Horváth et al., 2016).

**Table 1. Frontier AI business model archetypes along the AI value chain**

| Archetype | Value-chain position | Value proposition | Typical revenue architecture | Defensibility (moats) | Key constraints/risks |
|---|---|---|---|---|---|
| A1 Compute & infrastructure enablers | Upstream (hardware/cloud) | Scalable compute, acceleration, deployment primitives | Usage-based, reserved capacity, enterprise contracts | Scale, performance, distribution, ecosystem integration | Capital intensity, cyclical demand, geopolitical/regulatory exposure |
| A2 Data & governance providers | Upstream (data/governance) | High-quality data access, data governance, provenance, synthetic data, compliance tooling | Subscription, licensing, per-query, compliance services | Data rights, network effects, trust, domain coverage | IP/privacy constraints, data drift, "garbage-in" reputational risk |
| A3 Foundation model providers | Core capability layer | General-purpose cognitive capability as a platform (APIs, model families) | Token/usage-based, tiered access, enterprise licensing | Scale effects, developer ecosystem, model quality + tooling | Commoditisation, safety/regulatory obligations, dependency on compute/data supply chain |
| A4 Adaptation & operationalisation platforms | Mid-layer (MLOps/RAG/eval) | Make frontier AI reliable in production: adaptation, evaluation, monitoring, guardrails | SaaS subscription + usage, enterprise licensing | Workflow embedding, integration depth, evaluation assets | Fast-moving standards, interoperability pressure, liability allocation |
| A5 Orchestration / agent middleware | Mid-layer (orchestration) | Coordinate multi-step work: tool integration, planning, routing, cost/quality optimisation | Usage-based + platform fees | Switching costs via workflow dependence; partner ecosystem | Auditability, cascading failures, accountability gaps in agentic execution |
| A6 Vertical AI applications / productised solutions | Downstream (apps) | Domain-specific outcomes (industry workflows, regulated processes) | Seat-based + usage; outcome-based where feasible | Domain expertise, compliance assets, integration into industry processes | Regulatory burden, costly domain adaptation, slower sales cycles |
| A7 Services & transformation partners | Downstream (services) | Strategy, redesign, integration, change management, governance setup | Time-and-materials, retainers, performance-based hybrids | Trust, delivery capability, industry relationships | Difficulty in scaling; margin pressure; reliance on partner platforms |

Taken together, existing AI-centric business model research exhibits strong internal coherence with respect to classical AI systems. Across economic theory, innovation studies and practitioner frameworks, AI-

enabled value creation is consistently framed around task-specific applications, bounded data sources, stable workflows, organisationally contained deployments and predominantly firm-level positions within the AI value chain (Agrawal et al., 2018; Naeem et al., 2025; Jorzik et al., 2024; Climent et al., 2024; Heeks & Spiesberger, 2024). These shared assumptions underpin a rich and increasingly mature body of knowledge on AI-enabled business models. At the same time, they delineate the implicit boundaries of current approaches, raising the question of whether and how these frameworks remain applicable as AI systems evolve beyond task-specific prediction toward more general-purpose, reusable and cognitively capable forms. These unresolved issues provide the point of departure for the research gap articulated later in this document.

## From Classical AI to Frontier AI: Conceptual Foundations and Research Implications

To assess whether existing business model frameworks remain applicable in the context of recent advances in artificial intelligence, it is first necessary to clarify how frontier AI differs conceptually from earlier, task-specific forms of AI.

### Artificial intelligence (AI)

Artificial intelligence (AI) can be defined broadly as the discipline of building intelligent agents that perceive their environment and act upon it. Within this discipline, machine learning (including deep learning) represents a major approach for constructing such agents by enabling them to learn from data or experience rather than relying on hand-coded rules (Mitchell, 1997; Russell & Norvig, 2020).

In this framing, an AI system is characterised by a perception–decision–action loop, where percepts from the environment are processed internally to determine actions. Different AI approaches can thus be understood as alternative methods for implementing the mapping from percept histories to actions (Russell & Norvig, 2020).

### Machine learning (ML)

A Machine Learning system is a computer program to learn from experience E with respect to some class of tasks T and performance

measure P, if its performance at tasks in T, as measured by P, improves with experience E (Mitchell, 1997).

This definition is useful in an article because it cleanly separates (1) the task, (2) how success is measured and (3) what data/experience drives improvement (Mitchell, 1997).

"Classical AI" (narrow, task-specific ML)

The term "classical AI" in this article means classical machine learning (including deep learning) for task-specific/narrow systems, i.e. ML systems trained to improve performance on a pre-specified task T under a pre-specified metric P, using a defined experience/data source E (Mitchell, 1997).

Frontier AI (general-purpose models at the capability frontier)

Frontier AI (FAI) is a policy-centric label for highly capable, general-purpose models at the state of the art; in practice it largely refers to advanced foundation-model-based systems already used in real applications (Bommasani et al., 2021; UK Government, 2025).

Policy-led definition

The UK government defines "frontier AI" as highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities of today's most advanced models; annex materials note that, as of late 2023, this category primarily encompasses foundation models (UK Government, 2025). OpenAI similarly uses the term for highly capable foundation models whose dangerous capabilities may be hard to anticipate or contain (OpenAI, 2023).

Frontier tech families

Three research families illustrate the capability frontier: (1) foundation models trained on broad data at scale and adapted across tasks (Bommasani et al., 2021); (2) agentic systems that interleave reasoning and tool-mediated action, exemplified by ReAct (Yao et al., 2022); and (3) world-model approaches that learn predictive models of environments for planning/control at scale, as in Dreamer (Hafner et al., 2023).

Versatile real applications

These frontier capabilities already underpin a non-exhaustive set of deployed applications in science and engineering, for example AlphaFold for protein structure prediction (Jumper et al., 2021), GraphCast for medium-range weather forecasting (Lam et al., 2023) and the Molecular Transformer for reaction prediction and uncertainty estimation (Schwaller et al., 2019). LLM-based simulated societies further suggest new "synthetic experimentation" modes: Generative Agents operationalise memory, reflection and planning around an LLM and show emergent social behaviors in an interactive sandbox (Park et al., 2023).

A potentially confusing related term is Transformative AI (TAI), which is conceptually distinct from frontier AI: while FAI is primarily a technology/capability label used in policy and governance (UK Government, 2025), TAI is an impact-focused term in economics referring to AI that could drive an economy-wide transformation by making intelligence abundant and reproducible rather than scarce (Agrawal et al., 2023). Accordingly, the two can overlap, but they are not synonymous.

Classical AI versus FAI

"Classical AI" is task-specific machine learning in the sense of Mitchell's canonical definition: a program learns from experience E with respect to tasks T and performance measure P if its measured performance improves with E (Mitchell, 1997). In contrast, "frontier AI" is regarded as highly capable general-purpose AI models that can perform a wide variety of tasks and "match or exceed" the capabilities of today's most advanced models (UK Government, 2025).

Conceptual difference (what is being optimised)

Classical ML systems are engineered to optimise performance P on a predefined task T using a defined data/experience source E, which drives narrow problem boundaries, stable metrics and well-scoped deployment contexts (Mitchell, 1997). FAI systems are better described as capability platforms built around broad pretraining and downstream adaptation, rather than trained end-to-end for a single task and metric (Bommasani et al., 2021).

From prediction machines to cognitive capability systems

Prior economic interpretations of AI—most prominently articulated by Agrawal, Gans and Goldfarb—frame classical AI primarily as a technology that dramatically reduces the cost of prediction; value creation therefore shifts toward redesigning decisions and workflows where prediction is the bottleneck (Agrawal, Gans, & Goldfarb, 2018).

Frontier AI extends this economic logic beyond prediction toward the partial automation of broader cognitive tasks, including content generation, reasoning, simulation and planning and tool-mediated action (Bommasani et al., 2021; UK Government, 2025; Yao et al., 2022; Hafner et al., 2020).

This shift does not remove the need for human judgment: humans remain responsible for setting goals, constraints and accountability (Agrawal et al., 2018). However, frontier AI can generate alternative courses of action, reason over trade-offs and execute bounded actions via tools under explicit objectives, which blurs the operational boundary between prediction and judgment (Yao et al., 2022; Kiss, 2016).

*How this partial automation of cognitive work reshapes the economics of AI adoption—particularly in terms of value creation, organisational design and business model structure—remains an open question and a central motivation for further analysis.*

Technology differences that matter for business models

Foundation-model literature explicitly frames these models as "critically central yet incomplete," meaning they require downstream adaptation and are used across many applications, which shifts value creation from "one model → one app" to "one model → many apps and complementors" (Bommasani et al., 2021). Policy treatments of frontier AI emphasise broad capability and rapid evolution of the "frontier," which implies shifting performance baselines and governance requirements over time rather than a static optimisation target (UK Government, 2025).

A value-chain perspective highlights that foundation models sit within a layered AI ecosystem (hardware → cloud → training data → foundation models → applications), which changes cost structure, bargaining power, and routes to market compared with many classical ML deployments (Gambacorta & Shreeti, 2025).

Classical AI versus Frontier AI: practical summary

- **Unit of design:** Classical AI optimises performance P for a predefined task T from experience E (Mitchell, 1997). FAI is a general-purpose capability core that is adapted across many tasks and contexts (Bommasani et al., 2021; UK Government, 2025).
- **Training logic:** Classical AI typically uses bounded datasets and metrics tied to one use case (Mitchell, 1997). FAI relies on large-scale pretraining + downstream adaptation; scale can yield emergent capabilities and widespread reuse (Bommasani et al., 2021; UK Government, 2025).
- **Scope of deployment:** Classical AI is often mapped to one workflow and one decision boundary. FAI is frequently reused across products, teams and domains, often through APIs and tooling layers (Bommasani et al., 2021; Gambacorta & Shreeti, 2025).
- **Failure propagation:** Classical AI errors are often localised if the model is not reused widely. With FAI, reuse/homogenisation can propagate shared defects across multiple downstream systems, making governance and risk management more central (UK Government, 2025).
- **New capability patterns:** Classical AI is primarily prediction/classification within a defined scope (Mitchell, 1997). FAI is associated with agentic, generative and reasoning-oriented patterns, including multi-step reasoning and tool-mediated action (Bommasani et al., 2021; Yao et al., 2022; Park et al., 2023; UK Government, 2025).

Crucially, the transition from classical, task-specific AI to frontier AI does not represent a quantitative extension of existing AI capabilities, but a qualitative regime shift with respect to how AI functions as an economic input. Classical AI systems are designed and evaluated at the level of individual tasks, workflows orproducts, enabling business model design to focus on localised efficiency gains and decision improvements. In contrast, frontier AI systems operate as general-purpose capability platforms that can be reused, adapted and recombined across multiple tasks, organisational functions and market contexts. This shift changes the unit of analysis for business model design—from isolated AI-enabled applications to shared cognitive infrastructures embedded in multi-layer ecosystems. As a result, business model logics developed for narrow AI

cannot be straightforwardly extended to frontier AI without reconsidering their underlying assumptions.

Business-model-relevant taxonomy of frontier AI configurations

Current definitions of frontier AI are primarily policy- and governance-oriented, focusing on capability thresholds and risk categories. While appropriate for oversight, this framing remains under-specified for business model analysis because it does not differentiate economically and organisationally distinct system configurations.

To preview the subsequent taxonomy-focused paper in this series, this section outlines a **working typology** of frontier AI configurations for business model analysis. The typology is **illustrative rather than exhaustive** and is intended to be refined through further conceptual development and case-based validation.

Specifically, the working typology distinguishes frontier AI configurations along four business-model-relevant dimensions that repeatedly emerge in the transition from task-level prediction systems to general-purpose cognitive capability platforms: (1) locus of value creation (capability core vs. application vs. orchestration), (2) degree of reuse across contexts (single workflow vs. multi-domain capability platform), (3) degree of agency and tool-mediated action (assistive vs. bounded autonomous action), and (4) dependence on complementary assets and governance (data access, integration, evaluation, safety, and accountability mechanisms).

Table 2 summarises five **illustrative** frontier AI configuration types. The types are not intended as an exhaustive technical classification; rather, they represent business-model-distinct configurations with **distinct** implications for value creation, value capture, organisational design, and ecosystem positioning.

Taken together, the typology supports the paper's central argument that the unit of design shifts from isolated AI applications to shared cognitive infrastructures embedded in multi-layer ecosystems. Consequently, business model design must explicitly account for reuse, complementarity, orchestration, governance, and ecosystem value capture.

In the remainder of this paper, these configuration types are used only as a **motivating lens** to clarify why business model logics developed for task-specific AI may not transfer directly to frontier AI.

**Table 2: Frontier AI configurations: business-model-relevant system types**

| Label | Core configuration | Locus of value | BM-relevant differentiator | Primary BM risks/constraints |
|---|---|---|---|---|
| FM-as-a-service capability core | General-purpose foundation model exposed via API (multi-tenant) | Capability core layer | Strong reuse across contexts; scale effects and ecosystem complementors | Commoditisation, margin pressure, regulatory risk; dependency on compute and data supply chain |
| Enterprise copilots / workbenches | FAI embedded in human workflows (chat + authoring + analytics) with enterprise controls/ connectors | Productivity and knowledge work augmentation | Value depends on organisational adoption, workflow redesign, and governance of human–AI boundaries | Low realised value despite adoption; security/compliance failures; change-management friction |
| Grounded knowledge systems (RAG + verification) | FAI + retrieval over curated corpora + citation/verification + evaluation loops | Knowledge-intensive task quality and speed | Differentiation through domain knowledge assets, content governance, evaluation, and reliability | Liability for errors; data rights; quality drift; "false confidence" without robust verification |
| Agentic tool-using systems (bounded autonomy) | FAI + planning + tool access (APIs, workflows) + constraints | Orchestration of tasks and execution chains | Shift from "content generation" to "tool-mediated action"; governance becomes central due to failure propagation | Misaligned actions, cascading errors, auditability and accountability gaps, *including failure propagation risks* |
| Domain-adapted frontier AI (vertical capability) | FAI adapted via fine-tuning / domain constraints / specialised interfaces | Vertical value propositions and industry outcomes | Differentiation via domain specificity, compliance, and integration into industry processes | High cost of domain adaptation; regulatory constraints; lock-in vs interoperability tension |

## Business Model Implications of Frontier AI

Building on the configuration types summarised in Table 2, frontier AI cannot be adequately understood using business model logics developed for task-specific machine learning systems. Classical AI applications are typically embedded in well-defined workflows, optimise a single performance metric, and create value by improving prediction or classification within existing organisational boundaries (Mitchell, 1997). In contrast, frontier AI increasingly functions as a reusable cognitive capability core that can be adapted, reused, and recombined across multiple tasks, functions, and domains (Bommasani et al., 2021; UK Government, 2025).

This shift has important consequences for how value is created and captured. When AI functions as a reusable cognitive infrastructure rather

than a task-level component, business model design must account for adaptation across contexts, orchestration of complementary assets, governance of shared capabilities and value capture across a multi-layer ecosystem (Bommasani et al., 2021; Gambacorta & Shreeti, 2025). As a result, familiar product-centric or workflow-specific business model patterns become insufficient to fully describe how frontier AI generates and distributes economic value (UK Government, 2025).

Despite the growing deployment of frontier AI systems, systematic approaches for analyzing and designing business models around such capabilities remain underdeveloped. Existing business model frameworks were largely shaped in contexts where AI systems were narrowly scoped, application-specific and organisationally contained (Mitchell, 1997). This gap motivates the need for business model research that moves beyond task-level AI productisation toward frameworks capable of capturing reuse, complementarity, governance and ecosystem dynamics associated with frontier AI (Bommasani et al., 2021; Gambacorta & Shreeti, 2025). *The present work positions itself as a first conceptual step toward such an AI-centric business model perspective.*

## Research Gap: From Cheap Prediction to Cheap Cognitive Task Execution

Building on the preceding analysis of business model theory, AI-centric business model research, the AI value chain and the conceptual distinction between classical and frontier AI, this section synthesises the resulting limitations and articulates the central research gap addressed by this study.

Business model research has established that value creation and value capture at the firm level are mediated through the business model rather than through technology alone (Baden-Fuller & Haefliger, 2013; Teece, 2010). While existing AI-centric business model frameworks provide important insights into how artificial intelligence can enable efficiency gains, innovation and competitive advantage, they are largely grounded in assumptions associated with classical, task-specific AI systems (Agrawal, Gans, & Goldfarb, 2018; Naeem, Kohtamäki, & Parida, 2025; Jorzik et al., 2024; Climent, Haftor, & Staniewski, 2024).

As shown in the preceding analysis, the dominant strands of AI-centric business model research conceptualise AI primarily as a prediction

technology that reduces the cost of forecasting within well-defined tasks and workflows (Agrawal et al., 2018). Within this paradigm, business model innovation focuses on redesigning decision processes, securing data across the AI lifecycle and embedding AI into stable organisational and market contexts (Agrawal et al., 2018; Naeem et al., 2025) These approaches are internally consistent and have proven useful for understanding AI-enabled value creation under conditions where AI systems are narrowly scoped, data-bounded and organisationally contained (Jorzik et al., 2024; Climent et al., 2024).

However, recent advances in artificial intelligence—referred to in this study as frontier AI—challenge these underlying assumptions. Frontier AI systems increasingly function as general-purpose, reusable capabilities that extend beyond prediction to the automated execution of broader cognitive tasks, such as reasoning, content generation, planning and bounded autonomous action (Bommasani et al., 2021; UK Government, 2025). This shift alters the economic role of AI from a technology of cheap and automated prediction to a technology of cheap and automated cognitive task execution (Agrawal et al., 2018; Bommasani et al., 2021).

This transition has important implications for business model research. By enabling the partial automation of cognitive work that was previously tightly coupled to human judgment and organisational roles, frontier AI affects not only decision-making costs but also the allocation of tasks, the design of workflows and the boundaries between human and machine agency (Agrawal et al., 2018; Climent et al., 2024). At the same time, existing business model frameworks provide limited guidance on how such changes reshape value creation and capture, particularly when AI capabilities are shared across multiple applications, embedded in broader ecosystems and positioned at different layers of the AI value chain (Heeks & Spiesberger, 2024).

Moreover, frontier AI remains under-specified from a business model perspective. Current definitions are primarily formulated at the level of capability thresholds and risk categories, which are appropriate for governance and policy purposes but insufficient for systematic analysis of economically and organisationally differentiated system types (UK Government, 2025). As a result, there is no coherent conceptual basis for linking frontier AI capabilities to specific business model design choices

or for comparing business model implications across different frontier AI configurations (Jorzik et al., 2024; Climent et al., 2024).

Taken together, these limitations point to a clear research gap. While existing AI-centric business model research offers a robust foundation for understanding value creation in prediction-centric AI systems (Agrawal et al., 2018; Naeem et al., 2025), it does not adequately account for the emerging economic and organisational implications of frontier AI as a technology of cognitive task execution.

Addressing this gap requires clearer conceptualisation of frontier AI, **business-model-relevant distinctions among frontier AI system types** and methodological approaches capable of capturing how such systems reshape value creation, value capture and organisational design; accordingly, this paper introduces an **illustrative working typology** and **value-chain archetype lens** as a first step toward that programme.

**Suggested Further Work in Subsequent Research**

Building on the conceptual foundations established in this paper, future research should focus on:

- Developing a business-model-relevant conceptualisation and taxonomy of frontier AI systems that distinguishes economically and organisationally meaningful system types, building on the preliminary working typology introduced in this paper and iterating it into a more complete, theoretically grounded and empirically informed framework
- Analysing how different frontier AI configurations map to distinct patterns of value creation, value capture and ecosystem positioning, and systematically assessing/validating the initial value-chain archetype lens proposed here as a useful structuring device
- Examining how business model design logics evolve when AI functions as a reusable, general-purpose capability rather than a task-level component
- Advancing methodological approaches for designing, evaluating and iterating business models enabled by frontier AI, informed by innovation theory and empirical evidence

These research directions are intended to support the development of a structured and cumulative body of knowledge on frontier-AI-enabled business model innovation.

## Contribution of This Paper

This paper contributes to the literature by demonstrating that existing AI-centric business model theories are internally consistent but regime-bound, resting on assumptions associated with classical, task-specific AI systems. It shows that frontier AI challenges these assumptions by introducing general-purpose, reusable and ecosystem-embedded capabilities that cannot be adequately captured using existing business model tools. By identifying the conceptual mismatch between frontier AI characteristics and current business model frameworks, the paper establishes the need for clearer conceptualisation and new methodological approaches to business model innovation in the context of frontier AI.

In addition, it provides two illustrative, non-exhaustive structuring devices—a working typology of frontier AI configurations and value-chain archetypes—to make the business-model implications of frontier AI analytically tractable and to scaffold the subsequent papers in this research programme.

## Conclusion

This paper examined why recent advances in frontier AI challenge existing approaches to business model innovation. It demonstrated that prevailing AI-centric business model frameworks are internally consistent but grounded in assumptions associated with classical, task-specific AI systems. By contrasting classical AI as a technology that enables cheap and automated prediction with frontier AI as increasingly enabling cheap and automated cognitive task execution, the paper highlighted a qualitative shift in AI's economic and organisational role.

The analysis further showed that frontier AI remains under-specified from a business model perspective, as current definitions focus primarily on capability thresholds and risk categories rather than on economically and organisationally differentiated system types. This under-specification limits the applicability of existing business model tools and constrains both theoretical progress and practical guidance.

In response, the paper argued for the need to develop clearer conceptualisations, taxonomies and methodological approaches tailored to frontier-AI-enabled business model innovation and introduced two

illustrative preview lenses—a working typology of frontier AI configurations and value-chain business model archetypes—to motivate and structure subsequent work. The research agenda outlined provides a foundation for future work aimed at systematically translating frontier AI capabilities into sustainable and economically viable business models.

**Acknowledgements**

The author thanks **Dr. Ferenc Kiss**, **R&D Director**, **Foundation for Information Society (INFOTA)**, for his constructive review and suggestions that helped improve the structure and positioning of this article and the planned subsequent papers in the series.

**References**

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence.* Harvard Business Review Press.

Agrawal, A., Gans, J., & Goldfarb, A. (2023). *The Turing transformation: Artificial intelligence, intelligence augmentation and skill premiums* (NBER Working Paper No. 31767). National Bureau of Economic Research. https://doi.org/10.3386/w31767

Baden-Fuller, C., & Haefliger, S. (2013). *Business models and technological innovation.* Long Range Planning, 46(6), 419–426. https://doi.org/10.1016/j.lrp.2013.08.023

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). *On the opportunities and risks of foundation models.* arXiv. https://doi.org/10.48550/arXiv.2108.07258

Bresnahan, T. F., & Trajtenberg, M. (1995). *General purpose technologies "engines of growth"?* Journal of Econometrics, 65(1), 83–108. https://doi.org/10.1016/0304-4076(94)01598-T

Caplinskas, A., Dzemyda, G., Kiss, F., & Lupeikiene, A. (2012). Processing of undesirable business events in advanced production planning systems. Informatica, 23(4), 563–579. https://doi.org/10.15388/Informatica.2012.375

Climent, R. C., Haftor, D. M., & Staniewski, M. W. (2024). *AI-enabled business models for competitive advantage.* Journal of Innovation & Knowledge, 9(3), 100532. https://doi.org/10.1016/j.jik.2024.100532

Gambacorta, L., & Shreeti, V. (2025). *The AI supply chain* (BIS Papers No. 154). Bank for International Settlements. https://www.bis.org/publ/bppdf/bispap154.htm

Hafner, D., Lillicrap, T., Norouzi, M., & Ba, J. (2020). *Mastering Atari with discrete world models.* arXiv. https://doi.org/10.48550/arXiv.2010.02193

Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2023). *Mastering diverse domains through world models.* arXiv. https://doi.org/10.48550/arXiv.2301.04104

Heeks, R., & Spiesberger, P. (2024). *Constructing an AI value chain and ecosystem model* (GDI Digital Development Working Paper No. 109). Centre for Digital Development, University of Manchester. https://doi.org/10.2139/ssrn.4960510

Horváth, A., Erdősi, P. M., & Kiss, F. (2016). The Common Vulnerability Scoring System (CVSS) generations—usefulness and deficiencies. In IT és hálózati sérülékenységek társadalmi-gazdasági hatásai (pp. 137–153). Információs Társadalomért Alapítvány. https://infota.org/kiadvanyok/alma-mater/it-es-halozati-serulekenysegek-tarsadalmi-gazdasagi-hatasai/

Jorzik, P., Klein, S. P., Kanbach, D. K., & Kraus, S. (2024). *AI-driven business model innovation: A systematic review and research agenda.* Journal of Business Research, 182, 114764. https://doi.org/10.1016/j.jbusres.2024.114764

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). *Highly accurate protein structure prediction with AlphaFold.* Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kiss, F., Horváth, A., Török, M., & Szanyi, I. (2015). Modern ICT technologies—situation and trends. In Tourism and ICT Aspects of Balkan Wellbeing—a Balkán jóllét turisztikai és IKT vonatkozásai (pp. 155–186). Foundation for Information Society (INFOTA). https://infota.org/en/publications/alma-mater-study-series/tourism-

and-ict-aspects-of-balkan-wellbeing-a-balkan-jollet-turisztikai-es-ikt-vonatkozasai-en/

Kiss, F. (2016). The education of information and knowledge management of cultural heritage. In L. Bassa & F. Kiss (Eds.), Proceedings of TCL2016 conference: Tourism and cultural landscapes: Towards a sustainable approach (pp. 316–325). Foundation for Information Society (INFOTA). https://infota.org/wp-content/uploads/2017/01/Proceedings_TCL2016.pdf

Kerzel, U. (2020). *Enterprise AI Canvas—Integrating artificial intelligence into business.* arXiv. https://doi.org/10.1080/08839514.2020.1826146

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., & Battaglia, P. (2023). *Learning skillful medium-range global weather forecasting.* Science. https://doi.org/10.1126/science.adi2336

Lupeikiene, A., Dzemyda, G., Kiss, F., & Caplinskas, A. (2014). Advanced planning and scheduling systems: Modeling and implementation challenges. Informatica, 25(4), 581–616. https://doi.org/10.15388/Informatica.2014.31

Mitchell, T. M. (1997). *Machine learning.* McGraw-Hill.

Naeem, R., Kohtamäki, M., & Parida, V. (2025). *Artificial intelligence enabled product–service innovation: Past achievements and future directions.* Review of Managerial Science, 19(7), 2149–2192. https://doi.org/10.1007/s11846-024-00757-x

OpenAI. (2023, July 6). *Frontier AI regulation: Managing emerging risks to public safety.* https://openai.com/index/frontier-ai-regulation/

Osterwalder, A., & Pigneur, Y. (2010). *Business model generation: A handbook for visionaries, game changers and challengers.* Wiley.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). *Generative agents: Interactive simulacra of human behavior.* In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23). Association for Computing Machinery. https://doi.org/10.1145/3586183.3606763

Russell, S. J., & Norvig, P. (2020). *Artificial intelligence: A modern approach (4th ed.)*. Pearson.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., & Lee, A. A. (2019). *Molecular transformer: A model for uncertainty-*

*calibrated chemical reaction prediction*. ACS Central Science, 5(9), 1572–1583. https://doi.org/10.1021/acscentsci.9b00576

Teece, D. J. (2010). *Business models, business strategy and innovation.* Long Range Planning, 43(2–3), 172–194. https://doi.org/10.1016/j.lrp.2009.07.003

UK Government. (2025, April 28). *Frontier AI: capabilities and risks—discussion paper.* https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/frontier-ai-capabilities-and-risks-discussion-paper

Weill, P., Sebastian, I. M., Woerner, S. L., & Benedict, G. (2025, October 16). *Business models in the AI era* (MIT CISR Research Briefing No. XXV-10). MIT Center for Information Systems Research. https://cisr.mit.edu/publication/2025_1001_BizModelsAIEra_WeillSebastianWoernerBenedict

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). *ReAct: Synergizing reasoning and acting in language models.* arXiv. https://doi.org/10.48550/arXiv.2210.03629

https://doi.org/10.60030/ALMAMS.c202501nLDqvQ

# Policing and Artificial Intelligence: Cooperation or a Technological Arms Race?

Gábor Ferenczi

Senior Government Counselor
Ministry of Interior
e-mail: gabor.ferenczi@bm.gov.hu
https://orcid.org/0000-0002-2522-9210

Éva Sütő

Chief Government Counselor
Ministry of Interior
e-mail: eva.suto@bm.gov.hu
https://orcid.org/0009-0006-1838-2096

## Abstract

The rapid advancement of artificial intelligence (AI) is fundamentally reshaping policing, decision-making processes, and organizational functioning. The technology opens up new possibilities for detecting criminal patterns, conducting predictive analyses, and supporting strategic decision-making, while crime itself is expanding into increasingly digital domains. Particular significance is gained by the examination of linguistic evidence, online communication, and texts designed to obscure authorship or identity, where the collaboration between forensic linguistics and AI offers novel perspectives. Rather than signalling a reduction in police personnel, future developments point toward a transformation of professional roles: alongside traditional policing competencies, digital forensics, information technology, forensic linguistic analysis, and legal-ethical expertise are becoming increasingly vital. The integration of intelligent technologies may inaugurate a new era, one that prioritizes knowledge-sharing and interdisciplinarity over hierarchical organizational models. At the same time, offenders also exploit AI through cyberattacks,

manipulation, and concealment techniques, contributing to an emerging technological arms race. The future of criminalistics will thus be shaped by the depth and balance of cooperation between humans and machines.

***Keywords:*** *artificial intelligence, policing, criminalistics, forensic linguistics, linguistic evidence, organizational transformation*

## Absztrakt

A mesterséges intelligencia (MI) rohamos fejlődése alapjaiban alakítja át a rendészetet, a döntéshozatalt és a szervezeti működést. A technológia új lehetőségeket nyit a bűnügyi mintázatok felismerésében, a prediktív elemzésekben és a stratégiai döntéstámogatásban, miközben a bűnözés is digitális dimenziókkal gazdagodik. Kiemelt jelentőségűvé válik a nyelvi bizonyítékok, az online kommunikáció és az anonimitást célzó szövegek vizsgálata, ahol az igazságügyi nyelvészet és az MI együttműködése új távlatokat nyit. A jövő nem a rendőri létszám csökkentését, hanem a feladatkörök átalakulását vetíti előre: a hagyományos rendőri kompetenciák mellett felértékelődnek a digitális kriminalisztikai, informatikai, igazságügyi nyelvi elemzői és jogi-etikai készségek. Az intelligens eszközök bevezetése új korszakot nyithat, amely a hierarchikus modell helyett a tudásmegosztásra és interdiszciplinaritásra épít. Ugyanakkor a bűnözők is kihasználják az MI-t kibertámadások, manipulációk és rejtőzködési technikák révén, és ez egyfajta technológiai versenyfutást indít. A kriminalisztika jövőjét így az ember és a gép közötti kiegyensúlyozott együttműködés mélysége határozza meg.

***Kulcsszavak:*** *mesterséges intelligencia, rendészet, kriminalisztika, igazságügyi nyelvészet, nyelvi bizonyítékok, szervezeti átalakulás*

## 1 Introduction

The rapid development of artificial intelligence (hereinafter: AI) is opening new perspectives across several domains of policing and law enforcement studies. This paper focuses on the intersection of criminalistics and forensic linguistics – two young and dynamically

evolving scientific fields – and examines how AI may support their interaction. Our aim is to highlight the criminological relevance of applied linguistics and, through an interdisciplinary approach, to identify those points of convergence that may contribute to the renewal of analytical methodologies.

The integration of AI into Hungarian policing is currently in its early stages. Wider practical implementation is constrained primarily by financial limitations, gaps in data protection, legal and ethical regulation, and the strongly hierarchical organizational culture characteristic of law enforcement agencies. At the same time, domestic research initiatives and participation in international projects may facilitate the gradual development of an AI-aware policing culture that combines technological capabilities with human expertise. Although the pace of change may presently appear slow, incremental adaptation and pilot projects suggest that over the next 5–10 years AI is likely to become an integral part of everyday police work in an increasing number of areas – particularly in the analysis of linguistic evidence, communication patterns, and digital networks.

## 2. Foundations and Methodology

### 2.1 Criminalistics, Forensic Linguistics, and AI

Criminalistics is one of the foundational pillars of police and law enforcement studies: through scientific procedures – crime scene trace detection, evidence analysis, and expert examinations – it directly supports crime investigation, evidentiary processes, and lawful actions performed by authorities. Its core lies in an evidence-centred approach, that is, the professional recording and evaluation of evidence capable of withstanding judicial scrutiny. It also introduces interdisciplinary innovation by integrating findings from the natural sciences (biology, chemistry, physics) and the social sciences (linguistics, psychology, sociology) with modern technologies (digital forensics, information technology, AI). As a bridge between theory and practice, it operationalises the theoretical frameworks of policing for everyday law enforcement work, thereby increasing efficiency and public trust; without it, the credibility of prevention, investigation, and evidentiary processes would not be ensured (see, among others, Saferstein–Roy 2020; in Hungarian see Kovács 2021; Fenyvesi et al. 2022).

A semiotic perspective is particularly significant within criminalistics, as traces function as signs that serve as the fundamental carriers of evidential value. Iconic signs may offer hypotheses at most; they become procedurally admissible indices only when demonstrable causal mechanisms are established alongside similarity (e.g. temporal and spatial data, the reconstruction of relational chains, exclusion of alternative explanations). Linguistic signs hold equivalent importance in forensic linguistics: patterns of speech and writing, lexical choices, and textual structures can be interpreted as signs of communicative intent, social identity, and the speaker's or author's orientation toward reality. Thus, indexical linguistic material can equally be treated as part of the evidentiary process (see Szívós 2012).

The intersection of criminalistics and forensic linguistics opens new dimensions in law enforcement by enabling a multifaceted examination of evidence. This process is supported by AI through large-scale data analysis and pattern recognition, significantly enhancing the efficiency of analytical procedures. Within this system, linguistics has a distinctive role in analysing texts, voices, and speech samples designed to conceal identity, as well as in authorship attribution—tasks that enable more precise, rapid, and reliable forensic conclusions by combining technological tools with human expertise (for forensic linguistics see, among others, McMenamin 2002; Coulthard–Johnson 2007; Olsson 2019; for Hungarian overviews see Kontra 2003; for interdisciplinary links see Tolnainé Kabók 2015).

With the practical integration of AI, the role of linguistic analysis is becoming increasingly prominent both in supporting investigations and in processing evidence intended for judicial proceedings. Language use is one of the most characteristic and least easily manipulated human activities; therefore, the examination of written and spoken texts can reveal numerous traces that assist in identifying offenders, uncovering organisational networks, or demonstrating communication contexts related to criminal acts. Advances in AI have ushered in a new era in this field: stylometric, phonetic, and interpretive text analyses—once conducted manually or with limited statistical tools—can now be supported by large language models, machine-learning algorithms, and network-analytic methods (see Argamon et al. 2009; Kicsi et al. 2022).

## 2.2 AI-Supported Forensic Linguistic Analysis and Criminalistic Competencies: Research Background

In recent years, forensic linguistic analysis has become one of the most dynamically developing areas of law enforcement, particularly through the integration of artificial intelligence. AI enables the efficient processing and analysis of large volumes of digital data – such as emails and social media posts – as forms of linguistic evidence (for the nature of linguistic data and data informatization see Ferenczi 2023). At the same time, this requires substantial technological infrastructure: without adequate hardware and software capacity, and without modern database systems, such methods cannot be effectively applied.

Human expertise remains crucial. AI can realize its full potential only when trained linguists, criminalistic experts, and data analysts work collaboratively to interpret analytical results. Ethical and legal frameworks are also decisive: the use of AI-based analyses in judicial or investigative contexts requires strict regulation in terms of data protection and legal admissibility. The development of ethical protocols and verification procedures often slows down practical implementation (see Billah 2025).

At the international level, forensic linguistic analysis is already an established practice, widely used in both law enforcement and crime prevention. In Hungary, by contrast, the field is still largely in an experimental and research-oriented phase, although interest in AI-supported linguistic examinations is steadily increasing. The core aim of this domain is to integrate linguistics, Natural Language Processing (hereinafter: NLP), and machine learning in order to support law enforcement agencies in investigations, evidence evaluation, and the examination of digital communication. One of the most important application areas is the authorship attribution of anonymous texts (see Kicsi et al. 2022; Ürmösné–Kudar 2025).

In Hungary, linguists are typically employed for specialised and well-defined tasks: the analysis of linguistic evidence, the identification of communication patterns, and the interpretation of written and spoken evidentiary material. Within expert circles, document analysis is common, and research collaborations between universities and research institutes frequently focus on the criminalistic applicability of speech and text analysis. The linguist's interdisciplinary role in linguistic profiling and

investigative groundwork is especially important and may at times be indispensable (see Ránki 2021). A well-trained forensic linguist is capable of combining linguistic expertise with AI technologies, thereby enhancing the effectiveness of investigations.

Practical applications cover a wide spectrum. In the investigative phase, the analysis of unidentified or coded messages aids in determining the profile of suspects or authors. Based on writing patterns, lexical choices, grammatical structures, slang usage, and specialised code languages, inferences may be drawn regarding the offender's social background, psychological state, or regional affiliation (see Herke 2021). The analysis of speech and audio recordings makes it possible to identify dialect regions and estimate native language and educational level. Such linguistic findings can serve as the basis for expert testimony in judicial proceedings, for example when assessing the authenticity of writing samples or audio recordings.

The role of the linguist extends beyond evidence evaluation to crime prevention and criminological research. Identifying criminal communication patterns and behavioural tendencies can support preventive strategies and contribute to counter-radicalization efforts. Online criminalistics is also a key area, in which monitoring the darknet, social media, and online forums enables the identification of dangerous communication and organised illicit activity (Lontai et al. 2024).

The AI toolkit supporting forensic linguistic analysis is highly diverse. Contemporary NLP models offer context-sensitive processing. Stylometry allows the quantitative examination of patterns embedded in texts – such as lexical preferences, sentence length, and rare expressions (Olsson 2019). The combination of forensic phonetics and machine learning facilitates the comparison and identification of speech samples (McMenamin 2002). Sentiment analysis and topic modelling support the detection of underlying intentions and thematic structures, whereas network analysis and data visualization are suitable for mapping the communicative structures of criminal organisations (Coulthard–Johnson 2007).

## 3 Context and Risks

### 3.1 Predictive Policing and Criminal Adaptation – Opportunities and Risks

Traditionally, policing has not been regarded as a direct shaper of geopolitical developments. With the emergence of artificial intelligence, however, this situation may change. Predictive policing systems are capable of estimating the likelihood of criminal events and identifying potential offenders. Yet their methodology involves significant risks: individuals may be wrongly targeted solely on the basis of their online activity or communication patterns. Such scenarios can lead to extreme outcomes, including unjustified arrests or the political manipulation of these systems. These risks underscore the need for robust human oversight, without which AI-driven decisions may have severe social and foreign policy consequences (see Egbert–Leese 2020, Grimm et al. 2021).

AI may also be adopted by organized criminal groups, which could employ the technology to support their operations. AI has the potential to revolutionize crime: aspects of criminal activity could be partially or fully delegated to automated systems, from identifying victims and selecting offenders to profiling potential perpetrators. The technology could enable the monitoring, ranking, and even coordination of individuals for the execution of criminal acts, particularly in the domain of cybercrime. This raises a wide range of legal and ethical concerns, including the protection of personal liberty and fundamental rights. AI-enabled crime is expected to emerge across national borders, potentially generating international tensions. These developments present governments and law enforcement agencies with new challenges: in this technological race, the balance of power will be determined by whether law enforcement or criminal actors succeed in deploying AI more rapidly and effectively (see Mehrotra et al. 2025, and see also the discussion of Prakash 2020 in Sütő 2023).

### 3.2 International Outlook

International practice shows that AI-based linguistic analysis is already widely established and routinely applied. Numerous studies report the use of AI for authorship attribution on anonymous darknet forums (see, among others, Sennewald et al. 2020, Ranaldi et al. 2022). Other research focuses on the automatic detection of radicalization discourses on social media, an

essential element of counterterrorism strategies (see Al-Shawakfa 2025). In the United States and Western Europe, law enforcement agencies regularly employ NLP tools under strict conditions to monitor online communication, for example to filter texts associated with terrorism or organized networks.

Experiments conducted by Europol and the FBI have demonstrated that AI can detect distinctive linguistic features in the writing patterns of anonymous forum users, thereby facilitating their identification. According to a previous Europol report, AI promises substantial efficiency gains in policing, although it also carries significant ethical and legal risks (bias, data protection, fundamental rights). The European Union's AI Act explicitly aims to balance innovation with legal safeguards through regulated experimentation, transparency, accountability, and trust-building (see Europol 2024). According to last year's report by the U.S. Department of Justice, the FBI may use AI only under clearly defined conditions: explicit purpose specification, designated responsibilities, transparent data sources, mandatory human review, and realistic pre-deployment testing, followed by continuous oversight and community consultation. For high-risk decisions, AI outputs can never serve as sole grounds for action or evidentiary conclusions (see Artificial 2024).

Judicial admissibility also presents a major challenge: profiles generated by machine-learning algorithms are often accepted only as supplementary evidence because of concerns regarding methodological transparency and error margins. Despite these limitations, international trends are clear: forensic linguistic analysis increasingly relies on AI, particularly in the mapping of communication within digital environments (Grimm et al. 2021).

Internationally, the implementation of AI-supported forensic linguistic methods is considerably more advanced, with well-established practices. A few prominent examples include:

- Aston University (United Kingdom), Institute for Forensic Linguistics, a global leader in authorship attribution, stylometry, and courtroom applications.
- Shlomo Argamon (United States), one of the most recognized experts in computational stylometry and forensic linguistic profiling.

- Marilu R. Madrunio (Philippines), a prominent international representative of forensic linguistics, widely published on legal and forensic linguistic analysis.

### 3.3 The Situation in Hungary

In Hungary, the development of the field is progressing at a more moderate pace, yet it remains consistently present in both academic and practical law enforcement contexts. Domestic research and participation in international projects create an opportunity for the gradual emergence of an AI-aware policing culture that draws simultaneously on technological innovation and human expertise.

Within the university sphere, the primary institutions involved in forensic linguistics and the policing applications of digital data analysis are Eötvös Loránd University (ELTE) and the University of Public Service (NKE). At ELTE's Faculty of Humanities, the Department of Applied Linguistics and Phonetics conducts research in the field and also offers a dedicated forensic linguistics course. At NKE, AI features prominently in the curriculum, especially in policing and legal education. Further courses and research projects are expected to be launched in the coming years, aimed at deepening and operationalising AI applications in law enforcement. At the research-institute level, the Expert Institute of the Special Service for National Security, the National Centre for Expert and Research (NSZKK), and ELTE's Research Centre for Linguistics (formerly: HUN-REN Research Centre for Linguistics) are all engaged in studies that apply speech and text analysis for criminalistic purposes.

In recent years, several domestic conferences have addressed forensic linguistics (see, among others: I. Applied Linguistics Doctoral Conference, Hungarian Academy of Sciences, Institute of Linguistics, Budapest, 2007; Forensic Linguistics Conference, ELTE Hungarian Language Strategy Research Group, Budapest, 2019; 16th Hungarian Conference on Computational Linguistics, University of Szeged, Institute of Informatics, 2020; Bozontvágó vagy bozótvágó – avagy bevezetés az igazságügyi nyelvészetbe, NKE RTK Szent György College for Advanced Studies, Budapest, 2023). Increasing attention is being paid to the automated processing of linguistic evidence as well (see Főző–Tatár 2022, Kicsi et al. 2022). The demand for developing AI-based criminalistic applications has already appeared within domestic specialised research,

indicating that in the coming years a gradual convergence with international trends can be expected (for early initiatives see Főző 2023).

## 4 Organisational Implementations and Knowledge

### 4.1 Knowledge Sharing in Policing Supported by AI

Effective law enforcement and public safety today rely not only on human experience and traditional data collection, but also on modern technologies capable of structuring and making accessible accumulated knowledge. In policing practice, knowledge sharing previously meant the transmission of investigative experience, statistical data, and local or national crime reports. AI, however, elevates this process to a new level. It can identify patterns that human analysts alone would not, or only significantly more slowly.

AI is able to process and structure information from multiple sources – criminal records, CCTV footage, and online communication patterns – and synthesise these into a knowledge base that integrates practical experience with real-time data. Predictive analyses can provide forecasts of where, when, and what types of crime are likely to occur; this knowledge can then be rapidly disseminated to other organisational units.

The process also fosters interdisciplinary cooperation: linguists, criminalistic specialists, IT professionals, and data analysts jointly interpret AI-generated analyses, enabling both horizontal knowledge flows (across different police units) and vertical knowledge flows (between research and practice). Large international organisations such as Europol and the FBI already rely on AI-based databases and knowledge-sharing systems. For Hungary, adopting similar systems will be crucial in combating transnational and international crime.

The trust factor, however, remains indispensable: policing decisions must continue to be overseen by human expertise. AI does not replace but complements and strengthens knowledge-sharing processes by structuring vast quantities of information, facilitating the rapid exchange of experience and analysis, and contributing to a preventive approach to policing. Continuous communication with international partner agencies is also essential, as adapting their proven methodologies can provide substantial advantages in domestic practice (for AI applications in policing

see, among others, Döbó–Gyaraki 2021, Máté 2024; for comprehensive discussions of knowledge sharing and knowledge management see Noszkay 2023).

### 4.2 The Impact of AI on Policing Organisational Culture

The introduction of AI is expected to exert a dual influence on the organisational culture of policing. On the one hand, it reinforces a data-driven perspective in which decisions increasingly rely on objective analyses, statistical models, and predictive forecasts. This enhances organisational efficiency and accountability, as decision-makers no longer rely solely on personal experience or intuition but draw upon large volumes of real-time data. An evidence-based policing culture can also strengthen public trust by making decision-making processes more transparent and reproducible.

On the other hand, AI integration presents major organisational and cultural challenges. The traditionally hierarchical, experience- and routine-oriented operation of policing agencies must be reshaped to accommodate technological innovation. This includes the acquisition of new competencies, openness to continuous learning, and the strengthening of interdisciplinary collaboration: cooperation among diverse specialists fosters a new organisational culture in which horizontal knowledge sharing becomes as important as the vertical chain of command.

In the long term, AI may contribute to the development of a more open, innovation-oriented organisational culture. This shift is not only technological but also attitudinal: policing organisations must move from rigid structures toward flexible and adaptive modes of operation. As a result, bureaucratic rigidity may diminish and the organisation's ability to respond rapidly to evolving security challenges may increase.

International research suggests that AI integration in policing will fundamentally transform leadership decision-making, professional roles, and organisational identity. AI does not replace the human factor but reshapes its functions: the tasks of officers and specialists will increasingly involve overseeing, interpreting, and ethically monitoring technological tools (for visions of future policing see Egbert–Leese 2020; for practical approaches and experiences see Europol 2023, Artificial 2024).

## 5 Conclusion

Within policing and law enforcement studies, AI represents not merely a technological innovation but a catalyst for complex change, transforming practice on multiple levels – from analytical methods to organisational culture. At the intersection of criminalistics and forensic linguistics, AI can function as a comprehensive toolkit that enables more rapid and accurate authorship attribution, the examination of communication patterns, and the mapping of digital networks. At the same time, automated procedures raise legal, ethical, and data-protection concerns, especially regarding the judicial admissibility of evidence and the risks of algorithmic bias or misidentification.

The role of forensic linguistics in the future of criminalistics will be highly significant. Language use is one of the most resistant human behaviours to manipulation and reveals valuable traces regarding offenders' social backgrounds, regional affiliations, psychological states, and identities. AI-supported stylometric, phonetic, and textual analysis methods are capable of identifying subtle patterns that would be difficult or impossible to detect through human analysis alone. Such knowledge carries significant value not only for investigations but also for courtroom evidence and preventive strategies.

The criminalistic application of linguistics also forms a bridge between science and practice. Linguists working in interdisciplinary frameworks provide added value by integrating linguistic expertise with AI technologies. International trends – from Europol and FBI's AI-based knowledge-sharing systems to the forensic linguistics school at Aston University and the NLP-driven practices of Western European agencies – demonstrate that the methodology is rapidly becoming standardised, while requirements for usability and transparency are simultaneously strengthening. These experiences can serve as a compass for domestic implementation.

Overall, the future does not forecast a reduction in police personnel but rather a transformation of competencies: traditional policing experience will be complemented by digital, linguistic, IT, and ethical skills. The key to success lies in gradual and controlled integration, where the advantages offered by AI are always paired with human expertise. Equally crucial is the evolution of organisational culture: knowledge sharing and horizontal

collaboration gain increasing importance, replacing rigid hierarchical structures. AI is capable of structuring and making accessible accumulated experience, supporting swift and transparent decision-making. This flow of knowledge may form the basis of a new policing culture built on innovation, interdisciplinarity, and preventive orientation. In this way, the relationship between policing and AI need not be a race but a balanced cooperation, ultimately yielding more effective, open, and trust-enhancing law enforcement practice.

## References

Al-Shawakfa, E. M. – Alsobeh, Anas M. R. – Omari, Sahar – Shatnawi, Amani (2025). An Ensemble Approach for Radicalization Detection in Arabic Tweets. In *Information*, 16. 7: 522–555. https://doi.org/10.3390/info16070522

Argamon, Shlomo – Koppel, Moshe – Pennebaker, James W. – Schler, Jonathan (2009). Automatically Profiling the Author of an Anonymous Text. In *Communications of the ACM*, 52. 2: 119–123. https://doi.org/10.1145/1461928.1461959

Artifical ˙(2024). *Artificial Intelligence and Criminal Justice. Final Report.* U. S. Department of Justice, USA.

Billah, Maruf (2025). Developing an Explainable AI System for Digital Forensics: Enhancing Trust and Transparency in Flagging Events for Legal Evidence. In *Journal of Forensic Science Research*, 9. 2: 109–116. https://doi.org/10.29328/journal.jfsr.1001089

Coulthard, Malcolm – Johnson, Alison (2007). *An Introduction to Forensic Linguistics: Language in Evidence.* Routledge, London–New York.

Döbó, Judit – Gyaraki, Réka (2021). A mesterséges intelligencia egyes felhasználási lehetőségei a rendvédelmi területeken. In *Magyar Rendészet*, 21. 4: 67–81. https://doi.org/10.32577/mr.2021.4.3

Egbert, Simon – Leese, Matthias (2020). *Criminal futures. Predictive policing and everyday police work.* Routledge, London. https://doi.org/10.21428/cb6ab371.17e3e7ab

Europol (2024). *AI and policing. The benefits and challenges of artificial intelligence for law enforcement.* An Observatory Report from the Europol Innovation Lab. European Union, Luxembourg.

Fenyvesi, Csaba – Herke, Csongor – Tremmel, Flórián eds. (2022). *Kriminalisztika.* Ludovika Egyetemi Kiadó, Budapest.

Ferenczi, Gábor (2023). Nyelvi adat és adatinformatizálás: adalékok egy nyelvföldrajzi kutatás elméletéhez. In *Magyar Nyelvőr,* 147. 4: 427–441. https://doi.org/10.38143/Nyr.2023.4.427

Főző, Eszter – Tatár, Zoltán (2022). Gépi becslés a szerzőségvizsgálatban, a SZIRA gépi szerzőazonosító funkciójának validálása. In Borza, Veronika – Főző, Eszter – Kóré, Veronika – Markó, Alexandra – Mell, Péter – Nyiri, Miklós – Tóth, István László (eds.), *Aktuális kérdések a szakértésben a biometriától a vegyészetig.* Nemzetbiztonsági Szakszolgálat, Budapest. 52–81.

Főző, Eszter (2023). Kommunikáció 4.0: A ChatGPT nyelvi képességeinek vizsgálata nyelvész szakértői szemmel In Borza, Veronika – Főző, Eszter – Kóré, Veronika – Markó, Alexandra – Mell, Péter – Nyiri, Miklós – Tóth, István László (eds.), *Régi és új kihívások az igazságügyi szakértői munkában.* Nemzetbiztonsági Szakszolgálat, Budapest. 56–85.

Grimm, Paul W. – Grossman, Maura R. – Cormack, Gordon V. (2021). Artificial Intelligence as Evidence. In *Northwestern Journal of Technology and Intellectual Property,* 19. 1: 8–106.

Herke, Csongor (2021). A mesterséges intelligencia kriminalisztikai aspektusai. In *Belügyi Szemle,* 69. 10: 1709–1724. https://doi.org/10.38146/BSZ.2021.10.2

Kicsi, András – Sánta, Péter – Horváth, Dániel – Kőhegyi, Norbert – Szvoreny, Viktor – Vince, Veronika – Főző, Eszter – Vidács, László (2022). Computer-Aided Forensic Authorship Identification in Criminology. In *Computational Science and Its Applications – ICCSA 2022 Workshops. Lecture Notes in Computer Science,* 13381. Springer International Publishing. 1: 576–592. https://doi.org/10.1007/978-3-031-10548-7_42

Kontra, Miklós (2003). Nyelv és jog. Igazságügyi nyelvészet. In Kiefer, Ferenc (ed.), *A magyar nyelv kézikönyve.* Akadémiai Kiadó, Budapest. 551–556.

Kovács, Gyula (2021). Kriminalisztikai alapkérdések. Kérdések és válaszok a nyomozás során. In *Glossa Iuridica,* 8. 1–2: 183–197.

Lontai, Márton – Pamzsav, Horolma – Petrétei, Dávid (2024). Mesterséges intelligencia a törvényszéki tudományokban. Revolóció vagy invázió? In *Belügyi Szemle,* 72. 4: 577–592. (I.), 8: 1355–1369. (II.)

Máté, Szilveszter (2024). Mesterséges intelligencia és a rendészetelmélet. In *Belügyi Szemle*, 72. 8: 1387–1403. https://doi.org/10.38146/bsz-ajia.2024.v72.i8.pp1387-1403

McMenamin, Gerald R. (2002). *Forensic Linguistics: Advances in Forensic Stylistics.* CRC Press. https://doi.org/10.1201/9781420041170

Mehrotra, Siddharth – Gadiraju, Ujwal – Bittner, Eva – van Delden, Folkert – Jonker, Catholijn M. – Tielman, Myrthe L. (2025). "Even explanations will not help in trusting [this] fundamentally biased system": A Predictive Policing Case-Study. In *33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '25), June 16–19, 2025, New York City, NY, USA.* ACM, New York, NY, USA. 51–62. https://doi.org/10.1145/3699682.3728343

Noszkay, Erzsébet ed. (2023). *Tudásmenedzsment a következő két évtized határán. Az MTA GB TM Munkabizottság 20 éves jubileumára készült tanulmánykötet.* Akadémiai Kiadó, Budapest. https://doi.org/10.1556/9789634549130

Olsson, John 2019. *Forensic Linguistics.* 3rd edition. Bloomsbury Academic, London.

Prakash, Abishur (2020). *Go.AI: A mesterséges intelligencia geopolitikája.* Pallas Athéné Könyvkiadó, Budapest.

Ranaldi, Leonardo – Ranaldi, Federico – Falluchi, Francesca – Zanzotto, Fabio Massimo 2022. Shedding Light on the Dark Web: Authorship Attribution in Radical Forums. In *Information*, 13. 9: 435–451. https://doi.org/10.3390/info13090435

Ránki, Sára (2021). Nyelvi profilalkotás. In *Belügyi Szemle*, 69. 12: 2155–2166. https://doi.org/10.38146/BSZ.2021.12.6

Saferstein, Richard – Roy, Tiffany (2020). *Criminalistics. An Introduction to Forensic Science.* Pearson, London.

Sennewald, Britta – Herpers, Rainer – Hülsmann, Marco – Kent, Kenneth B. (2020). Voting for authorship attribution applied to dark web data. In Shaddick, Lily – Jourdan, Guy–Vincent – Onut, Vio – Ng, Tinny (eds.), *CASCON '20: Proceedings of the 30th Annual International Conference on Computer Science and Software Engineering. Toronto Ontario Canada November 10–13, 2020.* IBM Centre for Advanced Studies (CAS), Canada. 217–226.

Sütő, Éva (2023). Prakash, Abishur: Go.AI: A mesterséges intelligencia geopolitikája. Pallas Athéné Könyvkiadó, Budapest. 2020. 204. o. In

*Külügyi Szemle*, 22. 2: 144–150. https://doi.org/10.47707/Kulugyi_Szemle.2023.2.9

Szívós, Mihály (2012). *A jeltől a kódig. Rendszeres szemiotika.* Loisir Kiadó, Budapest.

Tolnainé Kabók, Zsuzsanna (2015). Interdiszciplináris kapcsolatok a rendészettudományok és az alkalmazott nyelvészet között – különös tekintettel a törvényszéki nyelvészetre. In *Magyar Rendészet*, 15. 5: 131–145.

Ürmösné Simon, Gabriella – Kudar, Mariann (2025). The Role of Forensic Linguistics and Case Studies, Demonstrating the Effectiveness of Linguists' Contribution to the Investigations. In *Magyar Rendészet*, 25. Special Issue: 159–169. https://doi.org/10.32577/MR.2025.KSZ.1.10

https://doi.org/10.60030/ALMAMS.c202501efkEfb

# Artificial Intelligence, Explicit Online Knowledge, and Tacit Knowledge – From Knowledge Boundaries to the First Knowledge Asymmetry

**Mihály Szívós DSc**

szivosml@yahoo.de
https://orcid.org/0009-0003-5388-5447

## Abstract

The study connects three themes. In relation to AI, anthropomorphizing tendencies can be observed, as many people overestimate its capabilities. These views stem from the insufficiently recognized fact that AI relies exclusively on online explicit knowledge resources and cannot, or can only to a very limited extent, access tacit knowledge, its subtypes, or network knowledge. The second part of the study presents these types of knowledge. Since AI can process only online explicit knowledge with increasing intensity, a significant knowledge asymmetry emerges: the utilization of the other two knowledge domains does not grow at a comparable rate. As a consequence of this asymmetry, new inequalities and other societal problems may develop. Among these, the most significant appears to be the further strengthening of the resources of various functional centres and their impact on society.

***Keywords:*** *knowledge boundary, type of knowledge, knowledge asymmetry, anthropomorphization, network knowledge, AI-visibility of knowledge, efficiency of AI knowledge processing*

## *Motto*

*„It is worth bearing in mind…the general rule that we must not expect to find simple labels for complicated cases…however well-equipped our*

*language, it can never be forearmed against all possible cases that may arise and call for description: fact is richer than diction."*

*John Langshaw Austin*

## Introduction

Artificial intelligence (hereinafter AI) uses language models both to interpret the questions posed to it and to construct the answers it provides. For this reason, it is particularly important for AI users to be aware of the warning issued by the well-known English philosopher of language, John Langshaw Austin – both when formulating their questions and when interpreting the answers. Without this awareness, they may easily fall into the error of overestimating AI's capabilities. The motto indirectly reminds us of the importance of tacit linguistic knowledge as well. I will return to this point when discussing tacit knowledge.

The work of defining and systematizing different types of knowledge began already in the first half of the 20th century. The emergence of new disciplines, such as the sociology of knowledge – which has significant Hungarian roots through Karl Mannheim and other members of the Sunday Circle – pointed in this direction. The most significant milestone in the philosophy of science and theory of knowledge, however, proved to be Michael Polanyi's theory of tacit knowledge. This sharply formulated theory (Polanyi, 1958) not only confirmed the suspicion that multiple forms of human knowledge exist, but also made it necessary to describe and define them within philosophy – especially the philosophy of science – as well as within the sociology of knowledge. These results later became foundational for the development of knowledge management.

The present study consists of three main parts. The first part deals with the possibility of de-anthropomorphization, which is needed because AI's capabilities are often overestimated. The structure of the second and third parts is based primarily on the fact that although the knowledge resources available to AI are enormous, they are nevertheless limited. This limitation arises from the fact that certain types of knowledge are not accessible to AI at all, or only to a very limited extent. To explore the relationship between AI and the various knowledge resources, several key concepts must be clarified. In developing these concepts, it will inevitably be necessary to outline some properties of these knowledge types. From this, it will also become clear why they are inaccessible – or only partially accessible – to artificial intelligence.

Given that AI uses only certain types of knowledge and processes them with very high intensity, it follows that the status of other knowledge types also begins to change. It would be a mistake to underestimate the social consequences and significance of these changes. The fourth part of the study attempt to map these emerging difficulties.

## 1. The Importance and Limits of De-anthropomorphization in the Use of AI

I believe that a multidirectional analysis of AI's use of knowledge can also help us find reference points for reducing the anthropomorphization that surrounds AI. It is evident that, largely due to the film industry and science-fiction literature, the similarity of AI to human thinking is greatly overestimated. The paradoxes emerging within this wave of overestimation recall an old computer joke: "You should never anthropomorphize computers, because they really don't like it." This joke wittily points out that a kind of spontaneous anthropomorphization had already been taking place among those who worked with computers. However, this may amplify misunderstandings related to AI, especially in everyday life, where this highly significant technological achievement is increasingly used. Unfortunately, even a small portion of developers and professional users are also prone to fall into this anthropomorphizing error. These are the main reasons why persistent fantasies continue to circulate about AI "becoming self-aware," "turning against humans," and so on. Radical anthropomorphization also obstructs clear thinking regarding the important question of what the knowledge boundaries of AI are, and how the relationship between the types of knowledge accessible to AI and those inaccessible to it is being transformed.

Expressions used in relation to AI often include linguistic turns that were previously associated with "ordinary" computers as well. For this reason, it is difficult to avoid anthropomorphizing words, as they have become deeply embedded in the world of computing. Thus, I too will use expressions such as "AI can see this or reach that," or "AI uses this," even though more precise, de-anthropomorphizing formulations would be preferable. For example: "In AI's designed circuits, the following process takes place at this moment." Such expressions, however, would be far more cumbersome. Since these anthropomorphizing expressions have already become standard in the literature, it follows that many of them will remain in use in the emerging AI-related academic discourse as well. By

recalling the old computer joke and its original context, I simply wish to signal that although I also use such anthropomorphizing expressions, I strive to remain aware of their nature and encourage others to do the same. I hope this may be a small contribution to fostering a clearer understanding of AI's real capabilities, their character, and simultaneously their limitations, and to avoiding the spontaneous encouragement of exaggerations related to them.

## 2. Key Concepts

Among the many key concepts associated with AI, I examine here only those that are closely related to AI-related knowledge management. A further limitation is that the types of knowledge discussed here – tacit knowledge and network knowledge (i.e. the knowledge flowing through hidden social networks) – play a somewhat more important role in the life sciences, social sciences, and humanities than in the natural sciences. The difficulty mentioned by Austin, which arises when verbally describing facts, also appears more prominently in the fields of life sciences, social sciences, and humanities.

To discuss AI's knowledge management, we also need several concepts that denote knowledge units of different sizes. A knowledge package is what AI provides to the user in response to a single query. But a knowledge package is also what AI uses internally to construct its answer. A knowledge set is the collection from which AI assembles the individual knowledge packages needed to answer similar questions. The largest units of knowledge are knowledge worlds, which consist of specific knowledge sets. Missing from these knowledge worlds are those knowledge sets that AI can access only minimally or not at all. Thus, in every domain, AI possesses only limited knowledge sets due to various constraining factors.

### 2.1 The Degree of Richness of a Fact

The essential elements of the first key concept can already be found in the motto. The concept of a "social fact" came to the forefront during the emergence of the social sciences – particularly sociology – at the end of the 19th and the beginning of the 20th century. This classical term was developed by Émile Durkheim, one of the founders of the discipline, in his study The Rules of Sociological Method, first published in 1895. The concept gained prominence because, although statistical data are important

in this field as well, the elements of social phenomena that come into focus are much more difficult to delineate and interpret. Durkheim lists among the primary examples laws, moral norms, religious beliefs, institutions, as well as money, fashion, and language – all of which determine individuals from the outside.

The question of defining facts came to the forefront in sociology rather than, for example, in history, because historians often work with facts whose significance has already been confirmed over time through their consequences. Sociologists, however, typically deal with present-day problems, making the selection and delineation of facts a far more complex task. Moreover, the separation and definition of facts never yields a final, unchanging result; it remains open to revision. Over time, not only the evaluation of facts change, but also their definition and their selection according to importance. Thus, the concept and evaluation of social facts underwent significant transformation from the late 19th century to the first quarter of the 21st century, in which the more precise differentiation of various types of knowledge also played an important role.

Since I also examine here the kinds of knowledge on which AI relies, the question of what we consider a fact cannot be avoided. It is therefore crucial to understand what kinds of factual descriptions AI takes into account when constructing its answers. This also means that we must point to the long list of problems associated with these questions. Due to space limitations, I can mention only the most fundamental one: according to Austin's epistemological and linguistic-philosophical thesis, facts are richer than what linguistic descriptions can express, even when those descriptions are highly nuanced and carefully considered. AI can therefore approach the richness of facts – and especially social facts – only to a limited extent, because it cannot access those elements of knowledge that we now commonly classify as tacit knowledge. Humans connect these tacit elements to the facts they perceive, and indeed interpret the facts through them, but cannot articulate them in words. The use of large language models, in which AI is particularly strong, also suggests that Austin's thesis applies here as well, since the knowledge-accessing capabilities of these models are likewise limited.

At this point, it may also undoubtedly be raised whose perspective determines the importance of social facts: one may examine which decision-makers, strata, classes, or social groups have considered these facts significant and have delineated them accordingly. Thus, there would

still be room here for certain criticisms from historians and sociologists as well.

The critical insights that arise in sociology concerning facts can be applied to all forms of knowledge based on verbalization, which reduces the reliability of these knowledge forms and, consequently, the probabilistic value of the conclusions that rely on them.

### 2.2 The Degree of AI-Visibility of Different Types of Knowledge

This concept refers to the varying accessibility of different types of knowledge from the perspective of AI. When considering collaborative work with AI, this concept is absolutely fundamental. In such collaboration, we must know in what areas our tool is strong and in what areas it is weak. Let us take an immediate example: works published online, which are available explicitly, possess a high degree of AI-visibility. This means that AI's language models are able to process them. This constitutes the enormous body of knowledge that AI can, in principle, access fully, since it is digitized and generally available. Thus, its AI-visibility is theoretically 100 percent. We will later see which factors may reduce this value.

AI is trained to know and communicate that it can access only online explicit (verbalized) knowledge sources. This could serve as an important reference point for separating from AI those strongly anthropomorphizing and capability-exaggerating assumptions that tend to attach to it. In this regard, however, the creators of marketing strategies associated with AI may even have opposing interests.

On a map of knowledge types, however, we also find categories that, as we shall see, cannot be said to be readily accessible to AI. Such are, for example, knowledge packages that are already public but not yet available electronically. Thanks to various digitalization programs, these too are gradually being transferred into online knowledge bases. Worldwide – and in Hungary as well – digitalization programs are underway through which vast quantities of publications from earlier centuries are becoming available online. Among Hungarian digitalization initiatives, one of the most important is that of the Library of the National Assembly. Digitalizing library collections is an important program everywhere, but capacities are, of course, limited. The light at the end of the tunnel is not yet visible. The universally accessible World Library project remains a

very distant plan, and it is far from certain that it can be realized. Owners of AI systems may, of course, commission affordable digitalization for specific tasks, and such expansion may even represent an additional source of revenue for them.

Copyright is also among the factors that strongly restrict accessibility. For seventy years after an author's death, the copyright holders retain control over the protected intellectual works, and it is not in their interest to diminish or relinquish these rights, as doing so would cause financial loss. Thus, the digitalization of an author's works – if not permitted during their lifetime – may be further delayed due to the rights of the heirs. Yet this grace period represents a significant limitation for AI, since in most cases the knowledge produced in the past seventy years is the most relevant for solving contemporary problems.

Among the types of knowledge with limited visibility is the so-called network knowledge, which, according to its preliminary definition, arises and circulates within informal human networks that operate spontaneously alongside institutions and enable knowledge exchange. Since the knowledge sets that emerge here rarely become known to the public and even more rarely appear online, they do not – or only barely – enter AI's sphere of accessibility. Yet these knowledge sets are extremely important for understanding the functioning of human societies, smaller and larger communities, and networks.

Tacit knowledge is typically acquired by a person through working with a tool, a material, an environment, an institution, or another individual. According to Michael Polanyi's classical thesis, this type of knowledge cannot be articulated, and thus every person knows far more than he or she can verbally express. This knowledge is tied to its possessors – the individual persons – and manifests and grows through their activities. Due to its non-verbalizable nature, tacit knowledge is fundamentally inaccessible to AI.

Given the operation of AI, linguistic tacit knowledge deserves special emphasis. It is our shared experience that a healthy six-year-old child, even before entering school, already speaks his or her mother tongue – or even multiple languages – fluently, without being aware that he or she is following linguistic rules. In addition to linguistic rules, native speakers possess tacit knowledge of the scope of meaning of every word they know, as well as the various nuances of those meanings. Relying on these

elements of knowledge, linguistic intuition operates, and it has contributed greatly to enabling speakers to understand sentences containing words they themselves may never have used before. This linguistic intuition supports the verbal recording of facts even when we can no longer cross the boundaries mentioned by Austin.

### 2.3 The Efficiency of AI's Knowledge Processing

In discussing this concept, we are still dealing with those types of knowledge that AI can access. Later, I will also address those that, for the reasons mentioned earlier, are no longer – or only barely – accessible to it. The accessibility of knowledge packages and knowledge sets is not equivalent to their optimal processability. Through this simple distinction, we arrive at another key concept: the efficiency AI can achieve when processing the knowledge sets available to it. As we have already seen in connection with the concept of fact, our linguistically formulated knowledge of certain parts of reality is incomplete due to linguistic limitations.

Another limitation arises from the fact that the value of knowledge packages is greatly influenced by their time of origin. The date of scientific results is important in every discipline, though not to the same degree. Not knowing – or disregarding – the time of origin of available knowledge may reduce AI's processing efficiency. This problem persists even when the knowledge package contains facts constructed through machine perception, because perceptual limitations exist there as well. Although the accuracy of machine-constructed facts may improve with technological development, certain limitations will always remain. Proper weighting of the validity and value of individual knowledge sets and knowledge packages therefore requires taking into account when they were created.

Each scientific discipline has a different so-called "half-life" of scientific results, meaning the period – measured in years or months – during which the validity of a given knowledge package decreases by half. In chemistry, for example, this half-life is very short. This decline occurs due to the rapid progress of science, as new results – while building on earlier ones – overwrite or even completely invalidate parts of previous findings. Depending on the average half-life characteristic of each discipline, one can determine how necessary it is to index scientific results by their time of origin when solving particular problems. AI can be trained to index its

knowledge sources by the time of discovery, and even more importantly, by the time of publication. The question, however, is how effectively it can weight these results when synthesizing them.

For example, AI can be contrasted with a researcher who has at least twenty years of experience and who has personally participated in the developmental processes within their discipline over those two decades. While AI may be able to estimate this researcher's explicit knowledge based on statistical averages, his tacit knowledge is neither accessible nor reconstructible for AI.

AI is capable of indexing the knowledge elements it uses according to their time of origin and of weighting them based on certain estimations. However, this does not mean that its evaluative weighting according to time of origin will necessarily be adequate. What further aggravates the difficulty is that AI cannot accurately determine the degree of insufficiency in its own knowledge with respect to the time of origin of the knowledge elements it uses.

## 3. A More Detailed Presentation of the Types of Knowledge That Fall Outside AI's Search Focus

The primary reason for discussing these types of knowledge in more detail is that they are far less well known than other forms of knowledge. Secondly, I can only illuminate the problem I have called AI's knowledge asymmetry if I outline in greater detail the differences between the two main groups of knowledge types. The situation has arisen in which some types of knowledge enter AI's search focus immediately or relatively easily, with only minor delay, while others enter it only minimally, indirectly, or not at all. The knowledge types belonging to these latter categories are, above all, tacit knowledge in its personal form and network knowledge.

### 3.1 Tacit Knowledge

The motto from Austin already hinted at one form of tacit knowledge without naming it explicitly. This refers primarily to the tacit linguistic knowledge already mentioned briefly, which we are unable to make explicit in our spoken or written sentences. Austin formulated this idea one year before Michael Polanyi published his major work on the theory of

tacit knowledge, Personal Knowledge. Austin's description resonates strongly with Polanyi's earlier, well-known insight that tacit knowledge cannot be made explicit. Polanyi later summarized this in a concise thesis: "… we know more than we can tell." (Polanyi, 1966:4). Austin in fact slightly expands Polanyi's thesis by linking the philosophical problem of formulating knowledge – making it explicit – to that part of knowledge that language cannot reach or represent in words. In this way, he indirectly connects it to that form of knowledge – tacit knowledge – that largely remains outside the process of explicit knowledge acquisition. This problem thus leads us to one of the key concepts relevant to AI: tacit knowledge, which we know is either inaccessible or only minimally accessible to AI.

Some researchers have attempted to make tacit knowledge explicit, or in other words, to externalize it. However, the structural characteristics of tacit knowledge contradict this idea. Since its structure differs from that of explicit knowledge in several respects, these characteristics are lost during the externalization process. One can therefore argue convincingly that the resulting knowledge may be interesting, but it is certainly no longer tacit knowledge, because its structure differs from that of explicit knowledge. Although descriptions exist of what various forms of tacit knowledge contain, and although their structure can be explored through complex research, the content of individual tacit knowledge packages cannot be made explicit. Thus, they do not become accessible to AI.

### 3.1 Types of Tacit Knowledge

Before the emergence of the theory of tacit knowledge, it would have been difficult to discuss the types of knowledge inaccessible to computers – and thus to AI. Polanyi attempts to define tacit knowledge in several domains, two of which are particularly important. One concerns its relationship to explicit knowledge, highlighting differences and connections. The other concerns its use: Polanyi illustrates through numerous examples how we necessarily rely on this form of knowledge even though we cannot articulate it or verbally transmit it to others.

Descriptions of the use of tacit knowledge led us to the insight that one possible classification of tacit knowledge arises from the nature of the object or domain to which it becomes attached in the human mind and through which it takes on a specific structural form. A further common feature of the subtypes of tacit knowledge briefly presented below is that

their definition is linked to the teleological concept of action, insofar as they relate to its components: goals and means. Here I can discuss these subtypes only briefly, as I have provided much more detailed definitions in several of my earlier writings (Szívós, 2017).

### 3.1.1 Tool-Centred Tacit Knowledge

Depending on their occupation – but also in everyday life – people regularly use tools whose smooth handling requires experience and, within that, specific tacit knowledge. This knowledge becomes activated when a person begins to use the tool: the violinist picks up the violin and bow, the lathe operator stands at the machine, the doctor takes an instrument in hand, and so on. The musculature and nervous system of the hand adapt to the regular and purposeful use of these tools, and based on this adaptation, the person integrates them into their own body schema. This is how this subtype of tacit knowledge arises.

### 3.1.2 Material-Centred Tacit Knowledge

Just as tacit knowledge is an essential part of the experience required to use tools, it is also necessary for understanding and mastering the behaviour of materials. This, too can be acquired only through practice. For example, a sculptor gradually learns the physical properties of the chosen material and how it can be shaped most effectively.

### 3.1.3 Environment-Centred Tacit Knowledge

The tacit component of our experience in any given environment is indispensable for carrying out quick and effective actions. A mountain guide can spare an inexperienced tourist many difficulties thanks to detailed knowledge of the terrain. This type of tacit knowledge is more complex than the previous two because of the diversity of environments.

### 3.1.4 Institution-Centred Tacit Knowledge

This subtype of tacit knowledge is acquired through working within one or more institutions. As a subordinate component, it may include elements belonging to the other four subtypes. Research in this area goes back to the topic of "organizational knowledge," which is essential for the successful operation of companies. It was then recognized that part of organizational

knowledge consists of non-verbalizable elements – that is, tacit knowledge.

### 3.1.5 Person-Centred Tacit Knowledge

Finally, the most difficult and most complex subtype of tacit knowledge is the one acquired through regular cooperation with another person. The state of cooperation in which two people "understand each other even from half a word" indicates the presence of significant personal and shared person-centred tacit knowledge.

These subtypes of tacit knowledge also differ in that, in the order listed above, they become progressively less accessible to artificial intelligence and robotics. It is now common practice to develop robots by modelling human bodily movements for them. In this way, part of tool-centred tacit knowledge can be externalized, encoded, and placed into a robot's memory. But even this has limits: for example, if a movement lacks its connection to other types of movements – connections maintained by other forms of tacit knowledge – then no well-usable movement experience is created.

## 3.2 The Temporal Structure of Tacit Knowledge

It is worth complementing the description of tacit knowledge with a third perspective. This arises from the fact that tacit knowledge develops in a time flow just like explicit knowledge. However, unlike the latter, whose final form does not reflect the phases of temporal development, in the structure of tacit knowledge individual development events leave a lasting mark. We know even less about this structure than about the segmentability and reorganizability of explicit knowledge. However, a few things have already become inferable from important learning examples. One of these examples concerns the fact that the first stage in the development of a certain type of tacit knowledge can become very important. I take this example from the history of tool-centered tacit knowledge.

This is a story from antiquity, whose hero is a teacher named Antigenides, who in the 4th century BC was an excellent master of a wind instrument called the aulos in Thebes, where a famous music school with a long tradition operated. Antigenides had the habit of charging double teaching fees from any new student who had started learning this instrument with

another teacher. The reason for this was that it took much more effort to eliminate the bad habits already acquired than to teach students starting with a clean slate. This realization, in terms of tacit knowledge, means that since we cannot consciously access these previously acquired faulty habits, it takes longer for the teacher to develop other, parallel but correct skills in us.

From this example drawn from the history of music, my main conclusion regarding the structure of tacit knowledge is that newly acquired knowledge is placed alongside the problematic knowledge first learnt, and cannot automatically overwrite or invalidate it. In the realm of explicit knowledge, however, once we have recognised and acknowledged an error, we largely free ourselves from it, and it may fade into oblivion. No such mechanism of recognition and forgetting operates in the case of tacit knowledge. Earlier knowledge is rather bracketed – over time – and may later re-emerge occasionally and assert itself again as a bad habit.

The second conclusion is that, in this way, an internal knowledge conflict develops within the domain of tacit knowledge, which may have further consequences. One such consequence is that the knowledge conflict continues to evolve and eventually parts of the two differing bodies of knowledge become explicit, that is, they become externalised. This transformation is not complete; it affects only certain opposing elements of the conflicting knowledge contents. At this point, the pupil consciously understands and can articulate one essential difference between the way he used to handle their instrument and the way he learnt to do so under the new, better master.

The third conclusion, building on the previous two, is that extended learning processes over time play a significant role in determining what structural form this type of tacit knowledge takes. The dual structure outlined above carries fewer advantages than a knowledge structure built upon one dominant, initial, high-quality experience.

By means of Antigenides' example, I have sought to illuminate two aspects of the structure of tacit knowledge. Certain features of this structure are not present to the same degree in all five types of tacit knowledge. Thus, differences may arise in this respect between tool-centred, material-centred, environment-centred, institution-centred and person-centred knowledge. At the same time, a similar structural characteristic appears in person-centred tacit knowledge, as suggested by

the old saying formulated in connection with getting to know new people: "the first impression is very important", because later impressions cannot easily overwrite it. The fact that the two examples can be linked also means that the structure of tacit knowledge functions as a classifier, that is, as a criterion whose varying presence can also be observed among the aforementioned types of this knowledge.

In connection with the structural characteristic of tacit knowledge just described, one structural property of explicit knowledge may also be identified: the body of knowledge acquired earlier does not form a lasting, relatively closed unit, but can be re-organised as appropriate according to explicit rules. Another property is that knowledge conflicts within this type of knowledge can be more easily exposed and therefore resolved more rapidly.

### 3.3 Network Knowledge

Compared with tacit knowledge, the concept of network knowledge is less well known. To introduce and define it, a very brief excursion into the history of science seems necessary. Many people are familiar with the expression "how many handshakes away someone is from another person", perhaps from a very famous individual. The "number of handshakes" refers to the number of acquaintances willing to act as intermediaries who, if asked, could connect the two people. The expression became established after Stanley Milgram's classic American sociological experiment known as the "small world" experiment. In this experiment, participants living in various parts of the United States were instructed to establish contact with unknown individuals living elsewhere solely through acquaintances and acquaintances of acquaintances. The concrete aim was for the participants to deliver a letter to the target individuals. The acquaintances first approached then sought further acquaintances who eventually – creating a particular chain of connections – reached one of the target person's acquaintances, who could then deliver the letter to the addressee. The task was thus solved collectively. The acquaintances formed a network between the seekers and the sought. The surprising average result was that 5–6 steps were sufficient for completion, that is, for the letter to reach the addressees from the participants entrusted with the task. This important experiment led to a deeper understanding of how social networks operate.

The next step in research laid the foundations for distinguishing a new type of knowledge: knowledge and opinion flowed through these networks of connections. This type of knowledge is network knowledge, a small part of which may become public, yet remains inaccessible to AI because it generally does not take an online form. Moreover, the larger part consists of personal and collective tacit knowledge, and as such is closed off from AI. At this crucial point, the other types of tacit knowledge connect with the tacit knowledge present within network knowledge.

Since smaller parts of network knowledge may reach certain points of the public sphere, its degree of AI-visibility is somewhat higher than that of individual or personal tacit knowledge. But overall, it remains very low.

## 4. The First Knowledge Asymmetry of AI

The first knowledge asymmetry of AI arises from the fact that, while AI can extract many kinds of results from the vast – though far from infinite – sets of explicit knowledge available to it, the bodies of tacit knowledge and network knowledge remain without such radical unification and use. With the strong development of AI and the rapid expansion of its use, the low level and limitations of the conscious utilisation of these other types of knowledge become even more striking. As a consequence, social positions are lost in comparison with AI, which supports centralising processes. In a broader sense, regionality and the relative independence of the subsystems of society may be endangered because the higher efficiency of knowledge utilisation made possible by AI tends to transform certain areas of production and social development in the direction of increased centralisation. It must not be overlooked that the emergence of AI itself was the result of enormous concentrations of capital, knowledge, technology and policy expertise.

The use of AI without a solid legal framework and appropriate social constraints fuels the danger that the problem-solving capacities of political and economic centres will once again be overestimated. Thus, “in the name of surplus knowledge”, the much more reliable and up-to-date tacit knowledge bases at local and regional levels, as well as the relative autonomy of the people working at these levels, are pushed into the background.

One of the main problems brought to the fore by the use of AI is the changing social status of tacit knowledge bases and network knowledge.

This change makes it necessary to examine regularly and thoroughly the social, communicative and interest-representation status of knowledge holders.

Let us consider a few sociological examples of the advantage AI provides to various centres. The surplus of knowledge spontaneously present in these centres and their simultaneous awareness of multiple aspects also give them an advantage in using AI. On the basis of statistical data and other available bodies of knowledge, AI can, for example, be used to outline an action plan for a small, medium or large region, while the tacit knowledge of local people remains outside the decision-making process or is pushed far into the background. Yet this knowledge is more up-to-date than the patchwork of facts from different periods on which the action plan is based. How might this contradiction be addressed? Materials produced by AI may serve rather as starting points for debate, being mainly useful in indicating which problems local communities will still need to address, alongside their most pressing concerns, on the basis of past trends.

The accumulation and learning processes taking place within AI may also lead to imbalances, since the strong sequences of questions posed by the centres force learning processes that tip the scales towards them and the fulfilment of their interests. Observing this, decision-makers working in the centres may fail to listen to the key interest groups of local stakeholders and ignore their current interests. Here, the introduction of constraints limiting the unrestricted assertion of the centres' interests, together with support for local communities and their leaders, may offer a solution.

## 5. Summary

In this study I briefly discuss three main topics in connection with one another. The phenomenon of anthropomorphisation draws attention to the fact that many people overestimate the capabilities of artificial intelligence (AI). Among the reasons for this, considerable weight lies in the insufficient awareness of the fact that AI has little or no access to several types of knowledge. Therefore, in the second part, I present some key concepts of AI's knowledge management. I then provide a concise summary of those types of knowledge – tacit knowledge and its subtypes, as well as network knowledge – that AI either cannot access at all or only to a very limited extent. In the fourth part, I examine the first knowledge asymmetry of AI, whose primary source is that this new tool of humanity

radically renews and extends the use of certain types of knowledge available to it. Meanwhile, there is currently no prospect of exploiting to a similar extent the other types of knowledge that are not, or barely, accessible to AI. From this, new inequalities and problems arise in society, the most significant of which appears to be that the resources of centres with various functions, and their impact on society, may further intensify, raising questions also in relation to democracy.

## References

Benczúr A., Gyimóthy T., Szegedy Balázs (2025) Mesterséges intelligencia témájú kutatások Magyarországon [Research on Artificial Intelligence in Hungary]. Magyar Tudomány 186/4 663-675. https://doi.org/10.1556/2065.186.2025.4.7

Bernáth L. (2025) Miért nem rendelkezik felelősséggel a mesterséges intelligencia? [Why Does Artificial Intelligence Not Have Responsibility?] Magyar Tudomány 186/8. 1574-1581. https://doi.org/10.1556/2065.186.2025.8.11

Hoffman M. (2025) Mesterséges intelligencia-alapú tudományos felfedezések – kreativitás, szelekció és heurisztika. Magyar Tudomány 186/8. 1589-1595. https://doi.org/10.1556/2065.186.2025.8.13

Polanyi, Michael (1958) Personal Knowledge. London Routledge & Kegan Paul.

Polanyi, Michael (1966) The Tacit Dimension. Chicago: The University of Chicago Press.

Szívós M. (2005) A személyes és a hallgatólagos tudás elmélete. Előzmények, keletkezéstörténet és új távlatok [The Theory of the Personal and the Tacit Knowledge. Antecedents, Origin and New Perspectives] Budapest, Nemzeti Tankönykiadó – Universitas.

Szívós M. (2017) Fordulópontok a hallgatólagos tudás és a tudattalan felfedezés-történetében. A hallgatólagos tudás általános elmélete [Turning-points in the Discovery of Tacit Knowledge and Subconsciousness. The General Theory of Tacit Knowledge] Budapest, MSZT-Loisir.

Szívós M. (2017) A hallgatólagos tudás négy alaptípusának szerepe az innováció formáiban, valamint az innováció tudásgazdálkodási formájának meghatározása. [The role of the four basic types of tacit knowledge in the forms of innovation, and the definition of the knowledge-management form of innovation.] In: Noszkay Erzsébet

(szerk.) (2017) Közösségi tudások – tudásközösségek: Tudásmenedzsment elméleti és módszertani megközelítésben. Az MTA Gazdaságtudományi Bizottság Tudásmenedzsment Munkabizottságának III. gyűjteményes kötete 2012-2016. Budapest: N & B Kiadó. 192-199.

https://doi.org/10.60030/ALMAMS.c202501OAtH5E

# Extended PALWONS Diagnosis of Systemic Disorders of Organisations – Immune Diseases and an Approach to Their Treatment Using AI Assistance

**Prof. Dr. habil. Erzsébet Noszkay**

professor emerita
Budapest Metropolitan University
nomenb@t-online.hu

**Andrea Drobny–Burján**

PhD-student
University of Sopron
andrea.burjan@gmail.com

## Abstract

The appearance of artificial intelligence has given rise to revolutionary changes in the world around us thereby boosting the unpredictability of the ever-changing world. In this volatile environment the capability enab-ling us to forecast expectable changes and risk factors affecting operation is gaining increasing significance. This issue is examined from the aspect of small and medium-sized enterprises (SMEs). We intended to explore whether the involvement of artificial intelligence provides SMEs with a cheaply and simply available tool for predicting external risk factors? How could artificial intelligence help in forecasting future challenges and being prepared for them? Could the development of strategic thinking of SMEs be helped by the predictability of external risk factors? Would the involve-ment of artificial intelligence be sufficient for that, or would a method-logy be necessary for facilitating the conscious utilisation of this tool?

***Keywords*:** *Artificial intelligence, Large Language Models, external risk factors*

## Introduction

In the near future artificial intelligence and digitisation will exercise their impact on all areas of consultancy. In fact, digitisation and artificial intelligence will within the foreseeable future exercise such revolutionary impacts on employees as well as on various economic actors and organisations, which will compel changes in the professional consultancy domain. This perception incited us to start researches in cooperation with an artificial intelligence expert. Although it may be too early, we believe that the results of our experiments might rise interest and involve further considerations.

For a long time now, external risks form part of the corporate diagnosis (Noszkay, 2009), but the COVID epidemic shocked the focus of diagnostic thinking and warned about the appearance of increasingly dangerous external factors. (Noszkay, 2021) We thought that we would focus on a prominently actual problem for which the society is not sufficiently prepared, namely on the possibility of preliminarily mapping phenomena following each other overly quickly (changing climatic circumstances, natural catastrophes, recurring financial crises, epidemics, wars, etc. able to shake the global financial situation), which exercise increasingly serious impact on humans, organisations, societies and economies. And we approach this examination and the search for solution through the adequate area of business administration and management consultancy, i.e. by using the philosophy and problem solving set of means offered by the Turnaround management consultancy. However, in view of the fact that these phenomena appear not in a customary manner, we are not going to follow the track of traditional thinking patterns, but will rather cooperate with the most modern set of tools or our age, the artificial intelligence. Our studies focus on that feature of large language models that they are able to recognise patterns and early tokens that they were trained on. They are able to discover interrelationships that were hidden previously. Artificial intelligence enables the involvement of new aspects and the compilation of objective analyses best fitting the given branch of industry.

In view of the fact that the revolutionary consequences of artificial intelligence and digitisation furthermore the mechanisms of changes entailed by them cannot be forecasted in full detail, there are lots of

uncertainties, doubts and dilemmas. It is obvious, however, that we cannot manage these changes with traditional solutions, because not only new technologies are needed, but new jobs, new knowledge demands and totally different organisational solutions and production organisation methods are necessary. Our consultancy services must timely be prepared for that. Thus our study describes how to involve artificial intelligence into a promising new approach to and an experiment for mapping and monitoring external risk factors that are of relevance for enterprises. Our study intends to discover not only theoretic considerations but to present the test case of our incipient experiment.

Inclusion of artificial intelligence into risk management, introduces a new era of efficiency and effectiveness. (Yazdi et al. 2024.) Appearance of large language models makes available a tool that analyses enormous information bases and is thereby able to recognise early tokens that serve as unique help in the preparations for future challenges.

In general it can be said that the new technology offers several solutions for economic enterprises for forecasting external risk factors by way of utilising dedicated artificial intelligence solutions adjusted to the specific features of the given firm, of the given branch of industry. The bottleneck in the project might be created by the expertise necessary for running dedicated forecasting systems, furthermore the information and the costs necessary for a feasible technology.

In the heart of our study we focus not only on the possibilities offered by artificial intelligence but we are searching for a solution that ensures for SMEs cost-efficiently and without specialised expertise a possibility for forecasting external risk factors, with the involvement of artificial intelligence. For some time now we deal with the strategic thinking necessary for the efficient application of artificial intelligence, and with a complex approach beyond simple prompts, as well as with the modelling of the same. (Drobny-Burján, 2025) We assess the ways how could the simplest available artificial intelligence solution or any of the large language models be involved in the mapping of challenges expected in the future. What kind of thinking framework could we elaborate in order that an easily applicable methodology would be available for SMEs for forecasting risk factors?

Involvement of large language models enables us to map external risk factors impacting a given SME, let them be political, economic, societal,

technological, environmental, legal etc. factors. However, in order that we would arrive at a really firm-specific analysis, we should perform an analysis goes beyond simple prompts, which takes into consideration the individual features and branch-specific factors of an SME, and is able to take several steps to summarize, weight and evaluate future challenges that are of relevance. In order to forecast future risks factors, a dialogue is needed with the large language model, in which a possibility is given for tracking these specific features. In cooperation with the large language model we build a risk forecasting information field that assists in predicting future challenges, from several aspects.

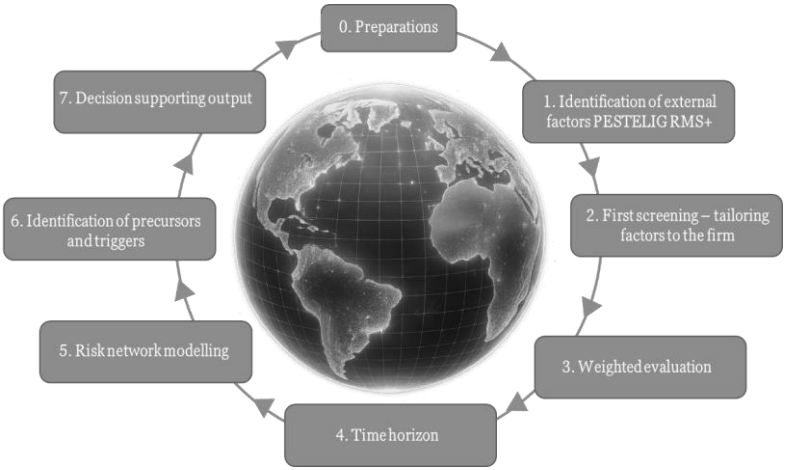



**Figure 1: Steps of the thinking framework that, in cooperation with large language models, serves for forecasting external risk factors faced by SMEs,**

In the following part of this study we are going to summarise the details of the dedicated thinking framework and the experiences gained from the process.

Our set target, i.e. the application of an artificial intelligence solution most simply available for SMEs, unambiguously suggests that in order to forecast external risk factors we should involve a large language model in the process, which is accessible either for free or for monthly subscription fee. In the course of our studies we used subscription ChatGPT Plus (models 5 and 5.2).

## 1. Main characteristics of a thinking framework serving for forecasting external risk factors:

Man principles relating to a thinking framework elaborated with the involvement of artificial intelligence, to be used by SMEs for forecasting external risk factors, are the following:

- The target set for the thinking framework is not that in cooperation with the artificial intelligence it would foretell what will happen, but it should assist an SME in the timely recognition and understanding of the relevant external risks and thus make them manageable. SMEs are not able to prepare for every risk, but if in cooperation with artificial intelligence they could timely recognise the signs of changes, this gives them competitive advantage.
- The model concentrates exclusively on external factors (macro, market, regulatory, technological, environments, societal), but it always examines where and how would the same impact a given enterprise and what solution possibilities are available.
- Key features of the thinking framework are the screening of the external risk factors; identification of the many external factors and the selection of those that from among them are relevant for the enterprise; identification, weighting and mapping risk factors that on the short run are most critical for the given SME.
- Within the model we examine interrelationships among external risk factors, whether or not there are factors that intensify each other's impact and even could accelerate a specific process that is of importance for the given SME, too.
- At the end of the process a summary analysis is elaborated about the external risk factors, which assists decision makers in preparing for future challenges.
- An important feature of the thinking framework is that it can be run without any special expertise. Mapping of the external risk factors can be repeated if necessary, which enables tracking of temporal changes in external risk factors forecasted by artificial intelligence. Thus a new tool is available for improving strategic thinking of SMEs. A seven-step prompt chain (thinking framework) should importantly be highlighted, which by way of exceeding a simple prompt, enables SMEs to elaborate a complex map of future challenges that are of relevance for them, and to manage the same.

## 2. Steps of a thinking framework serving for forecasting external risk factors:

The essence of a thinking framework is a preliminarily compiled prompt-chain that identifies expected external risk factors on the basis of the specific features of the SMEs. On the basis of the prompt-chain a weighted list is prepared showing the relevant risk factors, which is an evaluation in the form of a management summary, an analysis of the most important external risk factors and triggers that are of relevance for the SME, as well as a strategic proposal offering solution for the management of risk factors. Let us have a look on the steps of the prompt-chain, as follows:

### Step 0: Preparations:

The Preparation phase is made of two steps:

1) External risk factors can be retrieved in "agent mode", because "agent mode" of ChatGPT is able for the step-by-step execution of a given prompt-chain. The first step should be the elaboration of the Meta prompt that includes the most important aspects of the query, the prompt-chain, in agent mode. Its aim is to provide the large language model with the operative framework and the bans, in order that the prompt-chain compiled for forecasting the external risk factors could run appropriately.
   Meta prompt is a guiding framework that defines the way how the large language model should think and make decision before responding to a test.
   It does not define the content but rather the rules for thinking, managing uncertainties and undertaking responsibility, thus ensuring consistent and reliable operation.
2) Preparatory phase includes the completion of the EWS (Early Warning System) Profile datasheet (Figure 2) serving for the identification of the SME. Its purpose is to give those most important information that enable the large language model to identify the relevant risk factors. The input must be a collection of the identifiers of the firm, specific to its area of operation, the characteristics of its activity and of its branch of industry, i.e. all information that helps in understanding the operation of the firm, and in identifying unique features necessary for forecasting external risk factors.

When prompting ChatGPT, we must start forecasting external risk factors, with the completion of the Meta prompt and the EWS Profile datasheet.

## EARLY WARNING SYSTEM (EWS) PROFILE DATASHEET

### Firm and industrial branch identification

This datasheet serves for the quick identification of the environment of the enterprise and the industrial branch. It is aimed at giving structured input for the AI-based Early Warning System for mapping risk factors. This datasheet is not an analysis but rather a context and profile defining tool.

**1. Basic data for identifying the firm (compulsory)**

- Official name of the firm:
- Short name / Brand name:
- Official website:
- Official e-mail domains:
- Country of registration (ISO):
- Country of operation: (ISO, multiple choice)

**2. Branch of industry / regulatory environment**

- Primary branch code (TEAOR/NACE/ NAICS):
- Secondary branch code(s) if relevant:
- Name of the branch (text):
- Main regulatory framework: (e.g., GDPR, ISO, MDR, other)
- Regulatory exposure: (Low/Medium/ High)

**3. Business model**

- Main products / groups of services:
- Role played in the supply chain: (Producer/Distributor/Service provider/Mixed)
- Customer types: (B2B/B2C/Mixed)
- Main revenue resource(s):
- Combined proportion of 3 top customers (% - if known):

**4. Market and geographical exposure**

- Main revenue producing markets (country / region):
- Rate of export (if relevant):
- Currency exposure (main currencies, estimated proportions):
- Macroeconomic exposure: (e.g., energy, inflation, interest rate)

**5. Supply and operational dependencies**

- Main supply countries/regions:
- Critical base materials/inputs:
- Energy intensity: (Low/Medium/High)
- Technology exposure (critical systems/technologies):

**6. Branch dynamics and external trends**

- Branch characteristics: (Stable/Cyclical/Quickly changing)
- Growth rate of the branch: (Decreasing/Stagnating/Increasing)
- Main competitive dimension: (Price/ Quality/Innovation)
- 1 to 3 determining branch or market specific trend(s):

**7. Presumptions critical from EWS aspect**

- Stability of which external factors are of vital importance for the operation?
- Which external changes might cause quick business shock?
- What type of events would imply largest external risk?

**8. Other enterprise specific factors (critical from EWS aspect)**

- Is there any external or internal factor that is not mentioned in the points above but that is of critical importance from the aspect of the firm's operation?
- Is there any exposure, sensibility or "tacit assumption" whose deterioration would entail serious business risk?
- Is there any factor which is expressly monitored by the management but which can hardly be measured or described formally?

**Figure 2: Thinking framework input datasheet serving for forecasting external risk factors impacting SMEs**

**Step 1: Identification of external factors with PESTELLIG RMS+ model:**

On the basis of the datasheet completed in Step 0, with the involvement of the large language model a relevant but not overly big risk universe is defined, which contains those risk factors that could appear in the course of the operation of the SMEs examined.

For the collection of the risk factors we have elaborated a new and extended model named PESTELIG RMS+ that is based on the PESTEL model (Yüksel, 2012) but widens the scope of the possible factors. In order to model the challenges created by changing circumstances, the political, economic, societal, technological, environmental and legal aspects are supplemented with specific features of the branch of industry, system-level shock effects, reputation risk, challenges stemming from the changes in the market and in consumer demands, and uncertainty factors stemming from the fragility of the supply chain. A model so extended ensures wider horizon for large language models for evaluating risk factors. A possibility is given for approaching the topic from more aspects and thus for the elaboration of a better grounded forecast. In fact, today the challenges originating from a quickly changing world cannot be evaluated comprehensively if for instance industrial features, shock effects on the system level, reputation risk influencing brand reputation would be neglected. Risk in changes of customer demands and inherent in the operation of the supply chain may not be neglected either. Challenges arising in the contemporary world justify the supplementation of the aspects in the PESTEL model with new criteria, as follows (see Figure 3).

Step 1, with the involvement of the PESTELIG RMS+ model, produces a rough list that may include as much as 15 to 30 risk factors, which is overly redundant from the aspect of further in merit processing. They include rather general factors and more firm-specific aspects, thus further processing is needed in order that these could form a basis for a material supporting managerial decisions, which would be of assistance for SMEs.

| | Sphere | Aspects | Supplementary summary |
|---|---|---|---|
| P | **Political** | Governmental stability; Sanctions; barriers to trade; Geopolitical tensions; Global power shift; Friendshoring; Deglobalisation | Quick changes in the political environment exercise impacts on the markets, supply chains and the regulation. |
| E | **Economic** | GDP, inflation, unemployment, currency exposure, interest environment; Raw material and energy prices; Digital economy, cryptocurrencies; Impact of automatization on the labour market | Economic fluctuation and financial innovations simultaneously create new possibilities and risks |
| S | **Societal** | Demographic trends; Migration, urbanisation; Changes in consumer habits; Generation gaps; Climate migration; Societal polarisation | Societal changes determine labour market, customer habits and long-term stability |
| R | **Technological** | Digitisation, cyber security; Automatization, robotization, IoT; Artificial intelligence bias; Quantum calculation; Data management, data protection | Fast pace of technological development entails competitive advantages and new risks simultaneously, specifically in the AI and quantum areas |
| E | **Environmental** | Climate changes; Extraordinary weather conditions; Scarcity of natural resources; ESG regulations; Water and soil crisis | Environmental factors produce strategic hazards on the long run, meanwhile supply and cost shocks on the short run. |
| L | **Legal** | GDPR, MDR, ISO; Intellectual property, competition law; AI Act and data sharing acts; Biotechnological regulations | The legal environment continuously changes specifically in the areas of technology and sustainability, which impact operation intensively |
| I | **Industry specific** | Sectoral regulation (e.g. pharmaceutical industry); Supplier dependence; Quality assurance rules; Consolidation; Takeovers | Industrial specificities determine the basic operational conditions and impact flexibility. |
| G | **Global shocks** | Black Swan events (e.g. COVID, 9/11); Grey Rhino risks (e.g. climate change); Pandemics, global cyber incidents; Crises of the financial system | Preparedness for rare but powerful shocks is of critical importance since it could start chain reaction in the economy and the society. |
| R | **Reputation/Media** | Brand image, reputation; Social media impacts; Fake news, disinformation; Deepfake hazards; Online boycott waves | Digital media accelerates the spreading of reputation risk and can cause substantial losses within a short time. |
| M | **Market / Customer** | Changes in customer demands; Competitors' actions; Sustainability expectations; Digital customer experience | Market and customer expectations change quickly and exercise direct impact on sales revenues and innovation demands. |
| S | **Supply Chain / Partners** | Supplier reliability; Logistic capacities; Friendshoring, reshoring; Cyber exposure through partners; Critical raw materials | Fragility of the supply chain is global and is a branch specific risks that is intensified by new geopolitical and technology trends. |

**Figure 3: PESTELIG RMS+ model**

**Step 2: Summarisation of risk factors of relevance for a given SME:**

Step 2 is aimed at the production of a manageable, transparent structure out of the external risk factors already identified. This step decreases noise, screens overlaps and organises risks of similar origin into logic clusters, without performing any evaluation or ranking. Its importance is inherent in the fact that it prevents overloading as well as any premature, spontaneous decision making. Its unique feature is that instead of trimming anything, it creates sensible patterns that ensure stabile ground for evaluations later. At this step the quantitative listing is replaced by qualitative evaluation. It is not aimed at reducing the number of risks, but rather at clarifying which factors are various manifestations of one and the same factor in the background. It helps to avoid that decision makers would react to a single problem as to several and seemingly different risks. Its uniqueness lies in the fact that already prior to the evaluation it is able to "think systematically" instead of managing risks in insulation.

In Step 2 the large language model acts primarily as a cognitive assistant. On the basis of meaning it recognises overlaps between differently described risk factors, unifies their abstraction levels and organises them into logic clusters alongside reasons in the background and effect-chains. This structuring decreases redundancy and noise, and at the same time saves information contents and helps in avoiding premature and spontaneous prioritisation. Impartial operation of the model contributes to the mitigation of cognitive distortions since it is not interested in organisational concerns or in justifying decisions made in the past. Meanwhile short and retraceable justifications are generated thus the risk structure becomes not only transparent but also explainable which can be defended on managerial level, too. The systematised risk-list generated as a result of step two, will be the input to the next step.

**Step 3: Weighted evaluation**

In step 3 the structured risks become worthy of comparison with each other with the help of a five-dimensional (Impact, Likelihood, Velocity, Detectability, Exposure) evaluation framework. The weighted scores do not conceive of a decision but rather create a ground for prioritisation and meanwhile make uncertainties visible. This step is of critical importance because it separates analyses from decisions and thereby reduces the hazard of fake accuracy and posterior rationalisation. Its uniqueness lies in the fact that weighting appears not as a static truth but rather as a conscious

managerial focus that could be reviewed from time to time. Step three clarifies the proportions represented by each risk, meanwhile consciously avoids judgments with non-appealable effect. The five-dimensional evaluation enables that the representation of a risk would illustrate not only the size of the expectable effect but also its velocity, transparency and the impact that the organisation is exposed to. This is of specific importance in the early warning systems where not the largest but the fastest developing risks deserve attention. The real value of this step lies in the fact that it creates a common language for analyser's logic and managerial intuition. Aspects of weighted evaluation are on Figure 4.

| Factor | Proposed weight | Explanation |
|---|---|---|
| Impact | 0,3 | This is the most important factor since one must by all means be prepared for risks with large impact. |
| Likelihood | 0,25 | Classic dimension: if something could occur frequently, it deserves extraordinary attention. |
| Velocity | 0,2 | Risks emerging quickly (e.g. exchange rate shock) might entail severer consequences. |
| Detectability | 0,15 | A hardly detectable risk is more dangerous, and scoring must reflect this. |
| Exposure | 0,10 | A given risk could menace different firms differently (e.g. currency exposure, market exposure) |

**Figure 4: Weighted evaluation of external risk factors**

| Dimension | Weight | Professional base |
|---|---|---|
| Impact | 30% | ISO 31000, COSO |
| Likelihood | 25% | ISO 31000, COSO |
| Velocity | 20% | HBR, Basel, crisis management |
| Detectability | 15% | FMEA (IEC 60812) |
| Exposure | 10% | COSO ERM, OECD |

**Figure 5: Professional base for weighting external risk factors**

Weighted risk evaluation is based on the customary practice of multi-criteria decision supporting and risk management, where the weights express not objective measuring but rather the transparent and reviewable managerial preferences. See the following table (Figure 5).

Calculation of weighted scores is a classic approach to decision supporting where weights express preferences and not objective facts. Weighting is a standard starting model based on the framework described in the literature, which could be modified by the management on the basis of enterprise-specific preferences. The greatest advantage offered by the possibility of modifying weight scores by the management is that the risk evaluation would in fact reflect the strategic reality of the enterprise and not a general scheme. Therefore the model will remain structured and meanwhile firm-specific, which improves managerial acknowledgment and the quality of decisions. Dependently of the branch of industry, in due consideration of its specific features, the weight scores might differ from each other. For instance, within the finance sector a bank or an insurer assigns larger weight to the Likelihood dimension than to Detectability, because frequent and hardly detectable risks (e.g. liquidity tensions) could entail severe consequences even in the case of slight differences. In contrast to that, an export oriented enterprise may increase the weight of Exposure and Velocity, because currency or supply chain shocks quickly and directly impact profitability.

In summary, the possibility of fine tuning weights ensures that the risk score will not be an abstract number but rather a decision supporting strategic mean formed consciously by the management. Starting values can be changed in the prompt to be fed to ChatGPT. It is suggested that in the first step the starting data are evaluated on the basis of industrial branch-specific features, and the weighting values are later annually reviewed in due consideration of changes.

**Step 4: Time horizon assigned to risk factors**

Step four, the assignment of the time horizon, enable us to handle risks already evaluated not as “timeless” menaces but rather to localise them within the time dimension of expected occurrences. This step is an answer to the managerial question concerning the time when a given risk deserves in merit managing: it may represent immediate (0 to 12 months), medium (1 to 3 years or long term (3 to 5+ years) effect. Assignment of the time horizon helps in separating risks demanding operative “wildfire fight”

from those which demand rather strategic attention and continuous monitoring, and thus reduces the risks inherent in overreaction and understatement. The uniqueness of this step lies in the fact that it handles time not as a static label but rather as a dynamic factor. Imputably to external events, the probability and the impacts of risks may change as time goes by, either acceleration or slowing down could be experienced. Supplementation of the time horizon with a short narrative description – which reveals the "life style" of the risks and the possible turning points – constitutes direct transition towards the identification of trigger points and the precursors. Thus this step is not an independent element but rather a contact point of key importance between the evaluation and the active early warning system (EWS), which supports managerial prioritisation and timely decision making.

The primary roles played by the large language model within this step are interpretation and pattern recognition. On the basis of risks identified and scored earlier, it is able to come to a conclusion in a structured manner concerning the time span within which a given risk could become relevant, in due consideration of the interrelationships among Impact, Velocity and external environmental context. The advantage ensured by the large language model is its capability of simultaneously integrating several sources, trends and narratives and thus it can not only assign a label to the time horizon but gives a justified and coherent explanation concerning the reason behind the time span of the risk, its short, medium or long term. In addition, the large language model assists in making temporal dynamics to be explicit. It assists in describing how would the probability of a risk change as time goes by, what external events would accelerate or postpone occurrences and where could the turning points be found. This is extremely important from the aspect of supporting managerial decision making because the large language model produces not the ultimate truth but rather an adaptive time frame that, in consideration of new information, precursors or trigger events, can be reviewed later. Thus the model does not make decisions instead of a manager but enforces structured temporal thinking exactly at the point where human intuition is often uncertain.

**Step 5: Risk network modelling**

In reality external risks rarely appear in isolation; they evolve alongside effect-chains strengthening each other and are in casual relationship. This is why step five, identification of risk networks and chains, is of extreme importance. The essence of this step is that the risks identified and

evaluated earlier would be examined not as independent points but within a scheme of cause-and-effect relationships (e.g. A → B → C). This approach helps in recognising domino effects, primer exciting factors and those "nodal" risks whose management could exercise disproportionally large impact to the entire system. From a managerial aspect this is the step that creates ground for real strategic prioritisation. Not the noisiest gets the most attention but the one that affects most of the risks.

In this step the large language model works as an analyser that supports system level thinking. It is able to recognise patterns among various risk factors and to suggest probable risk chains in structured form, meanwhile gives justifications concerning the logic of each relationship. An advantage of artificial intelligence is that it is able to simultaneously consider economic, societal, technological and regulative narratives, thus it is able to clarify interrelationships that in the course of human analyses could easily remain hidden. Meanwhile it is an important fact that the model does not "deliver the truth" but rather formulates hypotheses. It sheds light on alternative chains, on weak and strong connections that after a critical managerial review could ground decisions. Thus artificial intelligence does not replace but rather structuralises and accelerates the understanding of complex risk interrelationships.

**Step 6: Identification of precursors and triggers**

The aim of step 6 is the identification of precursors and trigger points, which is one of the most critical elements of the risk analysing process. This is the point where the analysis from passive evaluation turns into a forecasting system. The essence of this step is that the revealed risks are connected to measurable and continuously monitorable indicators that prior to the factual occurrence signalise the strengthening of the risk. Precursors (changes e.g. in economic indices, exchange rates, regulatory measures, market anomalies, media contents) and the trigger thresholds assigned to them enable the management to prepare for crises timely and not to react subsequently. This step ensures possibility for early intervention that often entails costs and risks that are significantly more modest than those of a late reaction. Within this step the large language model plays the role of a sensitive signal interpreter and a synthesiser. It is able to interconnect data sources of different types (news, policy announcements, market narratives, social media trends) and on that basis suggest precursor indicators that are relevant and risk-specific. The artificial intelligence assists not only in the identification of indicators but

also in their interpretation. It explains why would a given change be deemed to be a warning signal and what is the risk trend it signalised by it. In addition, it supports logical identification and continuous review of trigger points, since at the appearance of any new information it is able to review former hypotheses. Thus the artificial intelligence works not as an automated alarming system but rather as an adaptive, explainer and forecasting partner that helps in figuring out when and why would it be worthwhile to start managerial intervention.

**Step 7. Managerial decision supporting summary:**

Step seven of the process, summarisation to support managerial decision, is a process closing and meanwhile the most value-adding element, because this is the point where the complex risk analysis turns into factual, action supporting managerial cognition. The essence of this step is the organisation of risks, time horizons, networked interrelationships and precursors discovered in the former points, into a transparent narrative. Its aim is not the repetition of the analytic details but rather the support managerial focus. What are critical risks, what patterns come to light, and what strategic responses could be considered. This step ensures that the analysis appears for the management as a real decision supporting tool. In this phase the artificial intelligence plays synthesiser and narrative decision supporting role. By way of utilising structured outputs of the former steps, it is able to elaborate a short, easily understandable, meanwhile professionally well-grounded managerial summary, highlighting interrelationships and uncertainties that are really essential. An advantage of the large language model is that it is able to draft alternative strategic reactions, too, and not the only "right solution", but several possible directions together with their expectable consequences. Thus it supports conscious decision making and the taking of managerial responsibility, meanwhile grants possibility for critical reviews. Within this step the model does not make any decision but rather encloses the decision into an understandable framework, thereby helps to create a bridge between analytic thinking and strategic managerial actions. Despite the fact that a comprehensive analysis could be elaborated for forecasting external risk factors, this step does not disregard the personal managerial decision, elaboration and finalisation of strategic steps aimed at preparations for managing risk factors.

## 3. Operation of the thinking framework serving for forecasting external risk factors:

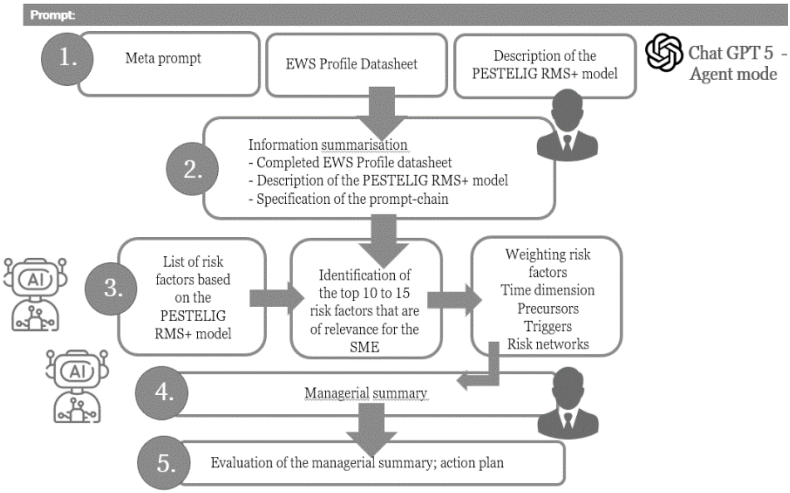



**Figure 6: Process of external risk factor forecasting with the help of artificial intelligence**

## 4. Summary:

Inclusion of artificial intelligence in the process of forecasting external risk factors, enables SMEs to think in a more structured way than earlier. The reason why large language models are extremely apt to support forecasting of external risk factors is that by their very nature they consider patterns, interrelationships and narratives and not isolated data. External risks, e.g. geopolitical tensions, changes in regulations, market mood or technology shifts, appear typically not in a single measure but rather in the form of scattered information, weak signals and stories mutually strengthening each other. From these fragmented information, the large language models are able to create coherent interpretation frameworks, to recognise rudimentary patterns even at points where there are no unambiguous statistical evidences. In addition, the large language models enable dynamic, adaptive thinking, which is of extraordinary importance in an uncertain and quickly changing environment. They operate not in accordance with predetermined rules, but in response to new information they are able to review their former conclusions and to draft alternative explanations and scenarios. This is extremely valuable in the cases of external risks, where the question “What could happen?” is often more important than the question “What will happen for sure?”. Thus the large

language models operate not as oracles, but as structured and clever forecasting partners. They assist in systematising uncertainties, visualising possible shifts and support timely passed conscious managerial decisions.

On the basis of research experiences it can be stated that artificial intelligence is an easily usable tool for SMEs when external risk factors should be forecasted. The tool itself is not enough for a solution; an elaborated thinking framework is necessary, which assists in the summarisation of aspects and steps necessary for forecasting risk factors. A thinking framework ensures possibility for running the process coherently and for repeated queries, and thus an easily and low-cost (monthly subscription) tool becomes available for SMEs, which enables the development of strategic thinking. In addition to forecast external risk factors, this tool assists in the elaboration of strategic suggestions leading to solutions, therefore the heads of SMEs are not left alone to face problems of challenges already forecasted.

Simultaneously with the elaboration of a thinking framework serving for forecasting external risk factors, a logic methodology has been created which can step by step be used in the agent mode of ChatGPT irrespective of the branch of industry. Irrespective of changes in the models, the thinking framework is time-resistant. Preliminary experiences are promising, since through the involvement of artificial intelligence, managerial level strategic decision preparatory materials can be compiled, which for smaller enterprises ensures possibility of future minded, conscious thinking, conscious preparations for future challenges. In summary we can say that the experiences gained so far are promising and grant secure starting basis for further refining this methodology.

**Bibliography:**

Drobny-Burján, A. (2025): Megoldásra Kódolt kreativitás - Hatékony üzleti döntéstámogatás és problémamegoldás a mesterséges intelligencia bevonásával, Bookship [Solution encoded creativity – Efficient support to business decision and problem solving with the involvement of artificial intelligence]

Yüksel, I. (2012): Developing a Multi-Criteria Decision Making Model for PESTEL Analysis - International Journal of Business and Management. https://doi.org/10.5539/ijbm.v7n24p52

Noszkay, E. (2009): Változás- és válságmenedzsment az alapoktól, N&B Kiadó [Change and crisis management from the ground up]

Noszkay, E (2021): Vállalati, vállalkozói válaszok, aktivitások a COVID-19 járvány fenyegető kihívásainak idején- Polgári Szemle, 17. évf. 1-3. szám, 2021, 95-116 [Answers and activities of enterprises, entrepreneurs]
https://polgariszemle.hu/archivum/187-2021-augusztus-17-evfolyam-1-3-szam/188-fenntarthatosag-regionalis-versenykepesseg/1146-vallalati-vallalkozoi-valaszok-aktivitasok-a-covid-19-jarvany-fenyegeto-kihivasainak-idejen
https://doi.org/10.24307/psz.2021.0708

Yazdi, M., Zarei, E., Adumene, S., & Beheshti, A. (2024) Navigating the Power of Artificial Intelligence in Risk Management: A Comparative Analysis, 15 March, 2024
https://www.mdpi.com/2313-576X/10/2/42
https://doi.org/10.3390/safety10020042

TANULMÁNYOK AZ INFORMÁCIÓ ÉS TUDÁSFOLYAMATOKRÓL 22. - ITA 2025; 153-185.

https://doi.org/10.60030/ALMAMS.c202501ZfkvSv

# A hibrid intelligencia a vállalatvezetői tudás új korszakában

Katits Etelka Éva

tudományos dékánhelyettes
Pannon Egyetem
e-mail: katits.etelka@zek.uni-pannon.hu
https://orcid.org/0000-0003-4108-7459

Fejes Judit Katalin

mestertanár
Pannon Egyetem
e-mail: fejes.judit@zek.uni-pannon.hu
https://orcid.org/0009-0000-8310-120X

## Absztrakt

A tanulmány célja annak feltárása, hogy a dual/tripla és az Edge mesterséges intelligencia, valamint az életszakaszokra igazított 4. generációs korai figyelmeztető rendszerek miként támogathatják a vállalati vezetői tudásmenedzsment folyamatokat. A vizsgálat kiemelten kezeli az információgyűjtés, tudásmegosztás és – megőrzés új lehetőségeit, valamint azt, hogy ezek a technológiák hogyan erősítik a stratégiai és operatív döntéshozatalt. A kutatás vegyes módszertant alkalmaz: kvalitatív eszközei között szerepelnek szakértői interjúk, esettanulmányok és Delphi-panel, míg kvantitatív oldalról vállalati adatbázisok és idősorok, statisztikai elemzések és kísérleti vizsgálatok támogatják a modell-alkotást. Az előzetes várakozások szerint a dual/tripla AI rendszerek a humán és gépi intelligencia összehangolásával növelik a tudás-felhasználás hatékonyságát, míg az Edge AI valós idejű feldolgozási képességei gyorsabb és pontosabb döntéstámogatást tesznek lehetővé. A korai figyelmeztető rendszerek életciklus-

alapú beépítése pedig fokozza a vállalati rezilienciát. A kutatás eredményei hozzájárulhatnak egy új, elméletileg megalapozott és empirikusan tesztelt keretrendszer kialakításához, amely tudományos szem-pontból gazdagítja a vezetéstudományt, gyakorlati oldalról pedig útmutatást nyújt a vállalatok számára az AI-technológiák bevezetéséhez és adaptálásához.

***Kulcsszavak:*** *életciklus, hibrid mesterséges intelligencia, korai figyelemeztető rendszer, reziliencia, tudásmenedzsment*

## Abstract

The study aims to explore how dual/triple and Edge Artificial Intelligence (AI), as well as 4th generation early warning systems (EWS) tailored to life stages, can support corporate leadership knowledge management processes. The study focuses on new opportunities for information collection, knowledge sharing and preservation, and how these technologies strengthen strategic and operational decision-making. The research uses a mixed methodology: its qualitative tools include expert interviews and case studies, while on the quantitative side, company databases and time series, statistical analyses and experimental studies support model creation. According to preliminary expectations, dual/triple AI systems increase the efficiency of knowledge use by coordinating human and machine intelligence, while the real-time process-sing capabilities of Edge AI enable faster and more accurate decision support. The lifecycle-based integration of EWS enhances corporate resilience. The results of the research may contribute to the development of a new, theoretically grounded and empirically tested framework that enriches management science from an academic perspective and provides practical guidance for companies in the introduction and adaptation of AI technologies.

***Keywords:*** *early warning system, hybrid artificial intelligence, knowledge management, life cycle, resilience*

## 1. Bevezetés

A tudásmenedzsment a modern vállalatok versenyképességének egyik meghatározó alappillére, mivel a szervezeten belül felhalmozott szakértelem, az információk hatékony integrációja és hasznosítása közvetlen hatást gyakorol az értékteremtésre és a hosszú távú teljesítményre. A tudásalapú vállalatelmélet klasszikus megfogalmazása szerint maga a vállalat olyan tudásintegráló intézmény, amelynek elsődleges funkciója a szervezeti tagok speciális tudásának koordinálása és alkalmazása az üzleti folyamatokban (Grant, 1996).

A 21. század digitális átalakulása során a korszerű technológiák – különösen az adattudomány és a mesterséges intelligencia (AI=Artificial Intelligence) – jelentősen kibővítették a tudásmenedzsment eszköztárát, új lehetőségeket teremtve a tudás feltárására, megosztására és szervezeti szintű hasznosítására. Rezaei (2025) is hangsúlyozza azt, hogy a tudásmenedzsment kritikus szerepet játszik a szervezeti versenyképesség fenn-tartásában, miközben az AI integrációja alapvetően átalakítja a KM-gyakorlatokat és a döntéstámogatási mechanizmusokat. Ehhez a technológiai fejlődéshez szorosan kapcsolódik az edge computing paradigmája, amely lehetővé teszi az adatfeldolgozás decentralizálását és a hálózat peremére helyezését. Shi et al. (2016) szerint ez a megközelítés jelentősen csökkenti a késleltetést, javítja a válaszidőt és növeli a rendszerek robusztusságát, ezáltal hatékonyan támogatja a valós idejű döntés-támogatást és az adatintenzív üzleti alkalmazásokat.

Itt érdemes hangsúlyozni azt, hogy az AI-technológiák önmagukban történő bevezetése nem garantálja a kívánt teljesítményjavulást. Egyet értünk Olan et al. (2022) megállapításával, amely szerint az AI kizárólagos alkalmazása nem elegendő a szervezeti teljesítmény tartós növeléséhez; fenntartható eredmények csak akkor érhetők el, ha az AI szorosan integrálódik a tudásmegosztási és tudásmenedzsment-folyamatokba. Ezzel a fel-ismeréssel összhangban Lim & Hwang (2025) a hibrid intelligenciát olyan kutatási és alkalmazási paradigmának tekintik, amelyben az emberi és az AI szinergikus együttműködése válik a döntéshozatal és a tudásalkalmazás központi elemévé. A hibrid intelligencia az ember+AI együttműködés új tudományos kereteként jelenik meg a jövőorientált döntéstámogatásban, ahol a gépi elemzés sebessége és az emberi ítélőképesség, kontextus-érzékenység egymást erősítve járul hozzá a szervezeti teljesítményhez. Ennek megfelelően a

legújabb kutatások egyre nagyobb hangsúlyt helyeznek az ember-központú, hibrid megoldásokra, amelyekben az emberi szakértelem és a gépi intelligencia kölcsönösen támogatják egymást a komplex üzleti döntések során (Ahdadou et al., 2025).

A pénzügyi kockázatok és szervezeti sebezhetőségek kezelése területén Tanaka et al. (2025) kimutatták azt, hogy a többlépcsős, gépi tanulásra épülő korai figyelmeztető rendszerek (EWS=Early Warning Systems) szignifikánsan növelik a vállalati pénzügyi nehézségek előrejelzésének pontosságát, és lehetővé teszik az időben történő vezetői beavatkozásokat. A gépi tanulásra épülő, többstádiumú EWS-modellek így hatékony eszközt kínálnak a pénzügyi instabilitás korai felismerésére, valamint a szervezeti reziliencia megerősítésére. Ráadásul a vállalatok egyre komplexebb és dinamikusabb működési környezetben kénytelenek helytállni, ahol a gyorsan változó technológiai feltételek – különösen az AI térnyerése – alapvetően új lehetőségeket és egyben jelentős kihívásokat teremtenek a vezetői tudásmenedzsment számára. Mindezek alapján a tartós versenyelőny forrása a tudás hatékony integrációjában és alkalmazásában rejlik (Grant, 1996). Napjainkban a hibrid intelligencia gyorsan fejlődő kutatási területté vált, és egyre meghatározóbb szerepet tölt be a döntéshozatal, az együttműködés és az intelligens automatizálás területén. Az említett folyamatokat tovább erősíti az adatközeli feldolgozásra épülő edge computing paradigma, amely az adatfeldolgozás decentralizálásával és a hálózat peremére helyezésével teremti meg a valós idejű, AI-alapú döntéstámogatás technológiai feltételeit (Shi et al., 2016). A tudásalapú vállalatelmélet és az adatközeli feldolgozás együttesen így szilárd szervezeti és technológiai alapot biztosít az AI-alapú döntés-támogató rendszerek számára (Grant, 1996; Shi et al., 2016).

A hagyományos tudásmenedzsment rendszerek általában túl statikusak és nem tudnak lépést tartani az AI gyors fejlődésével és a digitalizációval. A hagyományos EWS-modellek főként pénzügyi mutatókon alapulnak (pl. Altman Z-score) és késleltetett figyelmeztetéseket adnak, így kevesebb az esély a proaktív válságkezelésre (Fejes & Katits, 2025b). A hibrid intelligencia elméleti kerete és a gépi tanulásra épülő korai figyelmeztető rendszerek együttesen új szintre emelik a vállalatvezetői döntéstámogatást (Lim & Hwang, 2025; Tanaka et al., 2025). Ugyanakkor a vállalati életciklus is fontos keretet ad a válságok előrejelzéséhez, mivel minden cég különböző és ismétlődő életciklus-fázisokon megy keresztül (Adizes, 2023), s az életszakasz-váltások fordulatokkal járnak, amelyek menedzse-

lési megoldásokat igényelnek (Katits, 2024). A magyarázható AI ((XAI=eXplainable AI) és az ESG (Environmental, Social, Governance))-orientált AI-alkalmazások együttesen járulnak hozzá a felelős és fenntartható vállalatvezetői döntéshozatalhoz (Adadi & Berrada, 2018; Song et al., 2025). Az átmeneteknél tipikusak a kockázatok (pl. piacvesztés, finanszírozási gondok), de a hagyományos rendszerek nem integrálják ezt a tudást. A dual/tripla AI rendszerek bevezetése új kérdéseket vet fel a tudásfelhasználásban: Hogyan egészítheti ki a gépi tanulás a vezető tapasztalatát, és miként tudják együtt felerősíteni a tudás cseréjét, rendszerezését és fenntartását? Emellett az Edge AI technológiák robbanásszerű elterjedése lehetőséget ad a valós idejű döntéstámogatásra, de kihívásokat jelent az infrastrukturális fejlesztések terén is. A szervezeti életciklus-elméletek és a XAI együttes alkalmazása lehetővé teszi az életszakasz-specifikus, TAI-alapú döntéstámogatást (Quinn & Cameron, 1983; Lundberg et al., 2020). Tehát összességében hiányzik egy olyan átfogó keretrendszer, amely a hibrid intelligencia különböző formáit és az életciklus-alapú EWS rendszereket ötvözve erősíti a vállalati tudásmenedzsmentet és a szervezeti rezilienciát.

A kutatás elsődleges célja annak feltárása, hogy a dual/tripla AI és Edge AI rendszerek, valamint az Adizes-modellre épülő, életciklus-alapú EWS miként járulhatnak hozzá a vállalatok vezetői tudásmenedzsment-folyamatainak fejlesztéséhez. A vizsgálat középpontjában annak elemzése áll, hogy ezek a technológiai és elméleti megközelítések hogyan támogatják az információgyűjtést, a szervezeti tudás megosztását és megőrzését, valamint miként növelik a stratégiai és operatív döntéshozatal hatékony-ságát, átláthatóságát és megalapozottságát.

A prediktív döntéstámogatás módszertani megalapozásához a kutatás támaszkodik a fejlett idősorelemzési technikákra is. Az LSTM (Long Short-Term Memory)-alapú neurális hálózatok különösen alkalmasak komplex, nemlineáris gazdasági és üzleti folyamatok előrejelzésére, mivel hatékonyan kezelik a hosszú távú függőségeket az idősoros adatokban. Ennek megfelelően az LSTM-modellek kiemelt szerepet töltenek be a vállalati teljesítmény és kockázatok előrejelzésében (Hewamalage et al., 2021).

A kutatás interdiszciplináris szemléletet követ: elméleti szinten hozzájárul a hibrid intelligencia és a tudásmenedzsment kapcsolatának mélyebb megértéséhez, míg gyakorlati oldalról iránymutatást nyújt arra vonat-

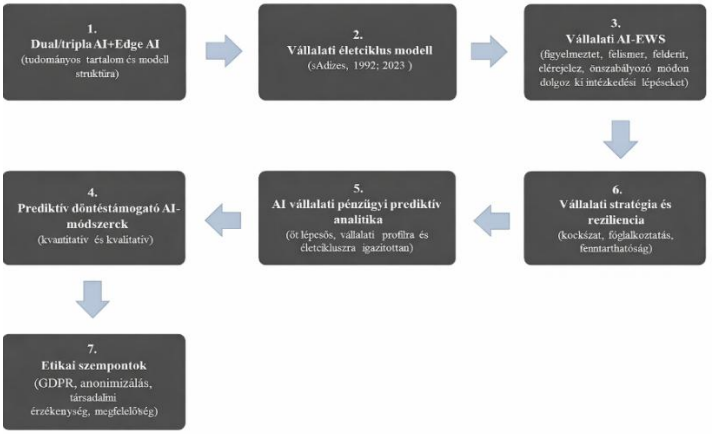
kozóan, miként vezethetők be és adaptálhatók az AI-alapú megoldások a vállalatok működésébe. Ezt a gyakorlati relevanciát erősítik a turizmus és a regionális foglalkoztatás területén megvalósított pilot projektek tapasztalatai (Fejes & Katits, 2025a; 2025b; 2025c), amelyek konkrét példákon keresztül szemléltetik az AI-technológiák szervezeti hasznosíthatóságát.

A hibrid AI-rendszerek alkalmazása lehetőséget teremt a tudásmenedzsment új architektúrájának kialakítására, amelyben a gépi intelligencia felgyorsítja az adatfeldolgozást és az előrejelzést, míg az emberi szereplők biztosítják az értelmezést, az irányítást és az etikai kontrollt. A prediktív modellezés technikai megalapozásában Chen & Guestrin (2016) munkája kiemelkedő jelentőségű, mivel az általuk bemutatott XGBoost (eXtreme Gradient Boosting) algoritmus hatékony és skálázható megoldást kínál nagy dimenziójú, strukturált adatok elemzésére, és mára az egyik legszéle-sebb körben alkalmazott ensemble tanulási módszerré vált a vállalati prediktív analitikában.

Tehát a cél egy olyan AI-alapú vezetői tudásmenedzsment- és döntéstámogatási keretrendszer kidolgozása (1. ábra), amely egyrészt gazdagítja a vezetéstudomány és a tudásmenedzsment elméleti diskurzusát, másrészt konkrét, alkalmazható iránymutatást nyújt a vállalatok számára az AI-technológiák hatékony és felelős integrációjához és az életszakasz-alapú EWS rendszerek együttes alkalmazására épül. Ez a célkitűzés szorosan illeszkedik azokhoz az empirikus eredményekhez, amelyek szerint az AI önmagában nem elegendő a szervezeti teljesítmény javításához; az AI valódi értéke akkor realizálódik, ha integrálódik a tudásmegosztási és tudásmenedzsment-folyamatokba, ami szignifikánsan javítja a szervezeti teljesítményt (Olan et al., 2022).

## 2. A KMDS keretrendszer szakirodalmi háttere

A vállalatvezetői tudásmenedzsment és döntéstámogató (KMDS= Knowledge Management and Decision Support for corporate executives) keretrendszer hét, egymással összefüggő tudományos-szakmai terület (T1-T7) összehangolt integrációján alapul. Itt azt vizsgáljuk meg, hogy ez a hét összetevője hogyan tükröződik a tudományos szakirodalomban (1. ábra).




**1. ábra: A vállalatvezetői tudásmenedzsment és döntéstámogató (KMDS=Knowledge Management and Decision Support for corporate executives) keretrendszer**
**Forrás: Szerzői szerkesztés, 2026.**

## T1. Dual/tripla AI+Edge AI

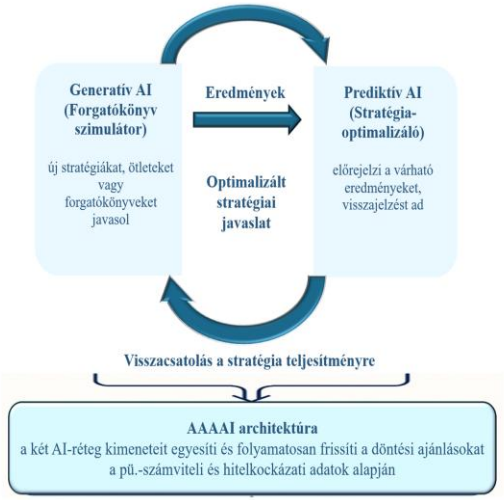
Az AI fejlődése a vállalati döntéstámogató rendszerekben három szakaszra bontható:

i. Az automatizálási célú alkalmazások időszaka, pl. adminisztrációs vagy logisztikai feladatok optimalizálására;
ii. A prediktív rendszerek, a gépi tanulás és a mélytanulás eszközeinek elterjedése, pl. turisztikai kereslet előrejelzésre szezonális adatok alapján;
iii. A XAI és transzparens AI (TAI=Transparent AI), mint a döntéstámogatás átláthatósága különösen fontossá vált a pénzügyi és turisztikai szektorokban, ahol a felhasználói bizalom kulcsfontosságú.

Napjainkban pedig a dual/tripla AI architektúrák a predikció, a magyarázhatóság és adaptivitás szinergiáját valósítják meg, lehetővé téve a regionális sajátosságok figyelembevételét is. Ezeknek a rendszereknek a bevezetése új kérdéseket vet fel a tudásfelhasználásban: Hogyan egészítheti ki a gépi tanulás a vezető tapasztalatát, és miként tudják együtt felerősíteni a tudás cseréjét, rendszerezését és fenntartását? Emellett az Edge AI technológiák (az adatfeldolgozás hálózati peremre helyezés) robbanásszerű elterjedése lehetőséget ad a valós idejű döntéstámogatásra,

ami lehetővé teszi a vezetők számára, hogy naprakész tudásra alapozott döntéseket hozzanak, de kihívásokat jelent az infrastrukturális fejlesztések terén is.

A dual AI olyan rendszereket jelent, ahol az automatizált tudásfeldolgozás és a humán-vezérelt elemzés egyensúlyban működik. Alkalmazásának az a célja, hogy az AI gyors feldolgozási kapacitását kombinálja az emberi döntéshozatal értelmező és kritikai képességeivel. Várható hatása a hatékonyság növekedése, mivel az ember-gép együttműködés révén pontosabb és gyorsabb döntések születhetnek. Ehhez képest a tripla AI egy lépéssel tovább megy: nemcsak a gépi és emberi intelligencia hangolását jelenti, hanem a vezetők céltudatos tréningjeit és soft skill-ek (pl. empátia, kritikus gondolkodás, etikai érzékenység) bevonását is. Így a technológia mellett az emberi tényezők is hangsúlyosabbak, különösen a felelősségteljes vezetésben. Várható hatása a fenntarthatóbb tudásgazdálkodás, hiszen nemcsak rövid távú hatékonyságnövekedést, hanem hosszú távú szervezeti tanulást és rezilienciát támogat (2. ábra).



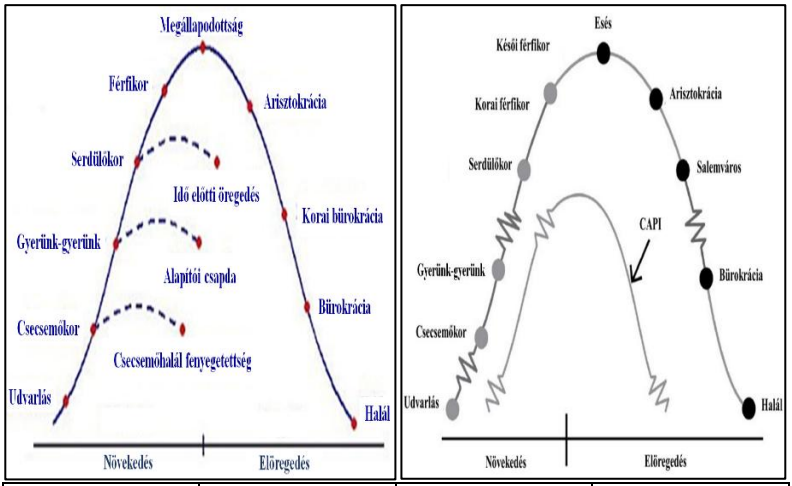

**2. ábra: A dual/tripla AI tartalma és működési logikája**
**Forrás: Szerzői szerkesztés, 2026.**

Az Edge AI alkalmazásával az adatközeli feldolgozás új minőséget ad a tudásmenedzsmentnek. Gyorsabb, valós idejű adatelemzés révén csökken az információs késleltetés, ami lehetővé teszi a vezetők számára, hogy naprakész tudásra alapozott döntéseket hozzanak (pl. a zöld ellátási lánc-

ban és az ipari automatizálásban). Várhatóan a magyar kkv-k számára az Edge AI bevezetése infrastrukturális kihívásokat jelent (pl. IoT=Internet of Things eszközök hiánya), de stratégiai javaslatként megfogalmazható az ipari IoT és 5G rendszerek fejlesztése. Az Edge AI-hoz kapcsolódó magyarázhatósági és transzparencia-kérdések értelmezésében a XAI és SHAP (SHapley Additive exPlanations)-alapú megközelítések is irányadók (Lundberg et al., 2020).

T2. Vállalati életciklus modell

A vállalati működés vizsgálata az adizesi életciklus modell szerinti felfogásban történik, illetve az adizesi életpályáról választunk céget vagy vállalati mintát. Mivel a gyakorlat nem vagy alig ismeri az adizesi megnevezéseket, ezért átírtuk egyszerű, könnyen érthető megnevezésekké. Jelöltük szürke háttérrel, ahol változott a megnevezés: Ami érdekes a stabilitás/megállapodottság szakasza kikerült a modellből (3. ábra).




| | | | |
|---|---|---|---|
| Csecsemőkor (CS)<br>Kezdő, induló | Megállapodottság (M)<br>Stabil | Csecsemőkor (CS)<br>Kezdő, induló | Késői férfikor (KéF)<br>Tartós, de mérsékelten növekvő |
| Gyerünk-gyerünk (GO)<br>Terjeszkedő | Arisztokrácia (A)<br>Stagnáló | Gyerünk-gyerünk (GO)<br>Terjeszkedő | Arisztokrácia (A)<br>Stagnáló |
| Serdülőkor (S)<br>Évről-évre növekedő | Korai bürokrácia (KB)<br>Hanyatló | Serdülőkor (S)<br>Évről-évre növekedő | Salemváros (SV)<br>Hanyatló |
| Férfikor (F)<br>Tartósan növekvő | Bürokrácia (B)<br>Válság, csőd | Korai férfikor (KoF)<br>Tartósan, „mindig” növekvő | Bürokrácia (B)<br>Válság, csőd |

**3. ábra: Az Adizes életciklus modell eredeti (1992) és aktuális (2023) változata**
**Forrás: Katits, 2024, 4. fejezet**

T3. Vállalati AI EWS

A pénzügyi EWS rendszerek első generációi az 1960-as években kezdtek elterjedni, amikor a vállalatok felismertek a pénzügyi stabilitás fenntartásának fontosságát. Ezek a rendszerek jelentős előrelépést jelentettek a vállalati csődök előrejelzésében, de még nagyon korlátozottak voltak a komplex pénzügyi helyzetek előrejelzésében. Tehát az EWS rendszerek a csődelőrejelző modellekből indultak, alakultak ki. Az 1980-as években a számítógépes technológiák fejlődése lehetővé tette a vállalatok számára, hogy nagyobb adatbázisokat és összetettebb statisztikai modelleket alkalmazzanak. Ezzel együtt megjelentek az első szoftveralapú EWS-ek, amelyek gyorsabb és pontosabb elemzést tettek lehetővé. Ezek a rendszerek azonban még mindig korlátozottak voltak a prediktív pontosság és a megbízhatóság tekintetében, mivel elsősorban a múltbeli adatokra és a lineáris előrejelzési módszerekre támaszkodtak. A 2000-es évektől kezdve, az AI alkalmazása lehetővé tette a valós idejű adatok elemzését is, ami különösen fontos volt a gyorsan változó gazdasági környezetben. Napjainkban ún. hibrid EWS rendszerek léteznek, vagyis az AI, a big data, a többváltozós statisztikai eljárások integrációja. A COVID-19 világjárvány által kiváltott gazdasági zavarok rávilágítottak a hatékony EWS-ek szükségességére, amelyek képesek előre jelezni a látens és kifejlett válságokat, valamint segíteni a vezetőket a preventív és proaktív menedzsmentben. négy generáció szerinti megállapításokat emeljük ki.

A vállalati EWS 1. generációja „hű a nevéhez”, mivel ellátja a korai FIGYELMEZTETŐ funkciót (pl. jelzi a működési veszteséget, az üzleti és a likviditási kockázatot). A 2. generáció már korai FELISMERÉS funkciót (pl. lehetőség a növekedésre, de ennek érdekében a vezetőségnek lépéseket kell tenni a költségek optimalizálása és a jövedelmezőség helyreállítása érdekében) teljesít. A 3-4. generáció elvégzi a korai FELDERÍTÉS ÉS ELŐREJELZÉS feladatát, amelynek keretében pl. három intézkedési programot ad, amit még tovább részletez intézkedési lépésekben a realizálás érdekében. Tehát EWS-megoldások kialakítására van igény, amelyek képesek a vállalatok fejlődési szakaszaihoz kötődő pénzügyi és működési kockázatok, és lehetőségek (válaszlépések korai azonosítására, ezáltal növelve a szervezeti rezilienciát és elősegítve a válságok megelőzését (Tanaka et al., 2025).

T4. Prediktív döntéstámogató AI-rendszerek

Az adat- és elemzésalapú prediktív modellek idősorelemzési és regressziós eljárások alkalmazásával, beleértve az olyan korszerű algoritmusokat, mint az XGBoost és az LSTM. Ezeket a modelleket magyarázó modulokkal (Bayes-hálózatok, XAI- és SHAP-megoldások) egészíti ki annak érdekében, hogy azonosíthatók legyenek a legfontosabb előrejelző tényezők – különösen a pénzügyi, foglalkoztatási, ESG-indikátorok – akár a régiók, akár az adott vállalat gazdasági és foglalkoztatási életciklusainak előrejelzésében (Fejes & Katits, 2025a; 202b; 2025c).

A kvantitatív prediktív modellek nagyrészt számszerű adatokon és algoritmusokon alapulnak, és számszerű előrejelzéseket adnak. A kvantitatív modellek közös jellemzője az, hogy adatvezéreltek és objektív előrejelzést adnak. Ugyanakkor az eredmények validálásában mindig fontos az emberi tapasztalat és kreativitás bevonása is, mivel a modellek a múlt vagy a jelen struktúráiból indulnak ki. A kvalitatív jellegű modellek inkább esettanulmányokra, szakértői tudásra, narratívákra és szimulációkra építenek, a stratégiai döntéshozók számára inkább keretrendszert vagy lehetséges forgatókönyveket vázolnak fel, semmint konkrét számokat. A prediktív modellezés eszköztára rendkívül széles: a kemény számoktól és algoritmusoktól kezdve a puhább, emberi kreativitást igénylő módszerekig sok minden ide tartozik. Gyakran a leghatékonyabb megoldás ezek kombinálása (pl. egy stratégiai szcenáriótervezés, amihez kvantitatív előrejelzéseket is mellékelnek).

A kutatási eredmények triangulációját és validitását szolgálja a módszertan kulcsa, vagyis az előzetes feltételezés szerint a módszerek kombinálása erősebb bizonyítékokat ad, mint bármelyik különálló


**4. ábra: Prediktív döntéstámogató AI-módszerek**
**Forrás: Szerzői szerkesztés, 2026.**

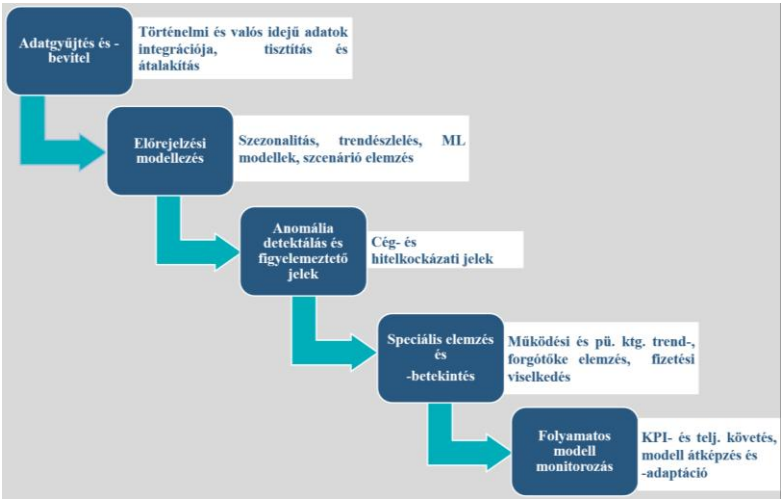
módszer önmagában. A statisztikai elemzések által felfedett trendeket az interjúk és esettanulmányok lehetőség szerint megerősítették vagy árnyalták. Például, ha a felmérés szerint egy adott AI-funkciót a vezetők előnyösnek ítélnek, az interjúkból kiderült, hogy konkrét példák is alátámasztják ezt (pl. AI alapú kockázatfigyelés bevezetése). A különböző adatforrások összevetése javítja a kutatás belső konzisztenciáját és külső validitását. A kvantitatív és kvalitatív eredmények triangulációja megerősíti a következtetéseink megbízhatóságát: a két módszerből nyert adatok, ha jó egyezést mutatnak, akkor együttesen árnyalt, konzisztens képet nyújtanak a vezető által vizsgált témáról (4. ábra).

T5. AI vállalati pénzügyi prediktív analitika

Ez a modul leszűkül a vállalati pénzügyek vezetői tudásanyagára.

A gépi tanulás, a természetes nyelvi feldolgozás és a prediktív elemzés révén a vállalatok hatalmas mennyiségű pénzügyi (és nem pénzügyi) adatot elemezhetnek, rejtett mintákat azonosíthatnak, és valós idejű betekintést generálhatnak a befektetési döntések irányításához. A vállalatok optimalizálhatják a tőkeallokációt, csökkenthetik a befektetési tervezés bizonytalanságát és maximalizálhatják a hosszú távú részvényesi értéket. Az AI nem csupán egy támogató eszköz, hanem egy átalakító erő, amely képes újra definiálni a vállalati pénzügyi stratégiát.

Az 5. ábra az ötlépcsős AI vállalati pénzügyi prediktív analitika menete „a vezető döntéshozó hurokban van”-helyzetre készült, s ez a helyzet nem ritka változó világunkban. Ezeket a lépéseket és tartalmakat testre szabottan, komplex feltételrendszerben kell elvégezni. A tartalmát itt most nem részletezzük (a háttéranyag Katits, 2024). Egyet kiemelünk: A 4. lépcső, amely a „Speciális elemzés és -betekintés”, avagy a „testreszabott”, életszakaszra és működési profilra igazított pénzügyi számítások megnevezést kapta.




**5. ábra: Az ötlépéses vállalati pénzügyi prediktív analitika készítése**
**Forrás: Szerzői szerkesztés, 2026.**

## T6. Vállalati stratégia és reziliencia

A reziliencia különböző tudományterületeken belüli definícióinak tartalomelemzése alapján a reziliencia a rendszerek viselkedésének egy közös értelmezése, amely olyan kifejezésekkel írható le, mint az alkalmazkodás, visszakeresés, teljesítmény, megtartás, konfrontáció, visszatérés, válasz és ellenállás (Zahedi et al., 2023). Az általános rendszerelméletre építve a vállalati reziliencia elmélete az emberi erőforrások, a társadalmi-kulturális értékek, az intézményi környezet, valamint a társadalmi és környezeti kérdések inputjait tartalmazza, lehetővé téve a szervezeti struktúrát, az érték- és hiedelem alrendszert, a kimenet a fenntarthatósági gyakorlatok és teljesítmény, az alkalmazkodó és pufferelő kapacitások (Kantabutra & Ketprapakorn, 2021). A vállalati stratégia arra irányítja a figyelmet, hogy a cégek milyen környezetben működjenek. A vállalati stratégiai menedzsment folyamat része annak biztosítása, hogy a vállalatok megfelelő rezilienciával rendelkezzenek. Zahedi et al. (2023) szerint három tényező járul hozzá a szervezeti rezilienciához: 1. Az ellenálló képesség a szervezet funkciója. 2. A szervezeti működésből (nevezetesen a szervezet által végrehajtott cselekvésből) eredő ellenálló képesség. 3. A rugalmasság a szervezet által tolerálható anomáliák mérőszáma.

Ruiz-Martín (2018) úgy véli, hogy mindezek a definíciók ugyanazt jelentik, és a szervezet túlélésére utalnak a sokkhatásokkal és kockázatokkal szemben. A kockázat a fenyegetés, a rugalmasság pedig a válasz erre a fenyegetésre.

A kockázati rugalmasság nemzetközi szabványa az ISO 31000, amely egy olyan keretrendszert biztosít a kockázatkezeléshez (pénzügyi, működési, stratégiai, informatikai kockázatok, kiberbiztonság és megfelelés), amelyhez minden szervezet alkalmazkodni tud. Az ISO 31000 alapelvei közé tartozik a kockázatok azonosítása, értékelése és folyamatos kezelése, valamint a kockázatkezelés integrálása a szervezet minden szintjén. Az ISO 31000 megközelítés szerint az AI technológiái (pl. dual AI-alapú prediktív modellek) kulcsszerepet játszhatnak a kockázatkezelés folyamatos ciklusában (tervezés, végrehajtás, monitorozás, fejlesztés) a fenyegetések előrejelzésében és a kockázatkezelés automatizálásában (TechTarget, 2025). A RISKRES (RISK-RESilience) modell együttesen kezeli a kockázatelemzést és a rugalmas stratégiák kidolgozását (Singh, 2022).

T7. Etikai szempontok

Az adatgyűjtést a GDPR-nak megfelelő eljárások alkalmazásával történik, beleértve a személyes adatok anonimizálását, a hozzájáruló nyilatkozatok dokumentálását és a társadalmi érzékenységi szempontok figyelembevételét a kísérleti régióban. Az AI modellekhez külön XAI és etikai megfelelőségi protokollt is be kell vezetni.

Tehát KMDS keretben végzett kutatásunk a dual/tripla AI-rendszerek és az Edge AI alkalmazását vizsgálja, összekapcsolva az Adizes-féle vállalati életciklus-modellt az AI-EWS rendszerekkel és a vezetői tudásmenedzsment folyamataival. Ez egy olyan integrált elméleti és gyakorlati keretrendszer, amely egyrészt hozzájárul a vezetéstudomány és a tudásmenedzsment elméleti fejlődéséhez, másrészt gyakorlati iránymutatást nyújt a vállalatok számára az AI-alapú, reziliens és életszakasz-érzékeny döntéstámogatási megoldások bevezetéséhez.

## 3. Anyag és módszer

A kutatás vegyes módszerre épül, amely ötvözi a kvantitatív felmérést és a kvalitatív elemzést, így a vizsgálat eredményei mindkét megközelítés előnyeit kamatoztatva komplex, megbízható eredményt hoznak. A

mintavétel során a hazai kontextusok széles spektrumát képviseli. Az 1. táblázatban közölt vállalati adatbázisokból választottunk.

A kvantitatív felmérésben magyarországi turisztikai vállalatok vezetői vettek részt, míg a kvalitatív esettanulmányok az összehasonlíthatóságot és az vegyes módszerrel kapott eredmények lefedettségét segítik. A mintaszerkezet összeállítását a 2024. év végi kérdőíves felmérés szolgálja. A mintavételi eljárás törekedett a turisztikai cégek méret és tevékenységek (szolgáltatások) szerinti eloszlásának megfelelő kiválasztására és a reprezentativitásra A mintavétel tudatosan rétegzett és célzott, amely lehetővé teszi a kutatásba bevont vezetők és szakértők körének kiválasztását előre meghatározott kritériumok (cégméret, pozíció, szektor) alapján. Így biztosított a releváns tapasztalattal bíró szereplők kutatásba történő bevonása: A célpopulációk eltérő méretű (kis-közép- és nagyvállalat), szektorális besorolású és vezetői pozíciót betöltő résztvevők, akik biztosítják a széleskörű tapasztalati hátteret.

**1. táblázat: Igazolás és tesztterepek – Szekunder és primer vállalati adatbázis**

| | |
|---|---|
| 1. | Csőd- és felszámolási eljárás alatt álló, 120 magyar társaság 1998-tól |
| 2. | A magyar TOP 5000 vállalkozás mérleg és eredménykimutatásai |
| 3. | TEÁOR kódszámok alapján, hazai és külföldi tulajdon szerint 1992-től |
| 4. | A magyar KKV-szektorba tartozó TOP 4 000 vállalkozás mérleg és eredménykimutatásai |
| 5. | TEÁOR kódszámok alapján, hazai és külföldi tulajdon szerint 1992-től |
| 6. | A magyar TOP 100 mérlegei és eredménykimutatásai, valamint szervezeti átalakulásai és formaváltásai 2007-től |
| 7. | A magyar TOP 500 mérlegei és eredménykimutatásai, valamint szervezeti átalakulásai és formaváltásai 2007-től |
| 8. | A magyar TOP 1000 mérleg és eredménykimutatás összesített adatai TEÁOR ’08 ágazatok szerint megyei bontásban 2007-től |
| 9. | 100 magyar családi vállalkozás mérlegei és eredménykimutatásai 2000-től |
| 10. | 90 konkrét nagyvállalati és kkv-esetanalízis 2007-től, pilot |
| 11. | AKI adatbázis |
| 12. | Magyar top 1000 kérdőív |
| 13. | Két reprezentatív vállalati minta a turizmus szektorba |
| 14. | 10 nemzetközi és 5 magyar EWS alkalmazás turizmus, pénzügy-számvitel, gépjármű. |

**Forrás: Szerzői szerkesztés, 2026.**

A kvantitatív felmérésben magyarországi turisztikai vállalatok vezetői vettek részt, míg a kvalitatív esettanulmányok az összehasonlíthatóságot és az vegyes módszerrel kapott eredmények lefedettségét segítik. A mintaszerkezet összeállítását a 2024. év végi kérdőíves felmérés szolgálja. A mintavételi eljárás törekedett a turisztikai cégek méret és tevékenységek (szolgáltatások) szerinti eloszlásának megfelelő kiválasztására és a reprezentativitásra A mintavétel tudatosan rétegzett és célzott, amely lehetővé teszi a kutatásba bevont vezetők és szakértők körének kiválasztását előre meghatározott kritériumok (cégméret, pozíció, szektor) alapján. Így biztosított a releváns tapasztalattal bíró szereplők kutatásba történő bevonása: A célpopulációk eltérő méretű (kis-közép- és nagyvállalat), szektorális besorolású és vezetői pozíciót betöltő résztvevők, akik biztosítják a széleskörű tapasztalati hátteret.

A kvalitatív kutatási design másik része a félig strukturált, szakértői mélyinterjúk (N=12), amelyek értelmezése reflexív tematikus elemzés keretében történt. A mélyinterjút szakértőkkel (szállodai vezetők, desztináció-menedzsment szakemberek, turisztikai startupok képviselői) készült. Az interjúk kérdései az AI gyakorlati alkalmazására, előnyeire, korlátjaira, valamint a válság megelőzés érdekében való hozzájárulására fókuszáltak. A szövegek tematikus elemzése azt jelenti, hogy induktív módon történő kódolással, majd ismételt iterációkkal fő kategóriák kerültek azonosításra az adatokban. Célzott/kritérium alapú mintavétel keretében olyan szállodai vezetők, desztináció-menedzsment szakemberek és turizmus-analitikai startupok képviselőivel készültek az interjúk, akik bizonyíthatóan döntési felelősséggel és AI-expozícióval bírnak. A 12 interjú a hazai ökoszisztéma különböző szervezeti érettségi fokait és technológiai kitettségét fedi le. Minden interjú rögzítése beleegyezéssel történt.

A kvalitatív kutatás harmadik komponense két esettanulmány feldolgozása az AI turizmusban való alkalmazásáról. Ezek valós piaci kontextust nyújtottak a kérdőíves eredmények értelmezéséhez. A kutatásunk kvalitatív részében két esettanulmányos designt választottunk, amely szemlélteti a dual AI-megoldások működését és hatásait. Az első, magyarországi eset az EWS alkalmazását mutatja be a turisztikai kereslet előrejelzésében és válságkezelésben. A második, balatoni eset, amely az intelligens desztináció tervezésről szól, különösen a kockázatkezelés eszközeiről). Mindkét esetben több forrásból származó információgyűjtés (adat trianguláció) történik annak érdekében, hogy megbízhatóbb következtetéseket levon-

hassunk. Az adatgyűjtés elemei: internetes esetleírások áttekintése, nyilvános statisztikák (Magyar Turisztikai Ügynökség (MTÜ) és Központi Statisztikai Hivatal (KSH), EUROSTAT adatok), vállalati jelentések és sajtóanyagok elemzése. Az esettanulmányok elemzésénél kombináltuk a leíró statisztika és az induktív tematikus elemzés mód-szereit. A kvantitatív adatokat grafikus és táblázatos formában is megjelenítettük. Ugyanakkor a minta korlátjai miatt a következtetések nem általánosíthatók az egész iparágra; a cél inkább mélyebb insightok nyújtása, semmint reprezentatív statisztikai bizonyítás. A kvalitatív adatok elemzését tematikus kódolással végeztük („előrejelzési pontosság", „döntéshozatal gyorsasága", „szervezeti kihívások", „fenntarthatósági hatások"), majd a

**2. táblázat. A vizsgált minta összetétel vizsgálata az országos arányok ismeretében**

| Jellemző | Kutatási minta | Országos adat | Forrás |
|---|---|---|---|
| **Mintanagyság (db)** | 1 157 | 104 600 | KSH (2024) |
| **KKV arány (%)** | 70% (810 db) | 73% (42 860 db) | KSH (2024) |
| **Nagyvállalat arány (%)** | 19.9% (230 db) | 17% (9 747 db) | KSH (2024) |
| **Nonprofit/állami/egyéb (%)** | 10.1% (117 db) | 10% (4 728 db+egyéb) | KSH (2024) |
| **AI/digitális használat (%)** | 49.9% (önbevallás) | 11% (szállás-helyeknél) | Statista (2024) |
| **Budapest arány (%)** | 23.3% | 25% | KSH, MTÜ (2024) |
| **Balaton arány (%)** | 16% | 28% | KSH, MTÜ (2024) |
| **Nyugat-Dunántúl (%)** | 9,5% | 11% | KSH, MTÜ (2024) |
| **Közép-Dunántúl (%)** | 9,5% | ~10% | KSH, MTÜ (2024) |
| **Észak-Magyarország (%)** | 12,1% | ~8% | KSH, MTÜ (2024) |
| **Dél-Dunántúl (%)** | 10,4% | ~10% | KSH, MTÜ (2024) |
| **Alföld és más régiók (%)** | 19,2% | ~8% | KSH, MTÜ (2024) |

**Forrás: A szerzői összeállítása alapján- Központi Statisztikai Hivatal, KSH. (2024); Statista Research Department. (2024); Magyar Turisztikai Ügynökség. MTÜ (2024)**

dokumentumok átolvasása után lettek finomítva a kódok és új témák azonosítása is történt. Ezzel a módszerrel sikerült feltárni az esetek közös nevezőit és az egyedi, kontextus-specifikus tanulságokat is.

A kvantitatív adatgyűjtés 2024 második felében történt, önkitöltős online kérdőív formájában. A kérdőív Likert-típusú többválasztós elemeket tartalmazott az AI-alkalmazás területeiről, az előnyökről és kihívásokról. A minta 1157 fő magyarországi turisztikai vállalatvezetőt foglalt magában. A minta elemeit a célcsoport arányaihoz igazítottuk, így például a kis-, közép- és nagyvállalati vezetők közel azonos arányban szerepelnek a mintában. A rétegzett mintavétel és a tudatos kvótaalkalmazás azt garantálja, hogy a minta tükrözi a valós vállalati struktúrát, ezáltal növelve az eredmények általánosíthatóságát és megbízhatóságát (2. táblázat).

A földrajzi, szektorális és volumetrikus összevetés alapján a kutatási minta bizonyíthatóan reprezentatív a magyarországi turizmus szektorára. A minta földrajzi és vállalati struktúrája arányos az országos megoszlással, a mintanagyság statisztikailag szignifikáns. Kissé magasabb a digitális/AI-innováció aránya, mivel a kutatás célja a fejlett megoldások elemzése is volt.

A dual AI rendszer teljesítményét az előrejelzési pontosság és a kockázatok mérésével lehet értékelni. A kutatásban kifejlesztett Dual AI rendszer két modult tartalmaz.

**AI1** – prediktív modul: regressziós és idősoros modellek (XGBoost, LSTM). A prediktív modul kimenete rendszerint jövőbeli trendek és igények vagy növekedési pályák formájában jelenik meg. A modellek pontosságát jellemző mutatók közé tartozik a Root Mean Square Error (RMSE) a predikciók szórása; a Mean Absolute Error (MAE) a tévedés nagysága; a determinációs együttható, mint magyarázott variancia ($R^2$) azt mutatja, hogy a modell mennyire képes visszaadni a célváltozó ingadozásait (Zhang et al., 2025). Egy modell elfogadhatósága, illetve két modell összehasonlítása során minél alacsonyabb a RMSE és a MAE, és minél magasabb az $R^2$, annál pontosabb a becslés.

A pontosság mérésekor érdemes külön vizsgálni a rendszer teljesítményét nyugodt időszakokban és krízisforgatókönyvek alatt. Ideális esetben a dual AI javítja a pontosságot mind nyugodt időszakokban, mind válságforgatókönyvek alatt azáltal, hogy a generált stressziteszt adatokra is kiképeztük. Ezt támasztja alá az, hogy szintetikus adatok bevonásával

egyes előrejelző modellek robosztusabbá váltak és kisebb hibával dolgoztak (Chatterjee & Byun, 2023). Fontos a valós pozitív arány és a valós negatív arány (beavatkozások a kisebb ingadozásoknál, ami túlzott reagálás) mérése is.

A pandémiát követően az EWS modellbe kerültek a pandemic dummy változók, illetve úgy lehet módosítani a rendszert, hogy külső források (pl. járványügyi adatok, nemzetközi utazási korlátozások hírei) is részét képezzék a riasztási mechanizmusnak. Ennek eredményeként a modell érzékenyebben reagál a szokatlan jelekre. Például 2022 elején a rendszer sárga figyelmeztetést adott (az orosz-ukrán háború kitörése után a külföldi foglalások számottevően visszaestek), mivel a modell az előző sokk (COVID-19) mintázata alapján gyorsabban „megtanulta" kezelni. **A**z AI-alapú prediktív modellek óriási értéket hordoznak a vállalati menedzsment számára. A történelem során először rendelkezünk olyan eszközökkel, amelyek a jövő „megsejtését" tudományos alapokra helyezhetik, ezáltal az üzleti döntéshozatal bizonytalanságát csökkenthetik – jóllehet a bizonytalanság sosem iktatható ki teljesen.

Összességében: ha javul az előrejelzések pontossága, csökken a volatilitás és a potenciális veszteség, valamint mindez nagyobb profitabilitással párosul, akkor a rendszer módszertanilag is megalapozott és üzletileg is indokolt.

**AI2** – metaértékelő modul: torzítás/elfogultsági (bias)-korrekció, hibadetektálás és validáció (Bayes-hálózat, SHAP/XAI modulok).

A tripla AI architektúra a két AI-réteg kimeneteit egyesíti, és folyamatosan frissíti a döntési ajánlásokat a regionális munkaerő- és turizmusadatok alapján.

A kvantitatív adatokat SPSS-ben és Pythonban elemezzük, valamint SMESBJ okos applikációt alkalmazunk (Katits, 2024).

Az alkalmazott statisztikai módszerek – többszörös regresszió: a dual AI predikciós pontosságának vizsgálata; faktoranalízis: a foglalkoztatási életciklusfázisok összehangolása; klaszterelemzés: régió ESG és turizmusfüggőségi profilja szerint; Monte Carlo-szimuláció: kockázati szintek validálására különböző forgatókönyvek mentén; ROC görbék és AUC értékek: az AI modellek diagnosztikai pontosságának értékelésére.

A kutatásban használt ROC (Receiver Operating Characteristic) görbe és AUC (Area Under Curve) értékek a prediktív modellek teljesítményének megbízható és szabványos mérőeszközei. A ROC-görbe az igaz pozitív arány (szenzitivitás) és a hamis pozitív arány (1-specifitás) kapcsolatát ábrázolja különböző küszöbértékek mentén. Az AUC-érték a ROC-görbe alatti területet jelöli, és numerikus mutatóként szolgál a modell általános teljesítményére vonatkozóan. Az 1-hez közelítő AUC-érték kiváló prediktív képességet jelez (Bradley, 1997; Huang et al., 2020; Muschelli, 2019), vagyis a dual/tripla AI rendszerek prediktív algoritmusai szignifikánsan megbízható döntéstámogatást nyújtanak a smart turizmus és regionális foglalkoztatás kontextusában (Fawcett, 2006).

A vállalati AI-EWS teszteléshez négy algoritmust is kipróbálunk: baseline modellként egy logisztikus regressziót, majd korszerűbb ML modelleket, mint a döntési fa, Random Forest és neurális háló. A rendszer teljesítményét először konfúziós mátrix segítségével szemléltetjük, amely bemutatja a valódi pozitív, valódi negatív, hamis pozitív és hamis negatív előrejelzéseket. Ez alapján számoltuk ki a legfontosabb osztályozási metrikákat. A modellek bemenetei a vállalati pénzügyi (mérleg- és eredménykimutatásból számított) ráták (SMESBJ applikációval, Katits 2024), kiegészítve turizmus-specifikus indikátorokkal (pl. szezonalitási mutató, regionális turizmus trendek). A kimenet egy előrejelzés a vállalat állapotára nézve a következő időszakban – tipikusan bináris osztályozás formájában ("stabil" vs. „veszélyeztetett"), de fontolóra vettük több kategória használatát is (pl. fizetőképes, fizetésképtelenség felé tart, nyereséges/veszteséges, átlagos havi bevétel nő/csökken/stagnál). Az EWS modelleket mérőszámokkal hasonlítjuk össze. A ROC-görbét mind a négy modellre ábrázoljuk. A várakozásunk az, hogy a legjobb modell teljesítménye meghaladja a hagyományos módszerekét. Vizsgáljuk a modell interpretálhatóságát is: a Random Forest esetében kiszámítjuk a változók *feature importance* értékeit, illetve alkalmazunk SHAP értékeket a neurális háló modellnél annak érdekében, hogy azonosítsuk a legfontosabb kockázati faktorokat.

## 4. Eredmények

Az AI1 prediktív modul teljesítményét illetően az XGBoost regressziós algoritmussal képeztük ki a rugalmasság és a foglalkoztatási életciklus-fázisok előrejelzésére. Az AI1 prediktív modul esetében kiszámítottuk az RMSE, MAE és $R^2$ értékeket is. A modellváltozatok (ARIMA vs. LSTM)

összehasonlításakor a 3. táblázatban szereplő mutatók szemléltetik a neurális hálózat és a lineáris modell közötti különbségeke. A prediktív modellek közül az LSTM teljesített jobban: RMSE=170, MAE=13,80, $R^2$=0,89, míg az ARIMA modell RMSE=220, MAE=18,50, $R^2$=0,73 értékeket mutatott.

**3. táblázat: Példa a modell illesztési mutatóira**

| Model | RMSE | MAE | $R^2$ |
|---|---|---|---|
| **LSTM** | 170 | 13.80 | 0.89 |
| **ARIMA** | 220 | 18.50 | 0.73 |

**Forrás: Szerzői szerkesztés, 2026.**

Az AI2 modul (szimuláció) kiértékelése némileg eltérő megközelítést igényelt. Itt elsősorban a generált forgatókönyvek valószínűségét vizsgáltuk, és szakértői ismeretekre támaszkodva értékeltük a rendszer reakcióinak hitelességét. Azt validáltuk, hogy a szimulációs kimenet összhangban van-e a valós megfigyelésekkel: például a modellnek egy adott típusú sokkhatás (legyen az gazdasági válság vagy természeti katasztrófa) hatásaihoz hasonló irányú változásokat kell-e mutatnia, ahogyan azt a történelminek számító, globális pénzügyi és pandémia válságban is tapasztalhattuk. Érzékenységi elemzést is végeztünk, azaz azt vizsgáltuk, hogy a bemeneti feltételek (pl. a GDP és a kereslet csökkenésének mértéke, a hőmérséklet-emelkedés intenzitása) kismértékű módosítása mennyire változtatja meg a kimeneti eredményeket, és hogy ezek az eltérések reális tartományon belül maradnak-e. Fontos követelmény volt, hogy a prediktív és a szimulációs komponensek konzisztensek legyenek. Ha ugyanazokkal a bemeneti paraméterekkel (pl. egy tipikus üzleti környezet) futtattuk őket, akkor hasonló trendeket kell produkálniuk; ha eltérés mutatkozott, akkor a modellek további finomhangolását végeztük. Az SHAP elemzés szerint az előrejelzések legfontosabb magyarázó változói a szezonalitás (SHAP=0,41), a legbefolyásosabb előrejelző, a likviditási mutatók (SHAP=0,33), a pénzügyi fenntarthatóság alapvető előrejelzője, valamint a foglalási átfutási idők (SHAP=0,29) voltak, amelyek a szezonális jelleg miatt is kritikus információk a turisztikai döntéshozók számára. Tehát ez a három tényező befolyásolta leginkább az előrejelzést. Ezek az eredmények segítik a döntéshozókat a kulcsfontosságú tényezők azonosításában és kezelésében.

A fenntarthatósági teljesítményt egy összetett ESG-index segítségével értékeltük, amely átlagosan 75/100-as pontszámot adott, ami azt jelzi,

hogy az ágazat közepes fenntarthatósági teljesítményt ér el, és jelentős fejlődési potenciállal rendelkezik a környezeti és társadalmi dimenziókban.

A Monte Carlo-szimuláció több mint 10 000 futtatása alapján a turisztikai bevételek várható eltérése ±12–18% között volt, míg a P90 kockázati szinten a veszteség esélye alig 10% alá csökkent. A RISKRES modell által számított átlagos kockázati kitettségi index 0,60 volt, míg a rugalmassági index 0,80, ami azt jelzi, hogy a turisztikai vállalatok viszonylag gyorsan képesek talpra állni a kedvezőtlen piaci események után. Az egyes szimulációs klaszterek jellemzőinek elemzésével (pl. enyhe vs. súlyos hatású esetek) rámutattunk arra, hogy mely rendszertulajdonságok felelősek a hasonló eredményekért: Például kiderült, hogy a legnagyobb veszteségeket mutató forgatókönyv-csoportokat alacsony szezonon kívüli kereslet és magas munkaerőköltségek jellemezték. Ezzel szemben a rugalmassági csoportok jellemzően diverzifikált bevételi forrásokkal és rugalmasabb munkaerő-struktúrával rendelkeztek. Ezek a felismerések az XAI-megközelítés részét képezték, növelve a modell üzleti értékét és elfogadottságát a turisztikai döntéshozók körében.

A vállalati AI-EWS modell tesztelését a négy gépi tanulási modell (logisztikus regresszió, döntési fa, neurális hálózat és véletlenszerű erdő) teljesítménye reprezentálta (öt kritérium szerint: pontosság, precizitás, visszahívás, FI-mutató, AUC), amely elérte vagy meghaladta a 70%-ot. A legjobb eredményt, és így a legnagyobb hatékonyságot azonban a random forest adta 85%-os pontossággal. Tehát a dual AI-EWS rendszer 85%-os pontossággal jósolta meg a kritikus pénzügyi eseményeket. A random forest jellemzőinek fontossági rangsora a következő: 1. (Turizmus) Szezonalitási index; 2. Likviditási ráta; 3. Foglalási átfutási idők; 4. Átlagos havi jövedelem; 5. Adósság/Összes forrás arány; 6. Munkavállalói fluktuáció. A SHAP-vizsgálat megerősítette azt, hogy az 1-3. nagyságrend/érték felelős az előrejelzés magyarázó részéért; ezek a legjelentősebb előrejelzők, ami összhangban van Tanaka et al. (2025) nemzetközi eredményeivel.

Összességében a reprezentatív és rétegzett magyar minta azt mutatja, hogy a dual/tripla AI rendszerek nemcsak elméletben, hanem a gyakorlatban is javíthatják a döntések minőségét és a válságkezelés hatékonyságát a turisztikai ágazatban. Ezek az eredmények tudományosan megerősítik, hogy a dual/tripla AI rendszerek statisztikailag megalapozottan képesek

támogatni a turizmus, a foglalkoztatás és az életciklusra igazított EWS-modellek fenntarthatóságát, a döntéshozatal prediktív és adaptív képességeinek erősítésével. Egyúttal az AI prediktív és adaptív képességét kontrollálni és validálni lehet az AI-EWS „jóságfokával" együtt (6. ábra)


**6. ábra: Az AI-prediktív és adaptív képesség kontrollálásának és validálásának módszerei**
**Forrás: Szerzői szerkesztés, 2026.**

A kvalitatív kutatásban a félig strukturált, szakértői mélyinterjúkból (N=12) származó eredmények a következők:

a. A technológiai függőség és humánhatás vizsgálatában az automatizáció és kiszervezett döntésképesség miatti kontrollvesztés-érzés jelentkezett, a humán szerepek átrendeződtek (front/back-office), amely képzésigényt vált ki. Ez összefügg a kvantitatív kutatásban látott negatív percepciók magasabb arányával (≈40%), különösen érett, standardizált folyamatoknál.
b. Adatvédelem és bizalom: Adatkezelés, harmadik féltől származó modellek, AI-as-a-Service átláthatósági dilemmái; XAI/TAI-igény explicit megjelenése. Különösen érzékeny a vendégpanasz-kezelés és dinamikus árazás; bizalom előfeltétel, amelyet a transzparencia növel.
c. A válságmegelőzés érdekében a korai előrejelzés (foglalási trendek, mobilitási mintázatok), kapacitás- és árstratégia adaptív hangolása szükséges, az Edge AI-jellegű lokális analitika gyors

reakcióidővel valósul meg. Ez illeszkedik a Turizmus 4.0 reziliencia-narratívához és a kvantitatív „közép-dominancia" gyakorisági mintához (adaptáció folyamatban).

d. A szervezeti kihívások és megvalósítás vizsgálatában az integrációs költség, az örökölt rendszerek, az adattisztaság, a belső ellenállás (munkahelyféltés) jelentkezett. A képességfejlesztés (adatkompetencia) és irányítás (adat- és AI-irányítás) feltétele a skálázásnak.

e. A fenntarthatósági és szolgáltatásminőségi metacélok tekintetében az energiagazdálkodás, a kihasználtság-optimalizálás, a környezeti lábnyom csökkentése, ugyanakkor etikai kérdések a megfigyelhetőség és megkülönböztetés kockázata körül kerültek előtérbe.

A kutatásunk kvalitatív részében a következő két magyar esettanulmányt (E1-E2) dolgoztuk fel eset leírása→működés és innováció→eredmények és hatás logika mentén.

E1. AI-alapú EWS a magyarországi turizmusban

Eset leírása – A Magyar Turisztikai Ügynökség (MTÜ) 2019-ben pilot jelleggel elindított egy AI-alapú előrejelző és riasztó rendszert a turisztikai kereslet monitorozására (https://mtu.gov.hu/cikkek/strategia/). A rendszer egy LSTM neurális hálózatot használ a havi vendégéjszakák számának előrejelzésére. Input adatok: a korábbi évek turisztikai forgalmi adatai, Google Trends keresési indexek, repülőjárat-foglalási adatai és gazdasági indikátorok. A rendszer időben jelzi, ha a turizmusban szokatlan trend-változás következhet be (pozitív, pl. egy kiugróan jó szezon; negatív, pl. kereslet-visszaesés). Az AI-EWS előrejelzései segítik a pontosabb pénzügyi tervezést (bevétel-előrejelzést) mind vállalati, mind nemzeti szinten.

Működés és innováció – A bevezetett EWS két komponenst tartalmazott: (1) előrejelző modul havi bontásban 12 hónapra előre prognosztizálta a vendégéjszakák számát; (2) anomália-detektáló modul valós időben figyelte a friss beérkező adatokat és riasztást generált, ha azok szignifikáns eltéréseket mutattak az előrejelzett trendtől. A rendszer kimenete egy online dashboard, ahol a döntéshozók grafikonokon látták az „alapeseti" előrejelzést és a becsült konfidenciasávot, valamint egy indikátort, ami zöld-sárga-piros színekkel jelezte a kockázati szintet (zöld – minden a várakozások szerint alakul; sárga – kisebb eltérés, figyelmet igényel; piros

– jelentős eltérés, beavatkozás szükséges). Az innováció abban rejlett, hogy mindezt megelőző jelleggel, nem csupán utólagos kimutatásként tette lehetővé. A rendszer indulásakor – 2019 végén – a 2020-as évre optimista előrejelzést adott, mivel semmi nem utalt akkor rendkívüli eseményre. 2020 februárjában azonban a modellt „meglepetésként" érte a COVID-19 gyors terjedése. Március első heteiben a rendszer piros riasztást adott, megerősítve a nemzetközi hírek alapján megfogalmazódó aggodalmakat, miszerint a turizmus súlyos visszaesés elé néz. A korai jelzésnek köszönhetően gyorsan ki kellett dolgozni egy válságkezelési tervet. Bár a válság mélységét a modell nem tudta kvantifikálni – hiszen addig példa nélküli helyzet volt – a riasztó funkció értékesnek bizonyult: hónapokkal korábban figyelmeztetett a problémára, mintha ahogyan a hivatalos statisztikák.

Eredmények és hatás – A magyar EWS tapasztalatai vegyesek voltak. Normál időszakban (amikor a pandémiás sokk már lecsengett) a rendszer előrejelzései egész jó pontosságot értek el: éves szinten ~5%-os átlagos eltérés volt a valós vendégéjszakák és a modell jóslatai között, ami jelentős javulás a korábbi, egyszerű trendextrapoláción alapuló előrejelzésekhez képest (10-15% hibával működtek). 2020 kivételes évében viszont a modell is drasztikusan tévedett, mivel a modell 110 fölötti indexet valószínűsített a járvány nélkül. 2021-től a modell illesztése javult, a pandémia utáni adatok bevonásával már képes volt követni a részleges fellendülést. Az integrált AI-megoldás korábban/időben jelzi az esetleges kereslet-visszaeséseket, így gyorsabban mozgósíthatóvá válnak a kockázatkezelő intézkedések.

A pandémiát követően az EWS modellbe kerültek a „pandemic dummy" változók, illetve úgy lehet módosítani a rendszert, hogy külső források (pl. járványügyi adatok, nemzetközi utazási korlátozások hírei) is részét képezzék a riasztási mechanizmusnak. Ennek eredményeként a modell érzékenyebben reagál a szokatlan jelekre. Például 2022 elején a rendszer sárga figyelmeztetést adott (az orosz-ukrán háború kitörése után a külföldi foglalások számottevően visszaestek), mivel a modell az előző sokk (COVID-19) mintázata alapján gyorsabban „megtanulta" kezelni. 2022 nyarán, amikor a belföldi turizmus a magas infláció miatt lanyhulni kezdett, az előrejelző modul ezt 1-2 hónapos előnyben jelezte, lehetővé téve egy időben elindított belföldi kampányt, ami végül segített stabilizálni a keresletet.

E2. Intelligens desztinációtervezés a Balaton térségében

Eset leírása – A Balaton Magyarország kiemelt turisztikai térsége, amelynek népszerűsége erősen szezonális és számos külső tényezőtől függ. A klímaváltozás, a globális utazási trendek, vagy épp egy világjárvány jelentős hatással lehet a régió turizmusára. Az elmúlt évek tapasztalatai rávilágítottak arra, hogy a hagyományos tervezési módszerek nehezen tudnak lépést tartani a gyorsan változó körülményekkel. A Balatonnak sürgősen lépnie kell az okos turizmus irányába: modern, digitális megoldások bevezetésével növelheti a vendégek elégedettségét és a régió versenyképességét, így biztosítva hosszú távú sikerét (Economx, 2025). Ez megteremti az igényt egy dual AI rendszer alkalmazására a desztináció-menedzsmentben, amely segít innovatív és proaktív módon kezelni a kockázatokat.

Működés és innováció – Az intelligens desztináció-tervezés alatt azt értjük, hogy a régió fejlesztési és marketing döntéseit adatalapon, előrejelzésekre és szimulációkra támaszkodva hozzák meg. A Balaton esetében a dual AI rendszer feladata a turisztikai szezon meghosszabbítása és a kereslet kiegyenlítése. A generatív AI szintetikus adatokat tud előállítani (látogatók száma egy szokatlan eseménysorozat vagy egy negatív médiavisszhang esetén). Ezek a forgatókönyvek (4. táblázat) segítenek megérteni, hol vannak a gyenge pontok.

A forgatókönyvekre épülve a prediktív AI különböző adaptív stratégiákat dolgozhat ki: dinamikus árkedvezmények, utalványrendszer, marketingkampányok, erőforrás-átcsoportosítás, alternatív programok kommunikációja, válságkezelési protokollok aktiválása, együttműködést más véleményvezérekkel, pozitív tartalmak generálása, speciális kedvezmények a különböző korosztályú turisták visszacsábítására, célpiac diverzifikálása.

Eredmények és hatások – A 4. táblázat összefoglal néhány lehetséges válságforgatókönyvet és az azokra adható AI-alapú válaszstratégiákat a Balaton térségében. A Balaton esetében egy ilyen rendszer hozzájárulhat ahhoz, hogy a régió proaktívan, innováció-vezérelt módon készüljön a jövő kihívásaira, és minimalizálja a kiszolgáltatottságát a külső sokkokkal szemben.

**4. táblázat: A lehetséges válságforgatókönyvek és dual AI-stratégiák a Balatonnál**

| Válság-forgató-könyv | Várható hatás a turizmusra | Generatív modell szimuláció | Adaptív válaszstratégia (prediktív AI) |
|---|---|---|---|
| **Kereslet hirtelen visszaesése (pl. globális járvány)** | Turisták számának drasztikus zuhanása egy szezonban; bevételkiesés és kihasználatlan kapacitások | Idősoros adatok szintetikus generálása a sokk és elhúzódó kilábalás trendjeinek modellezésére; több forgatókönyv (optimista, pesszimista, közepes) létrehozása Monte Carlo szimulációhoz. | Dinamikus árazás: jelentős kedvezmények a belföldi és regionális vendégeknek az azonnali keresletélénkítésért; marketing újracélzás a közeli piacokra (belföld, szomszédos országok); költségcsökkentő intézkedések bevezetése (energia- és munkaerő-megtakarítás) a válság idejére. |
| **Természeti katasztrófa (pl. árvíz, viharkár)** | Infrastruktúra sérülése, szolgáltatások átmeneti leállása bizonyos helyeken; lemondások hulláma és negatív média a károkról. | Térbeli adatok generálása (pl. kieső kapacitások térképe); ügynökalapú modell a turisták útvonalainak újratervezésére alternatív helyszínek felé; szimuláció a helyreállítás időigényére és az imázshatásra. | Erőforrás-átcsoportosítás: a sértetlen attrakciók és szállások felé terelni a vendégeket (transzfer, kedvezmények); válság-kommunikációs terv azonnali indítása (pontos tájékoztatás a helyreállításról, biztonság kiemelése); rugalmas foglalási feltételek (lemondási díj elengedése, halasztott utazási lehetőség) a vendégmegtartás érdekében. |
| **Reputációs veszteség (pl. influenszer-botrány)** | Egy fontos szegmens (pl. fiatalok) elpártolása; csökkenő foglalások a közösségi médiában terjedő negatív élmények miatt. | Közösségimédia-adatok és keresési trendek elemzése a negatív hír elterjedésének modellezésére; szentimentelemzés szintetikus generálása különböző narratívákra, hogy melyik ellenkampány lenne leghatásosabb. | Intenzív PR és marketing: pozitív kampány indítása a hírnév helyreállítására (elégedett vendégek történeteinek kiemelése); influenszer együttműködések a kedvezőtlen benyomások ellensúlyozására; új termék-fejlesztés vagy élmény a fiataloknak, amit promócióval támogatnak (pl. fesztivál, technológiai attrakció), hogy visszanyerjék a figyelmüket. |

**Forrás: Szerzői szerkesztés, 2026.**

## 5. Következtetések és javaslatok

A kutatási eredmények triangulációját és validitását szolgálja a módszertan kulcsa, vagyis az előzetes feltételezés szerint a módszerek kombinálása erősebb bizonyítékokat ad, mint bármelyik különálló módszer önmagában. A statisztikai elemzések által felfedett trendeket az interjúk és esettanulmányok lehetőség szerint megerősítették vagy árnyalták. Például, ha a felmérés szerint egy adott AI-funkciót a vezetők előnyösnek ítélnek, az interjúkból kiderült, hogy konkrét példák is alátámasztják ezt (pl. AI alapú kockázatfigyelés bevezetése). A különböző adatforrások összevetése javította a kutatás belső konzisztenciáját és külső validitását. Minden lépésnél figyelemmel voltunk az etikai és módszertani transzparenciára: a minta kiválasztásának elvei dokumentáltak. A bemutatott KMDS-keret koncepcionális jellegű: működőképességének igazolása még számos jövőbeli empirikus vizsgálatokat igényel.

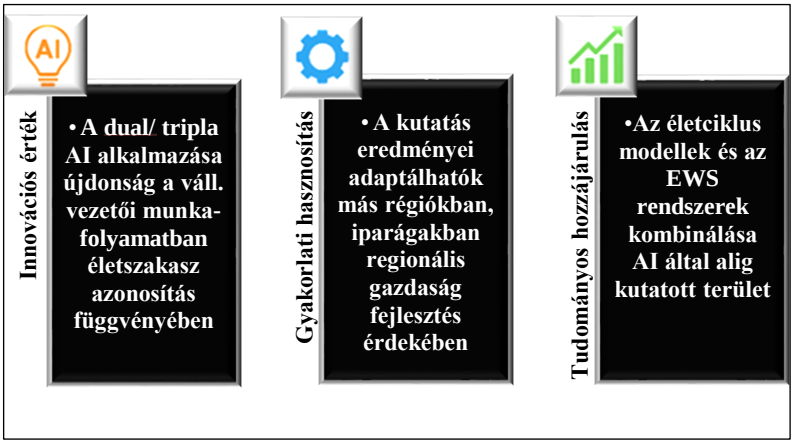
A kutatás eredmények megerősítik, hogy a hibrid intelligencia koncepciója újszerű lehetőségeket teremt a vezetői tudás menedzselésében. A három AI-modul együttesen olyan hibrid rendszert alkot, amely támogatja a folyamatos szervezeti tanulást és innovációt: a gépi előrejelzések révén új lehetőségek tárulnak fel, az AI-magyarázatok (XAI/SHAP) pedig megalapozzák a döntések átláthatóságát és a vezetői bizalmat.

Gyakorlati szempontból a vállalatoknak érdemes beruházniuk az AI-alapú infrastruktúrába – ideértve az IoT- és 5G-fejlesztéseket is – annak érdekében, hogy kihasználhassák az Edge AI előnyeit. Emellett kiemelten fontos a vezetők tréningje és a belső kultúra fejlesztése: a tripla AI modell keretében a menedzserek érzelmi és etikai intelligenciájának erősítése támogatja a felelős és fenntartható vezetést. A bevezetés során ajánlatos figyelembe venni az etikai és adatvédelmi szempontokat, ahogyan a biztonság és magyarázhatóság is fontosságát az AI-KM integrációjában.

Javasoljuk az említett AI rétegeket beépíteni az adizesi életszakaszokba. A 4. generációs AI-EWS-t érdemes már elkezdeni kiépíteni a cégműködés kezdetén. Mindenképpen javasoljuk alkalmazni az Edge AI-t és a tripla AI-t a tartós (mérsékelt) növekedés szakaszában, míg a dual AI-t a hanyatlásban. Két újabb AI (turnaround és a turnaround controlling)-réteg segíti mind a növekedési pályán tartást (különösen szakaszváltásnál), mind

pedig a növekedési pályára való visszatérést (látens és kifejlett válság időszakában).

Az eredmények alátámasztják a kutatásunk innovációs értékét, gyakorlati hasznosítását és tudományos hozzájárulását (7. ábra). A kutatásunk innovációs értékét az jelenti, hogy a dual/tripla AI alkalmazása újdonság a vállalatvezetői munka-folyamatban az életszakasz azonosítás függvényében. A gyakorlati hasznosítás abban van, hogy a kutatás eredményei adaptálhatók más ágazatokban, iparágakban a regionális gazdaságfejlesztés érdekében. A tudományos hozzájárulás az, hogy az életciklus modellek és az EWS kombinálása AI által alig kutatott terület.




**7. ábra: A kutatás innovációs értéke, gyakorlati hasznosítása és tudományos hozzájárulása**
**Forrás: Szerzői szerkesztés: 2026.**

A jövőben célszerű tovább vizsgálni a hibrid intelligencia hosszú távú hatásait a szervezeti kultúrára és tudásátadásra. Érdemes mélyebben elemezni azt, hogy hogyan segítheti a generatív AI és a collaborative AI beépítése az Adizes-modell egyes szakaszainak kezelését.

További irány lehet a különböző iparágak összehasonlítása: Hogyan lehet alkalmazni a dual/tripla AI-t más szektorokban (pl. gyártás, pénzügy), vagy a regionális különbségek feltárásában a magyar kkv-k körében. Érdemes kutatni további transzparencia- és etikai keretrendszereket is annak érdekében, hogy megalapozottabb legyen a bizalom az AI rendszerekben. A jövőbeni kutatásokban bővíteni lehet a mintavételt

külföldi vállalatokkal és nemzetközi adatbázisokkal, valamint kísérleti beavatkozásokkal érdemes tesztelni az AI-EWS-ek valós hatékonyságát.

## Összefoglalás

A tanulmányunk rámutatott arra, hogy a hibrid intelligencia – a dual/tripla AI és az Edge AI – jelentős mértékben képes gazdagítani a vállalatvezetői tudásmenedzsmentet. A vegyes módszertanú kutatás elért eredmények alapján a gépi és emberi intelligencia integrálása új szintre emeli az információfeldolgozást és döntéstámogatást. A vállalati EWS rendszerek életciklus-alapú beépítése tovább növeli a vállalatok rezilienciáját. A magyar mintavételben kimutatott összefüggések (mint az ESG és az előrejelzések megnövelt pontossága) alátámasztják, hogy az AI bevezetése a fenntarthatóságot és a szervezeti tanulást is ösztönzi.

A tanulmány bizonyítékokat szolgáltatott arra, hogy a vállalatvezetői tudásmenedzsment „új nyelve" a hibrid AI: a gép gyorsít, az ember irányt szab, és az EWS időben szól. A KMDS keretrendszerrel a vállalatok képesek hatékonyabban és proaktív módon felkészülni a jövő kihívásaira és lehetőségeire.

## Irodalomjegyzék

Adadi, A., & Berrada, M. (2018). Peeking inside the black box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Adizes, I. K. (1992): *Vállalatok életciklusai:* hogyan és miért növekednek és halnak meg a vállalatok, és mi az ezzel kapcsolatos teendő? Budapest: HVG Rt., 350. ISBN: 963752505x

Adizes, I. (2023): Vállalatok életciklusai. Budapest: HVG Könyvek Kiadó, 480. ISBN: 9789635652204

Ahdadou, M., Ahdadou, A., Aajly, A., & Tahrouch, M. (2025). Artificial intelligence in corporate boards: a dual-dimensional framework for integration across autonomy and structural levels. *Data Science and Management*. Online publication. https://doi.org/10.1016/j.dsm.2025.12.001

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Chatterjee, S., & Byun, Y. C. (2023). A synthetic data generation technique for enhancement of prediction accuracy of electric vehicles demand. *Sensors*, *23*(2), 594. DOI: 10.3390/s23020594

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fejes J. K., & Katits, E. É. (2025a). CityMindX: AI-based mobility and infrastructure optimisation in smart cities. *International Journal of Science and Social Science Research*, *3*(3), 146–168. https://doi.org/10.5281/zenodo.17960728

Fejes J. K., & Katits, E. É. (2025b). Developing early warning systems in the era of company crisis prevention. *International Journal of Science and Social Science Research*, *3*(2), 136–159. https://doi.org/10.5281/zenodo.17088464

Fejes J. K., & Katits, E. É. (2025c). *Double AI simulations supported by synthetic data in smart tourism innovation for corporate foresight*. In: Dar, H., Singh, K., & George, B. (ed.): Tourism technology, sustainability, and local community empowerment. London: Rotledge, 240. Chapter 11. ISBN: 9781041117261

Grant, R. M. (1996). Toward a knowledge-based theory of the firm. *Strategic Management Journal*, *17*(S2), 109–122. https://doi.org/10.1002/smj.4250171110

Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, *37*(1), 388–427. https://doi.org/10.1016/j.ijforecast.2020.06.008

Huang, G., Xu, C., Sun, X., & Zhang, S. (2020). ROC and AUC with applications in AI prediction reliability. *Journal of Artificial Intelligence Research*, 69, 1–23. DOI: https://doi.org/10.1613/jair.1.11234

Katits, E. (2024): VÁLLALATI TURNAROUND PÉNZÜGYEK a fordulatkezelési kutatások tükrében és a fenntartható gazdálkodás érdekében. Budapest: Akadémia Kiadó, MeRSZ kiadvány epub. ISBN: 9789636640811

Kantabutra, S., & Ketprapakorn, N. (2021). Toward an Organizational Theory of Resilience: An Interim Struggle. *Sustainability*, 13(23), 13137. https://doi.org/10.3390/su132313137

Lim, J., & Hwang, J. (2025). Exploring trends and topics in hybrid intelligence using keyword co-occurrence networks and topic modelling. *Futures*, 167, 103550.

https://doi.org/10.1016/j.futures.2025.103550

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2, 56–67. https://doi.org/10.1038/s42256-019-0138-9

Muschelli, J. (2019). ROC and AUC with a binary predictor: a potentially misleading metric. *Journal of Classification, 37*(4):696-708. DOI:10.1007/s00357-019-09345-1

Olan, F., Arakpogun, E. O., Suklan, J., Nakpodia, F., Damij, N., & Jayawickrama, U. (2022). Artificial intelligence and knowledge sharing: Contributing factors to organizational performance. *Journal of Business Research, 145*, 605–615. https://doi.org/10.1016/j.jbusres.2022.03.008

Quinn, R. E., & Cameron, K. (1983). Organizational life cycles and shifting criteria of effectiveness: Some preliminary evidence. *Management Science, 29*(1), 33–51.

https://doi.org/10.1287/mnsc.29.1.33

Rezaei, M. (2025). Artificial intelligence in knowledge management: Identifying and addressing the key implementation challenges. *Technological Forecasting and Social Change*, 217, Article 124183. https://doi.org/10.1016/j.techfore.2025.124183

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198

Singh, N. (2022). Developing Business Risk Resilience through Risk Management Infrastructure: The Moderating Role of Big Data Analytics. *Information Systems Management*, 39(1), 34–52. https://doi.org/10.1080/10580530.2020.1833386

Song, Y., Mi, L., Bian, Z., Tu, W., & He, J. (2025). How does artificial intelligence impact corporate ESG performance? The catching-up effect of digital technological innovation. *Journal of Innovation & Knowledge, 10*(6), 100843. https://doi.org/10.1016/j.jik.2025.100843

Tanaka, K., Higashide, T., Kinkyo, T., & Hamori, S. (2025). A multi-stage financial distress early warning system: Analyzing corporate insolvency with random forest. *Journal of Risk and Financial Management*, *18*(4), 195. https://doi.org/10.3390/jrfm18040195

Zahedi, J., Salehi, M., & Moradi, M. (2023). Identifying and classifying the financial resilience measurement indices using intuitive fuzzy

DEMATEL. *Benchmarking: An International Journal*, 30(4), 1300–1321. https://doi.org/10.1108/BIJ-07-2021-0395

Zhang, Y., Tan, W. H., & Zeng, Z. (2025). Tourism demand forecasting based on a hybrid temporal neural network model for sustainable tourism. *Sustainability*, *17*(5), 2210. https://doi.org/10.3390/su17052210

**Elektronikus forrás**

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system.* https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf?searchterm=Shrinkage
Letöltés ideje: 2025. 12. 07.

Economx (2025). https://www.economx.hu/magyar-gazdasag/kokorszaki-balaton-innovacio-fejlodes-technologia-generaciok-turizmus.804491.html#:~:text=%C3%89ppen%20ez%C3%A9rt%20a%20Balatonnak%20s%C3%BCrg%C5%91sen,turizmus%20piac%C3%A1n%20%E2%80%93%20%C3%A1llap%C3%ADtott%C3%A1k%20meg
Letöltés ideje: 2025. 12. 07.

MTÜ (2019). https://mtu.gov.hu/cikkek/strategia/
Letöltés ideje: 2025. 12. 07.

Ruiz-Martín, C. (2018). *A framework to study the resilience of organisations: a case study of a nuclear emergency plan.* Ottawa: Carleton University, 267, Thesis Paper. https://carleton.scholaris.ca/server/api/core/bitstreams/d65b2a33-a27e-426d-8c4b-46ea4b3dcaef/content
Letöltés ideje: 2025. 12. 07.

TechTarget (2025). *What is the ISO 31000 Risk Management standard?* https://www.techtarget.com/searchsecurity/definition/ISO-31000-Risk-Management#:~:text=ISO%2031000%20provides%20a%20framework,a%20consistent%20and%20structured%20approach
Letöltés ideje: 2025. 12. 07.

# SZERZŐK / AUTHORS