Bus Inf Syst Eng
https://doi.org/10.1007/s12599-025-00942-6

DISCUSSION

High-Risk Artificial Intelligence

Ali Sunyaev · Alexander Benlian · Jella Pfeiffer · Ekaterina Jussupow · Scott Thiebes · Alexander Maedche · Joshua Gawlitza

1 Introduction

Ali Sunyaev

Artificial intelligence (AI) technologies are increasingly permeating and transforming all walks of life, unlocking new potential for efficiency, automation, and innovation across most industry sectors (Yang et al. 2024). However, their implementation is not without significant challenges and risks. Integrating AI into information systems (IS), which are per se socio-technical systems, introduces risks that extend beyond purely technical concerns (Jussupow et al. 2021; Maedche et al. 2019; Pfeiffer et al. 2023; Wiener et al. 2023).

Given the immense opportunities and significant risks associated with AI, the topic has increasingly drawn attention from academia, industry, and legislators. As such, in April 2024, following three years of trilateral negotiations between the European Commission, the European Council, and the European Parliament, the European Union introduced a landmark regulation on AI. This so-called "AI Act" aims to establish the first comprehensive legal framework governing the use of AI technologies in the European Union (European Commission 2021; European Parliament 2024). A central component of the AI Act is its definition of four risk classes for AI systems (see Fig. 1).

With this risk classification, the AI Act puts particular emphasis on the regulation of high-risk AI systems (i.e., those AI systems that could impact and endanger the health, safety, or fundamental rights of individuals). These systems are subject to rigorous oversight, including mandatory internal conformity assessments by the providers and, in special cases, external reviews by notified bodies (European Parliament 2024; Hupont et al. 2023). Despite such regulatory measures being put forward, numerous questions remain open, especially regarding the integration of high-risk AI in IS.

While high-risk AI in IS hold tremendous promise for advancements in business and society, they also bring forth complex issues, highlighting technical and societal dilemmas and the need for balanced trade-offs to resolve such dilemmas (Thiebes et al. 2021). As a socio-technical discipline that integrates insights from computer science, management, and other domains to drive technological innovation in business and society, IS scholarship and practice play a critical role in addressing the challenges surrounding the responsible and sustainable development

A. Sunyaev (✉)
School of Computation, Information and Technology, Technical University of Munich (TUM), Campus Heilbronn, Bildungscampus 2, Heilbronn, Germany
e-mail: sunyaev@tum.de

A. Benlian
Chair for Information Systems & Electronic Services, Technical University of Darmstadt, Hochschulstraße 1, Darmstadt, Germany

J. Pfeiffer · A. Maedche
Institute for Information Systems (WIN), Karlsruhe Institute of Technology (KIT), Kaiserstraße 89, Karlsruhe, Germany

E. Jussupow
Chair of Information Systems, Technical University of Darmstadt, Hochschulstraße 1, Darmstadt, Germany

S. Thiebes
Department of Economics and Management, Karlsruhe Institute of Technology (KIT), Kaiserstraße 89, Karlsruhe, Germany

J. Gawlitza
InformMe GmbH, Weihenstephaner Straße 12, Design Offices, Munich, Germany

Published online: 09 May 2025

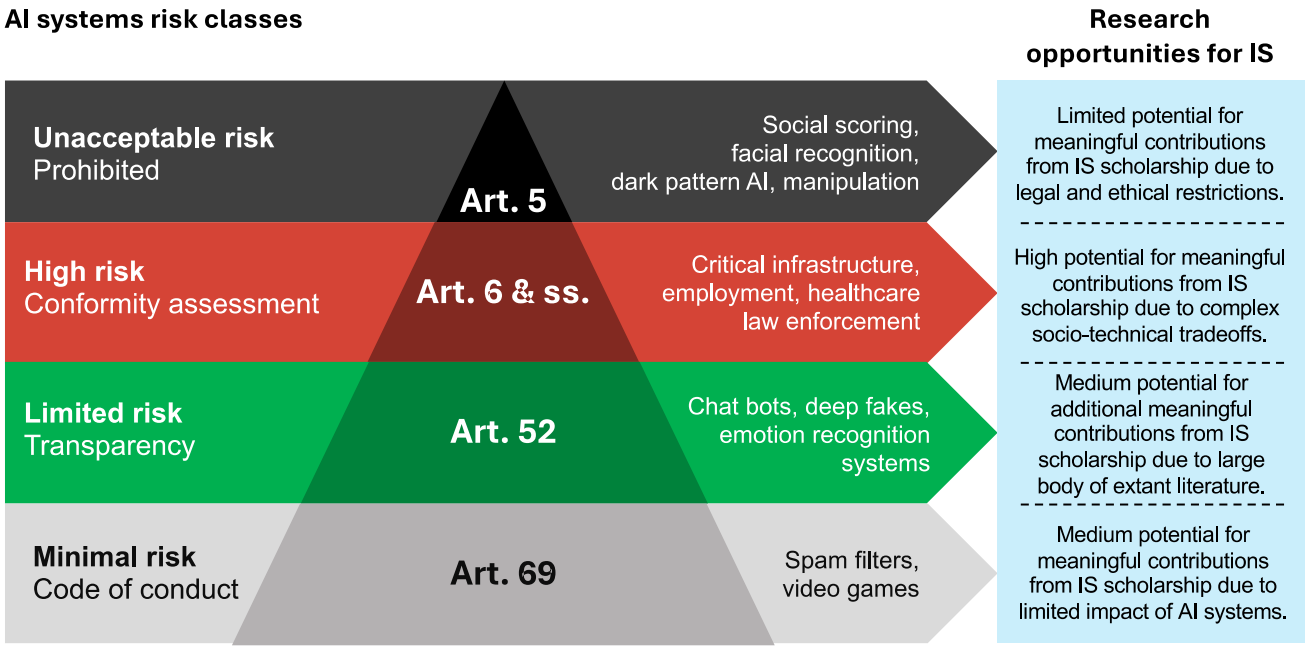

Fig. 1 The AI Act's risk-based approach to classifying AI systems (adapted from Edwards (2022) as cited in Tranberg (2023)) and emerging research opportunities for IS scholarship

and deployment of high-risk AI in IS and navigating emerging dilemmas or even trilemmas (cf. Fig. 2). For example, when balancing AI fairness with the need to collect sensitive data like age, gender, and ethnicity to develop more fair AI, as well as with potential impacts on the performance of such AI, all at the same time (Dolata et al. 2022). Likewise, from an academic point of view, high-risk AI offers substantial research opportunities for IS scholarship to make meaningful contributions, especially when compared to the other AI risk classes defined in the AI Act (cf. Fig. 1).

Against the backdrop of IS' unique position in addressing open challenges about high-risk AI, as well as the significant research potential high-risk AI presents for the field, the Business Information Systems Engineering (BISE) community organized a panel discussion on the

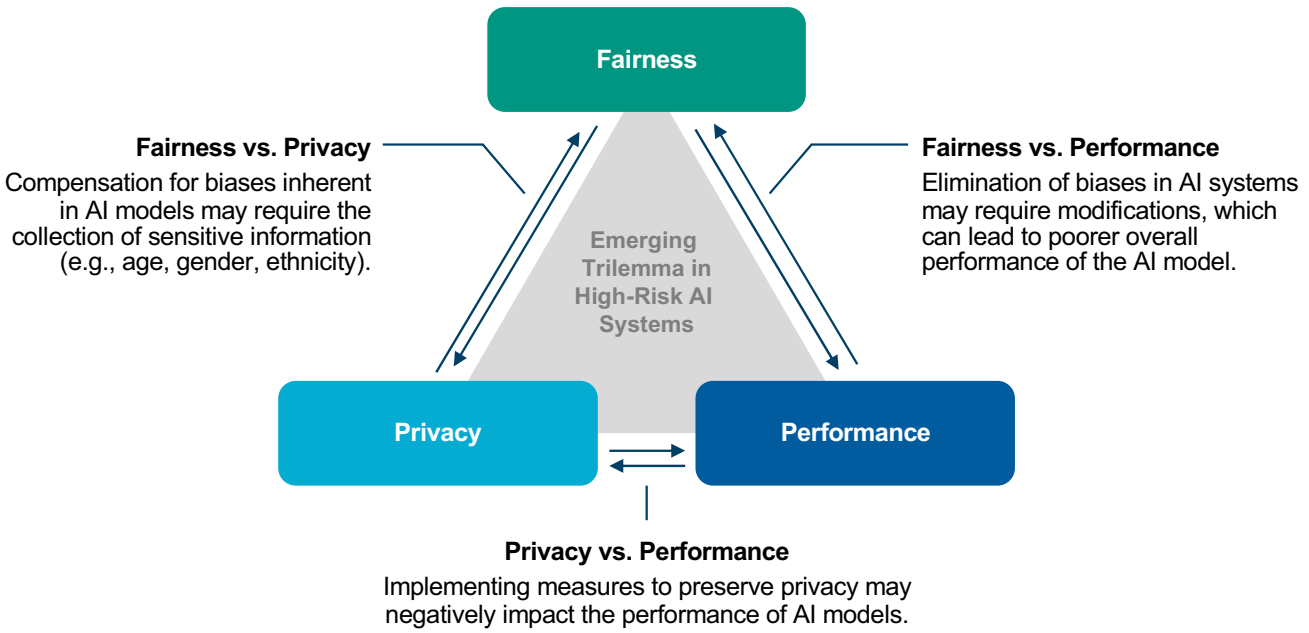

Fig. 2 Example of an emerging trilemma in high-risk AI systems

topic at the 19th International Conference on Wirtschaftsinformatik (WI 2024) in Würzburg on September 18, 2024. The panel brought together experts from academia and industry. It focused on IS' role in capitalizing on the opportunities while at the same time mitigating the risks of high-risk AI. Panelists explored key questions, including how high-risk AI systems should be understood from an IS perspective, because it is essential that IS as a discipline clarifies its understanding of high-risk AI systems, considering whether this aligns with or diverges from legal classifications. Panelists also examined the specific opportunities and risks posed by high-risk AI, the dilemmas that arise from implementing high-risk AI in IS, and how the field's unique socio-technical position can be leveraged to address these challenges. By doing so, IS as a discipline can contribute to resolving identified dilemmas and fostering the responsible design and deployment of high-risk AI.

With this discussion at BISE, we aim to continue the productive dialogue on the opportunities and challenges of high-risk AI in IS that was initiated by the BISE community at WI2024. Accordingly, in his contribution to this discussion, Alexander Benlian examines pertinent tradeoffs in high-risk AI systems and explores how IS researchers can contribute to their resolution. Jella Pfeiffer emphasizes the importance of leveraging existing IS expertise and explores how IS researchers can contribute to fulfilling regulatory requirements for high-risk AI systems, while Ekaterina Jussupow considers the role of human oversight in high-risk AI systems. Scott Thiebes elaborates on how high-risk AI in critical digital infrastructures can be a path for IS scholarship to navigate the disciplines ongoing identity crisis, whereas Alexander Maedche discusses the risks and potentials of emotion recognition systems in companies and educational institutions. In the final contribution to this discussion, Joshua Gawlitza explores the significance of high-risk AI for start-ups and SMEs.

Through this discussion, which spans a wide array of topics relevant to BISE readers, we hope to spark further dialogue and deepen the exchange on high-risk AI within the BISE community.

2 How Information Systems Researchers Navigate Tradeoffs and Dilemmas in High-Risk Artificial Intelligence

Alexander Benlian

The promise of high-risk AI is both thrilling and terrifying. Through its unique features, such as autonomy, learnability, and context-awareness (Berente et al. 2021; Schuetz and Venkatesh 2020), it has the potential to revolutionize sectors such as healthcare, finance, and transportation, solving problems that were previously deemed insurmountable. But with great power comes great responsibility. As these AI applications become more powerful, organizations are confronted with a thorny question: *How can we harness the full potential of AI in terms of performance, utility, or transparency without sacrificing fairness, privacy, or security?* This question is not just theoretical – it's a dilemma that cuts to the very core of the ethical and responsible design, development, deployment, use, and governance of AI. The IS discipline, with its socio-technical history and holistic view of technology, business, and society, is uniquely equipped to address these tradeoffs. This contribution examines the critical tradeoffs inherent in high-risk AI, raises vexing questions that persist in this domain, and discusses how the IS discipline can help address these challenges.

2.1 Performance Versus Fairness: Balancing Accuracy and Equity

Imagine an AI system used by a bank to assess loan applications. The model has been trained on historical lending data, which shows that past loan approvals were disproportionately granted to applicants from certain socioeconomic backgrounds, while applicants from marginalized communities were more frequently denied loans. If the AI system simply replicates these patterns, it may optimize for predictive accuracy – approving loans for individuals who, based on historical data, had lower default rates. However, this approach reinforces systemic bias and financial exclusion. This sets up the classic *dilemma of performance vs. fairness*. This issue is not just limited to lending. In my work with banks, insurance, and healthcare companies, I often experience how AI systems used in recruitment decisions, wealth management, underwriting, and medical diagnoses face the same predicament. If we tweak the AI model to be "fairer" – for example, by adjusting approval thresholds to ensure more equitable lending practices – we often end up sacrificing some level of predictive accuracy. As a result, organizations are stuck in a lose-lose situation: focus on accuracy and model performance and risk perpetuating disparities, or prioritize fairness and risk increased uncertainty in lending decisions.

What can the IS discipline offer in this context? IS researchers can, for example, advocate, develop, and evaluate the use of *fairness-aware learning methods and models* that do not treat performance and fairness as mutually exclusive goals. By integrating fairness constraints directly into optimization algorithms, IS can help create models that balance both aspects more effectively (Pfeiffer et al. 2023). But does this solve the problem entirely? Not quite. Fairness is subjective, and what seems

fair to one group might not appear fair to another (Teodorescu et al. 2021). To address this, IS research can play a critical role by developing context-specific fairness metrics, which allow organizations to define fairness within the specific legal, social, and ethical dimensions of their domain (Corbett-Davies et al. 2023). Additionally, *fairness auditing processes and tools* can provide ongoing assessments of fairness as AI systems evolve, addressing the challenge that fairness is a "moving target" and ensuring that these systems adapt to changing societal conditions (Simbeck 2024; Wilson et al. 2021). So, while fairness may never be entirely universal, IS can help create context-sensitive solutions, while also contributing to the development of effective governance models.

2.2 Privacy Versus Utility: The Data Dilemma

Picture a healthcare system that uses AI to predict disease outbreaks. Such a system would require access to vast amounts of personal health data, including patient histories, genetic information, and behavioral patterns. On the surface, the potential economic and social utility is enormous – early detection could save countless lives. But here's the catch: *At what cost to privacy?* The more data we collect, the more accurate the predictions become. Yet, every additional data point also increases the risk of privacy breaches, misuse, or even malicious exploitation. A breach could expose intimate health data and lead to identity theft or medical fraud. Collecting behavioral data to feed AI systems could also result in an invasive level of monitoring. In fact, we have already witnessed this dilemma in the realm of wearable health technology (Spiekermann et al. 2022). Devices like smartwatches and fitness trackers collect a wealth of personal data, including heart rate, sleep patterns, and even stress levels. While the data can be used to improve individual health outcomes – alerting users to potential cardiac issues or encouraging healthier lifestyle choices – the widespread collection of such intimate information has raised serious privacy concerns related to data breaches or unauthorized sharing of this information. As these devices become more advanced and capable of tracking increasingly detailed aspects of our lives, the tradeoff between utility and privacy continues to intensify.

How can IS contribute to addressing this data dilemma? One solution lies in *privacy-preserving technologies such as federated machine learning*, where models are trained across multiple decentralized devices holding local data samples, without transferring data to a central server (Kaissis et al. 2020; Sánchez et al. 2024). This approach may help organizations to benefit from data-driven insights while maintaining individual privacy. But federated learning is not a silver bullet – it is complex, resource-intensive, and often leads to reduced predictive accuracy, which can diminish its (economic) utility by limiting decision-making effectiveness, operational efficiency, and business value compared to traditional data aggregation methods. Thus, we are back to square one: *Is it possible to achieve both privacy and utility or will organizations always have to choose?* IS research could also increase their efforts to explore user-centric approaches to privacy, such as *personal data vaults and other forms of inverse transparency* (Adam et al. 2024; Gierlich-Joas et al. 2024), where users have more control over how their data is shared and what level of privacy risk they are willing to accept. These approaches would empower users and level the playing field while allowing organizations to still extract valuable insights, achieving more balance between privacy and utility.

2.3 Transparency Versus Security: An Ongoing Struggle

In high-risk AI systems like autonomous driving or financial decision-making, transparency is crucial for building trust and ensuring accountability (Thiebes et al. 2021). However, the more transparent an AI system becomes, the more vulnerable it is to security threats. For example, hackers could exploit the transparency of a financial AI model to reverse-engineer it, identify vulnerabilities, and engage in manipulative activities such as adversarial attacks (Cram et al. 2024). This creates another dilemma: while transparency is essential for trust and regulatory compliance, it can inadvertently expose the system to security risks. This leaves organizations in a precarious position: *Should they keep their AI models opaque and risk losing public trust, or should they be fully transparent and compromise security?* This dilemma is not hypothetical. Consider autonomous vehicle manufacturers, who must provide transparency on how their AI systems make real-time decisions during driving, especially when accidents occur. Regulatory bodies (e.g., through the EU AI Act) and the public demand detailed explanations to ensure safety and accountability. However, full transparency about how the AI navigates complex driving situations could allow hackers to identify weaknesses and exploit them to interfere with vehicle systems, creating serious safety risks. This leaves manufacturers in a bind: *how much transparency is too much?*

The IS field can play a pivotal role here in advancing the development of *hybrid explainability models* that combine local explainability (like LIME and SHAP) with global model transparency (Brasse et al. 2023). While local methods provide insights into individual decisions, they often fall short when applied to deep learning models that function as black boxes. By combining local methods with global interpretability techniques – such as decision

boundary and uncertainty visualizations or feature importance aggregations across decisions – IS researchers can provide a more holistic understanding of how deep learning models function without exposing proprietary or sensitive details. Unlike computer science, which primarily focuses on the theoretical and algorithmic aspects of explainability, IS research is uniquely positioned to ensure (hybrid) explainability models' practical viability by focusing on real-world deployment, usability across stakeholder groups, and seamless integration into organizational decision-making processes. Another approach IS researchers can pursue is promoting explainable AI techniques with *embedded security protocols*. This involves creating explanation mechanisms that can dynamically adjust the level of transparency based on the (critical) audience and context (Kemper and Kolkman 2019). For example, a financial AI system might provide detailed technical explanations to internal auditors and regulators, but a simplified, abstracted version to end users and the general public. These hybrid and audience-adaptive approaches would help strike a balance between transparency and security.

2.4 Striking the Balance in High-Risk AI

Returning to the initial question – *How can we harness the full potential of AI in terms of performance, utility, or transparency without sacrificing fairness, privacy, or security?* – the answer is anything but straightforward. The dilemmas inherent in high-risk AI are complex and multifaceted, requiring organizations to make tough choices. However, the IS discipline can serve as a compass, guiding these choices with tools and (governance, process) frameworks that help balance competing priorities. While no one-size-fits-all solution exists, IS can be a strong and confident contributor toward responsible high-risk AI by promoting dialogue, innovation, and ethical decision-making. In particular, interdisciplinary collaboration is essential. From my work at the Center for Responsible Digitality (ZEVEDI, www.zevedi.de), I have seen firsthand that only by bringing together diverse knowledge from law, ethics, computer science, and industry can we build frameworks that are both technically sound and aligned with societal values.

That said, beyond its core expertise in designing and assessing the impact of IS, the IS field must adopt a more proactive approach by providing normative guidance. This could involve engaging with policymakers and political stakeholders to propose regulatory frameworks for shaping AI governance (Pfeiffer et al. 2024). IS should embrace its responsibility not just to study the world, but to actively contribute to its improvement – particularly when dealing with the tradeoffs in high-risk AI. Ultimately, the question we face is not whether these tradeoffs can be solved entirely, but how we can navigate them responsibly and effectively over time.

3 Harnessing Information Systems' Expertise for High-Risk Artificial Intelligence Systems

Jella Pfeiffer

Many of the requirements for managing high-risk AI systems have been central to BISE research for years. The expertise of IS researchers can significantly contribute to the conceptualization and implementation of solutions necessary to meet these requirements and emerging regulations thereof. As IS researchers, it is our critical responsibility to identify the unique characteristics of high-risk AI systems compared to conventional IS, to adapt existing approaches to address these distinctions, and to share our specialized knowledge with companies, legislators, and policymakers. In the following, I will present specific examples illustrating how IS researchers can help meet regulatory requirements and reduce and manage risks associated with high-risk AI systems.

The EU AI Act outlines a set of requirements for managing high-risk AI systems, offering companies guidance on the measures needed to mitigate associated risks. These requirements are specified in Sect. 2, Chapter III, beginning with Article 9. For example, the Act mandates the establishment, implementation, documentation, and maintenance of a risk management system for high-risk AI systems (Article 9, EU AI Act). This system must then function across the entire lifecycle of the AI system (Article 9, Paragraph 2).

International standards such as the ISO 31000:2018 Risk Management and the COBIT 5 framework are widely employed in risk management. However, neither is specifically tailored to address the complexities of AI applications. AI systems often present unique challenges, such as opacity and dependence on large datasets, that may fall outside the direct control of system operators (Tjoa et al. 2022). Recent initiatives have sought to bridge this gap by adapting established frameworks to AI-specific contexts (e.g., National Institute of Standards and Technologies (NIST) 2023; Tjoa et al. 2022). Some researchers have concentrated on designing AI-specific risk assessment tools that align closely with established IS risk management principles (Nagbøl et al. 2021). However, despite these advancements, research specifically focused on adapting our existing risk management knowledge to the distinct characteristics and challenges of AI systems remains limited.

Following Article 9, the AI Act outlines further requirements, such as data governance and management

practices (Article 10) for datasets used in training, validation, and testing. These emphasize that data must be relevant, representative, and free from errors or biases that could compromise fairness. IS researchers have begun exploring fairness in AI algorithms, as reviewed in Kordzadeh and Ghasemaghaei (2022) and Dolata et al. (2022). Much of the existing work takes a technical perspective, such as Caton and Haas (2024), who systematically discuss approaches for mitigating biases in AI algorithms. However, addressing fairness from an interdisciplinary and socio-technical perspective is increasingly recognized as essential (Dolata et al. 2022; Pfeiffer et al. 2023). This emphasis on socio-technical approaches presents an opportunity for the IS field, which has a long-standing tradition of integrating technical and social perspectives. Moreover, concepts such as fairness also encompass broader ethical and legal dimensions, encouraging us as IS researchers to further expand our scope and engage with these critical aspects to contribute meaningfully to these broader, impactful discussions.

While several additional requirements for managing high-risk AI systems could be discussed – such as technical documentation (Article 11), record-keeping (Article 12), transparency and the provision of information to deployers (Article 13), and accuracy, robustness and cybersecurity (Article 15) – I choose to focus on one that, in my judgment, merits particular attention: human oversight (Article 14). This article stipulates that high-risk AI systems must include tools for human oversight tailored to their specific risks, autonomy, and context. These tools should then enable users to understand the system's capabilities and limitations, identify and address issues, avoid over-reliance (automation bias), and – if necessary – assess when to discontinue its use.

IS research has developed nuanced perspectives on the human-in-the-loop notion and human–machine learning augmentation, and, thus, interesting suggestions have been made for problems that may arise from AI usage. Teodorescu et al. (2021), for example, examine the issue of algorithmic fairness, proposing a typology of augmentation based on two dimensions: the difficulty of achieving fairness and the locus of decision-making in human–machine learning partnerships. They also propose managerial strategies for achieving fairness across the resulting types emerging from the typology. Similar approaches could be extended to other requirements for ensuring the trustworthiness of high-risk AI systems. Their analysis also highlights how certain particularities of machine learning challenge IS theories. For instance, addressing unfairness in AI systems often requires full retraining of a model if it produces unfair decisions rather than only implementing incremental changes, as is common in many IS solutions. Other challenges include user trust in opaque systems and the self-learning mechanisms of AI algorithms, which undermine traditional theories of technology-in-use. Baird and Maruping (2021), for instance, build on agent interaction theories, introducing the concept of delegation to account for the increasing role of IS artifacts to act as independent agents.

In summary, I hope to inspire IS researchers to harness the rich body of IS knowledge to advance the discussion on implementing regulations for managing high-risk AI systems, thereby contributing to the broader goal of fostering digital responsibility (Trier et al. 2023). Responding to the call by Butler et al. (2023) in their recent special issue in the Journal of Information Technology, I emphasize the need for more empirical and design science studies, as current approaches are predominantly based on secondary data and literature reviews. Future research should prioritize innovative approaches that leverage our deep understanding of IS systems, ensuring effective solutions to the pressing challenges presented by high-risk AI systems.

4 The Role of Human Oversight in High-Risk Artificial Intelligence Systems

Ekaterina Jussupow

Since high-risk AI systems, such as systems supporting medical decision-making, have a significant impact on individuals, the EU AI Act (Article 14) has introduced a key requirement for human oversight, i.e., humans' ability to monitor the AI output and intervene if the AI output were potentially harmful. This article focuses on how human cognitive processes influence the evaluation of AI system outputs. Additionally, it explores the role of these processes in fostering AI literacy – specifically, the capacity to assess AI systems critically (Long and Magerko 2020; Pinski and Benlian 2024). This section will elaborate on the requirements and implications for humans using the example of AI-augmented medical decision-making, where AI systems are implemented to support medical diagnostic decisions.

The first requirement of Article 14 specifies the need to develop appropriate human–machine interfaces that enable humans to oversee the AI effectively. The human's role is to mitigate potential harms from AI, such as incorrect medical diagnoses. The EU AI Act explicitly acknowledges that humans should remain aware of potential cognitive risks when working with AI systems, especially the risk of over-relying on the system's output and overlooking system errors as a result – i.e., automation bias (Goddard et al. 2012). Overall, in the context of high-risk AI systems, the regulation advocates that human decision-makers collaborate with AI to prevent harmful effects, e.g., from misdiagnoses.

Interestingly, prior research in IS suggests that human-AI collaboration poses cognitive challenges to human decision-makers, who have to engage in reflection practices in order to critically examine provided advice (Abdel-Karim et al. 2023; Fügener et al. 2022; Jussupow et al. 2022, 2021; Lebovitz et al. 2022; Taudien et al. 2024). Further, as explanations are provided for the AI advice, they influence which factors decision-makers consider and can reinforce prior incorrect beliefs of individuals about a specific decision (Bauer and Gill 2024). Thus, collaborating with AI can be challenging and requires additional metacognitive skills, i.e., the ability to monitor and control one's own cognitive processes (Nelson and Narens 1994), in order to assess and reflect upon the systems' outputs critically. Thereby, prior work has provided some evidence that it is unclear whether humans collaborating with AI outperform AI systems on their own (Fügener et al. 2021), raising the question of who can effectively oversee AI systems in medical decision-making.

This notion has been supported by a recent meta-analysis comparing over 100 experiments to systematically test performance differences across different configurations of humans and AI systems (Vaccaro et al. 2024). The results are astonishing; the human-AI group outperforms the AI model on its own in only 42% of all considered experiments, while the same group outperforms humans alone in almost 85% of the cases. The effects are especially strong in decision tasks, a set of tasks that contains medical decision-making. However, when humans on their own were better than the AI alone, the collaboration between humans and AI systems resulted in superior performance compared to that of the AI model. These findings suggest that humans could differentiate between correct and incorrect AI decisions and, by collaborating with AI, increase their performance. However, most decision-makers could not outperform the AI model alone. Thus, it remains open which cognitive and metacognitive abilities decision-makers must possess to benefit from AI and which skills need to be trained to effectively oversee AI systems in medical and other high-risk decision-making cases (Pinski and Benlian 2024). Furthermore, it is necessary to consider how the design of medical AI systems can help to achieve better reflection activities, i.e., providing transparency about the underlying data, model, and decision-processes.

In another example from the medical domain, Frazer et al. (2024) compare AI as a standalone reader against radiologists in screening mammograms, which is only one of the many possible configurations of human-AI collaboration. In this case, the AI system again outperformed most individual readers. However, it must be acknowledged that the standard of care often consists of two to three independent professionals who evaluate a case, who then outperform the AI model in terms of sensitivity, i.e., correct detection. Yet, replacing one human reader with an AI increased the configuration's performance. The author noted, however, that in their simulation physicians tended to overrule correct decisions more frequently, and the automation bias favored the overall accuracy (p. 6).

What do these findings mean for ensuring effective oversight of these systems and their output? I would argue that it is necessary to think beyond the configuration of one human and one AI system and instead consider configurations of different AIs and humans in high-risk decisions, thereby considering what organizational practices should be implemented in the case of high-risk AI, that allow for a broad range of human-AI configurations. Furthermore, it remains to be determined how to design AI systems to effectively foster effective collaboration and enable humans to correctly identify errors in these systems – which types of explanations are beneficial, and what other information should be provided to users to ensure effective oversight? How much transparency of the underlying data is necessary to help make the best decisions? In addition, especially in high-risk contexts such as medicine, it is necessary to determine who could oversee the results of the AI: for example, would it be possible to implement an AI system to provide personalized medication dosages for patients – if the patients cannot verify whether the suggestion is correct – or should we always keep a doctor in the loop after interacting with an AI system? How can we improve medical education to ensure that the medical professionals of the future are well equipped in working with and assessing AI outputs in high-risk contexts? How can the delegation of tasks be performed and regulated in the context of high-risk decisions, i.e., would a partial delegation be acceptable to reduce the workload? How should AI involvement be communicated to patients and other stakeholders, and what effects does it have on those? All these questions need to be addressed by regulations and future research.

5 High-Risk Artificial Intelligence Systems in Critical Digital Infrastructures

Scott Thiebes

Critical infrastructures are essential systems and resources that are vital to the functioning of societies and economies (Fekete 2011). Their disruption or failure can have serious consequences, including threats to public safety, economic damage, and social unrest (Hollick and Katzenbeisser 2019). Although there is no universally agreed-upon definition of critical infrastructures, and their exact scope can vary across nations, systems in sectors such as transportation, energy, telecommunications,

finance, water and food supply, healthcare, and defense are typically classified as critical (Hollick and Katzenbeisser 2019).

Traditionally, critical infrastructures have been associated with physical, technical systems. However, rapid advancements in information technology (including AI) and the increasing pace of digitalization over the past decades have led to many IS becoming essential components of modern society (Dehling et al. 2019). Some IS have now become so critical to the functioning of our societies that they are seemingly indispensable for the maintenance of basic social functions (Dehling et al. 2019). Consequently, the concept of critical infrastructures is no longer limited to physical, technical systems but has expanded to include non-physical, digital IS. Notable examples of such critical *digital* infrastructures include the internet, global financial transaction systems like SWIFT, smart grids, or nationwide health information exchanges.

Against the backdrop of the rapid proliferation of AI in all areas of society, we are also witnessing an increasing application of AI in critical (digital) infrastructures, where it brings both enormous opportunities and significant risks. With regard to the opportunities offered by the use of AI in critical infrastructures, there are two main advantages: First, increased efficiency: AI can, for instance, optimize processes in critical infrastructures through predictive maintenance. A relatively recent example of efficiency gains through the use of AI in critical infrastructures concerns automation in the energy sector, where smart grids use AI to balance supply and demand in real time (Khan et al. 2022). Second, improved security and increased resilience: AI-based early warning systems for cyberattacks or natural disasters, for example, can react faster and more precisely than conventional approaches or human actors (Albahri et al. 2024). Likewise, in autonomous vehicles and intelligent traffic management systems, AI helps optimize traffic flows and reduce accidents and disruptions (Karim et al. 2022).

The risks associated with the use of AI in critical infrastructures primarily include: (1) Technical malfunctions. Faulty algorithms can have catastrophic consequences in critical areas such as the control of energy supply systems. In the financial sector, for example, errors in high-frequency trading algorithms have already resulted in the loss of several hundred million dollars (Harford 2012). (2) Cyberattacks. AI systems in critical infrastructures are an attractive target for attacks that can have serious consequences for public safety (de Nobrega et al. 2024). This poses an enormous risk, particularly in light of current geopolitical challenges. (3) Discrimination. Automated, AI-based decision-making processes can reinforce unintended biases. This is particularly problematic in areas such as access to public services. For example, in the UK, an AI tool used in visa processing was suspended after being accused of racial bias (Ungoed-Thomas and Abdulahi 2024) while welfare fraud detection algorithms have been criticized for disproportionately targeting vulnerable groups (Stacey 2023). (4) Lack of transparency. Black box models make it difficult to understand decisions, which can be a major problem in safety-critical applications (Rudin 2019; Wang and Chung 2022). For instance, Tesla's approach to developing autonomous taxis relies on black box AI models, making it challenging to analyze failures and ensure safety in unforeseen scenarios (Shirouzu and Kirkham 2024).

Legislators as well have recognized the opportunities and risks of using AI in critical (digital) infrastructures and the resulting need for careful consideration. Toward that end, Annex III of the AI Act recently adopted by the EU specifies the rather broad definition of high-risk AI systems from Article 6(2) by naming eight specific areas in which the use of AI systems makes them high-risk AI systems (Future of Life Institute 2024). In addition to the use of AI in areas such as biometric data collection or employee management, Annex III of the EU AI Act explicitly includes the use of AI in critical infrastructures as high-risk AI systems. On the one hand, this underlines the need to regulate the use of AI in critical infrastructures in order to counteract emerging risks and dangers. On the other hand, the EU Act does not classify the use of AI in critical infrastructures as a prohibited application area (e.g., such as the use of AI to predict the likelihood of a person committing a crime), but explicitly permits the use of AI in critical infrastructures subject to certain conditions and requirements (see Sect. 3 for more details on the EU AI Act).

For IS as an interdisciplinary, socio-technical field of research with profound expertise in digital technologies, the tension between the opportunities and risks of using AI in critical digital infrastructures results in a wide range of research potential. In fact, there is already a lot of research within the IS field on the use of AI in application contexts that are directly related to critical digital infrastructures (e.g. energy (Schoormann et al. 2023), healthcare (Jussupow et al. 2021), mobility (Ketter et al. 2023)). Researching high-risk AI in critical digital infrastructures can be particularly fruitful for our discipline considering the debate about the identity of the research field, which has been ongoing for more than 20 years (Benbasat and Zmud 2003). One possible answer to the identity crisis of the IS field has always been seen in focusing our research efforts on high-impact areas (Agarwal and Lucas Jr 2005). Also today, with an increasing amount of research in our field revolving around AI and related phenomena, our discipline continues to face questions about its core and scope, as well as its distinction from related disciplines

such as (applied) computer science or human-computer interaction. Against this backdrop, research on the use of high-risk AI in critical digital infrastructures can be considered inherently high-impact research from an IS perspective.

In contrast to primarily technical disciplines such as (applied) computer science, we can, for example, leverage our strengths in the development of socio-technical solutions. The dilemmas arising from the opportunities and risks of using high-risk AI in critical digital infrastructures often cannot be resolved through purely technical solutions. Instead, they require socio-technical approaches that reconcile technical solutions with the sometimes conflicting ethical and regulatory requirements. In particular, as a discipline with a strong design-oriented background, we should utilize our strengths in the development of theoretical design knowledge to generate design knowledge (e.g., in the form of design principles or full-fleged design theories) for trustworthy high-risk AI systems in critical digital infrastructures. This knowledge should thereby contribute to resolving potentially diverging requirements among different stakeholder groups in society and aligning technical solutions with these requirements. In contrast to disciplines such as human-computer interaction, we can leverage our strengths in researching the organizational perspective and organizational issues related to the use of high-risk AI in critical digital infrastructures. Organizations involved in the operation, provision, and use of critical infrastructures are often subject to specific legal requirements and are frequently not solely driven by profit motives (e.g., as is the case for healthcare organizations). These circumstances open up promising research opportunities regarding the organizational perspective on the use of high-risk AI in critical digital infrastructures. For example, exploring the impact of explainable AI on organizational decision-making in the context of critical digital infrastructures.

To summarize, the use of high-risk AI in critical digital infrastructures offers enormous potential, but requires careful consideration of the associated risks (e.g., cyberattacks, discrimination), because their disruption can lead to serious negative consequences for all of society. I firmly believe that, thanks to its interdisciplinary, socio-technical approach, IS research can make significant contributions to striking this balance between the opportunities and risks of high-risk AI in critical digital infrastructures. Key tasks for IS scholarship include the generation of design knowledge for trustworthy high-risk AI systems, as well as investigating organizational perspectives on high-risk AI in critical digital infrastructures. Only through such an integrative approach can high-risk AI be used responsibly in critical digital infrastructures to maximize the benefits of the technology while minimizing the risks.

6 Emotion Recognition Systems in Companies and Educational Institutions – High-Risk or High-Potential or Both?

Alexander Maedche

Emotions are a fundamental part of human life. They help us to orientate ourselves in everyday life, they influence our motivation, our behavior and our ability to make decisions and perform. Emotions therefore play a central role not only in our private lives, but also during education and throughout our professional lives. For example, students experience a variety of emotions during their studies, such as the fear of failing a difficult exam. Numerous studies have shown that the ability to regulate emotions is crucial for academic success (García-Ros et al. 2023). At the same time, the ability of employees to regulate their emotions plays an increasingly important role in companies operating in a world full of uncertainty and constant change. For example, it is crucial to remain professional even in difficult situations and to keep calm when interacting with customers, superiors and colleagues. Emotion regulation, the ability to exert control over one's own emotional state, is a crucial skill for maintaining mental health and promoting positive interpersonal relationships (Gross 1998). The ability to regulate emotions is considered both a hallmark of emotional intelligence and an important resilience factor for all people. However, many people lack emotion regulation skills. Existing studies have shown that between 10 and 13 percent of the population in Germany are alexithymic (Franz et al. 2008). Those affected have problems perceiving or expressing their emotions in a differentiated manner, which leads to a lower level of productivity and well-being.

Based on their ability to automatically recognize emotions and provide adaptive interventions, biosignal-adaptive systems offer great potential to help people develop the ability to regulate their emotions (Schultz and Maedche 2023). For example, contemporary wearables such as smartwatches can measure heart rate variability (HRV) in real time, based on optical and electrical biosignals using photoplethysmography (PPG) or electrocardiography (ECG). By leveraging machine learning methods, emotions can be recognized on the basis of this data and adaptive interventions can then be derived (Slovak et al. 2023). Such applications reach their full potential when they are integrated as seamlessly as possible into daily life, e.g. in learning platforms of educational institutions or in video meeting systems of companies (Benke et al. 2022). However, emotion recognition systems are classified as high-risk systems by the EU AI Act and are subject to strict regulatory oversight. Furthermore, Article 5(1)(f)[1]

[1] https://artificialintelligenceact.eu/article/5/..

prohibits the use of emotion recognition systems in the workplace and educational institutions, except for applications serving medical or security-related purposes. The good news is that according to Article 2(6),[2] research on emotion recognition systems is still permitted but will certainly be more critically scrutinized due to the ban.

Article 3(39) of the EU AI Act defines an emotion recognition system as an AI system capable of identifying or inferring human emotions or intentions through the analysis of biometric data.[3] According to Article 3(34), biometric data means personal data resulting from specific technical processing relating to the physical, physiological or behavioral characteristics of a natural person. Therefore, this definition covers all forms of biosignal-adaptive systems targeting the recognition and processing of emotions. Recital 18[4] of the EU AI Act attempts to define human emotions more precisely and names specific emotional states such as joy, sadness, anger, surprise, disgust, embarrassment, excitement, shame, contempt, satisfaction and amusement. At the same time, cognitive states such as cognitive load, distraction, or fatigue are excluded. At first glance, this list or demarcation seems somewhat arbitrary. Furthermore, the relationship between cognition and emotion, including how they influence each other in the context of IS, is a topic of ongoing research (Seitz et al. 2024). In addition to the term emotion, other concepts used in the EU AI Act also have ambiguous definitions. For example, the boundaries between private and professional life, as well as between education and work, have become increasingly blurred. Thus, a central shortcoming of the AI Act in its current form is that many central concepts such as emotion, workplace, or education institution are not well defined. These definitional uncertainties make it almost impossible to clearly define the boundaries between "prohibited" and "high risk". This leads to a lot of uncertainty and, in the worst case, to stagnation in the important research field of biosignal-adaptive systems in general and emotion recognition systems in particular. This is problematic because, despite the EU AI Act, people continue to have emotions in the workplace and in educational institutions and are faced with challenges in their ability to regulate emotions.

I am not claiming that emotion recognition systems are risk-free. Recognizing and processing emotions comes with many challenges, e.g., proper definition and subjective nature of emotions, lack of representative and generalizable data, as well as quality of recognition methods (Hernandez et al. 2021). However, at the same time, most information technologies come with risks. Through systematic investigation and responsible design of information technologies, risks can be better understood and ideally eliminated or at least reduced. I believe that regulating information technology in general is an important step. However, by just banning emotion recognition systems, we are missing the opportunity to use the potential of biosignal-adaptive systems supporting emotion regulation for employees in companies and students in educational institutions in Germany and Europe. At the same time, research in this area is being driven forward in China and the United States.

I believe that we should take a human-centered approach to AI in Europe. As emotions are an inherent part of human nature, I firmly believe that we should not simply prohibit the use of AI methods for recognizing and processing emotions. Instead, we should understand how such systems affect people in their private and professional lives and how we need to design them in a targeted manner. For this purpose, corresponding systems must be deployed and researched in the real world, especially in companies and educational institutions. The EU AI Act proposes the concept of AI regulatory sandboxes (Article 57[5]) as well as testing of high-risk AI systems in real-world conditions outside AI regulatory sandboxes (Article 60[6]). My wish is that companies and educational institutions actively use these opportunities to design, test and successfully apply innovative, AI-based emotion recognition systems to support people.

7 High-Risk Artificial Intelligence – A Startup and Small to Medium Business' Perspective

Joshua Gawlitza

From a business perspective, building a company around a high-risk AI product poses different challenges: from the implementation of such a system over regulatory hurdles to different go-to-market approaches. With a growing number of startups and usage of such systems, I aim to provide a brief overview for younger and agile companies with smaller teams and budgets to understand the main challenges from a small to medium business perspective.

Looking at the current number of startups founded, AI-focused startups are on the rise. In Germany alone, the number of AI startups increased by 67% from 2022 to 2023 (appliedAI Institute for Europe gGmbH 2023). Examining the distribution of these startups, most of them are active in the category "Human Health". Especially, when considering the techniques used or being developed by the startups in this area, besides Large Language Models,

[2] https://artificialintelligenceact.eu/article/2/..

[3] https://artificialintelligenceact.eu/article/3/..

[4] https://artificialintelligenceact.eu/recital/18/..

[5] https://artificialintelligenceact.eu/article/57/..

[6] https://artificialintelligenceact.eu/article/60/..

predictive or computer vision models dominate. When looking at the potential of both techniques, it makes sense that most startups develop in this direction as prediction or computer aided-diagnosis systems can be beneficial in almost every medical discipline – from general practitioners to radiologists and from prevention or screening to post-mortal diagnosis (Kumar et al. 2023; Piraianu et al. 2023; Sangers et al. 2023). But bringing a high-risk AI product to market poses considerable challenges for startups and small businesses. In general, there are three main challenges: 1. access to reliable training data, 2. regulatory challenges, 3. go-to-market time.

7.1 Access to Reliable Training Data

One of the biggest challenges for a startup is access to reliable training data for either computer vision or prediction models. When viewing data from recent years, we see a constant decline in startups founded by academic researchers or as academic spin-offs (Kulicke 2023). However, access to reliable medical training data from, e.g., research projects from university hospitals is difficult to obtain for most startups founded outside the academic world. When looking at recent reviews of high-risk medical AI software and device safety, this problem appears to be eminent, as most articles related to high-risk AI safety in medicine see the "data acquisition process" as one of the major critical steps to define proper guidelines for AI safety evaluation (Fraser et al. 2023). Some startups mitigate the issue by partnering with private healthcare providers to access large datasets – one example here is the company Deepeye, which has partnered with a private ophthalmologist center in Muenster to train its models on optical coherence tomography data (Münsterland e.V. 2024). If and how data quality differs between academic and private institutions has not been evaluated yet, but especially for high-risk AI, guidelines for proper screening of the profound dataset are in desperate need, however, the importance of training data reaches beyond the quality of the final product. As shown by Bessen et al. (2022), there is a significant correlation between access to proprietary training data and future funding for AI startups.

7.2 Regulatory Challenges

As pointed out above, regulatory guidelines are essential, when it comes to high-risk AI. Besides the data acquisition process, most experts agree on further crucial criteria when it comes to define safety evaluation of high-risk AI: data preprocessing, model description, characteristics of study population, performance and benchmarking, and code/data availability (Fraser et al. 2023). Official certifications such as CE, ISO/IEC challenge certain aspects of these critical points in their audits. But despite the benefit of an officially certified system, these standards come with a major downside for young companies. When looking at the 2023 venture capital (VC) study from Pricewaterhouse Coopers, the number of VC deals decreased by 40% compared to 2022 (Honold et al. 2023). The most drastic decline affected pre-seed investments, i.e. very early investments for freshly founded startups and companies which typically do not yet generate any revenue. According to the VC study, investors are looking for more stable investments with a faster return on interest instead of high-risk early investments such as pre-seed. This immediately affects the capabilities of young companies to develop high-risk AI tools, as the certification processes, which are necessary to operate high-risk AI in most fields, are costly. For example, CE certification is typically quoted at around €50,000, and ISO27001 certification at between €6,000 and €40,000 (Alura Group UK Ltd. 2024; Vanta 2024). Further certifications and audits, such as ISO/IEC 42001, might be necessary for particular markets, raising the total initial investment to six figures, only for regulatory certifications. Thus, many high-risk AI startups release their products as a "research only" tools, until they can generate further revenue for the certification process. This is reflected in the low number of FDA-approved algorithms compared to the dramatic increase in high-risk AI startups (Benjamens et al. 2020).

7.3 Go-to-Market Time

Facing a more competitive investment market and strong expectations from VCs for a quick return on investment, young companies are increasingly forced to generate revenue faster. This is especially challenging for high-risk AI algorithms, which need time to be developed, evaluated and potentially certified. In comparison to traditional software development, this leads to a delay in revenue generation. The CE certification process alone can take up to a year or more, challenging business plans and go-to-market times (MedDev Compliance Ltd. 2024). As shown in a recent meta-analysis, at least the time-consuming process of AI development is mitigated by startups having relatively more employees in relation to revenue compared to traditional software service or platform technology companies (Schulte-Althoff et al. 2021). Given this, the more competitive investment environment and the importance of access to proprietary data into account, high-risk AI startups might face even more challenges in the very early design and funding phase.

In conclusion, despite the high potential for customers, consumers and investors, high-risk AI startups face major challenges when compared to traditional software businesses. Regulatory advances like the EU AI Act may seem

to complicate this process at first glance, but they provide certain guidelines that can be used by founders to set the right strategy for their business at an early stage. Academics should be encouraged to bring their research to the market, with access to valuable data and pre-development during their research time. Although the regulatory aspects of high-risk AI are currently under active research, further data is needed to evaluate the different aspects of high-risk AI startups from a business and business development perspective.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Abdel-Karim BM, Pfeuffer N, Carl KV, Hinz O (2023) How AI-based systems can induce reflections: the case of AI-augmented diagnostic work. MIS Q 47(4):1395–1424

Adam M, Kosin D, Benlian A (2024) From detractors to enhancers: harnessing the power of ad customization for user engagement on media websites. J Assoc Inf Syst 26(1):241–265

Agarwal R, Lucas HC Jr (2005) The information systems identity crisis: focusing on high-visibility and high-impact research. MIS Q 29(3):381–398

Albahri A, Khaleel YL, Habeeb MA, Ismael RD, Hameed QA, Deveci M, Homod RZ, Albahri O, Alamoodi A, Alzubaidi L (2024) A systematic review of trustworthy artificial intelligence applications in natural disasters. Comput Electr Eng 118:109409

Alura Group UK Ltd (2024) How much does CE certification cost? https://cemarking.net/what-are-the-costs-of-ce-certification/. Accessed 16 Nov 2024

appliedAI Institute for Europe gGmbH (2023) German AI startup landscape 2023. https://www.appliedai-institute.de/hub/2023-ai-german-startup-landscape. Accessed 9 Nov 2024

Baird A, Maruping LM (2021) The next generation of research on IS use: a theoretical framework of delegation to and from agentic IS artifacts. MIS Q 45(1):315–341

Bauer K, Gill A (2024) Mirror, mirror on the wall: Algorithmic assessments, transparency, and self-fulfilling prophecies. Inf Syst Res 35(1):226–248

Benbasat I, Zmud RW (2003) The identity crisis within the IS discipline: defining and communicating the discipline's core properties. MIS Q 27(2):183–194

Benjamens S, Dhunnoo P, Meskó B (2020) The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Dig Med 3(1):118–125

Benke I, Schneider M, Liu X, Maedche A (2022) TeamSpiritous – a retrospective emotional competence development system for video-meetings. In: Proceedings of the ACM on Human-Computer Interaction 6 (CSCW2)

Berente N, Gu B, Recker J, Santhanam R (2021) Special issue editor's comments: Managing artificial intelligence. MIS Q 45(3):1433–1450

Bessen J, Impink SM, Reichensperger L, Seamans R (2022) The role of data for AI startup growth. Res Policy 51(5):104513

Brasse J, Broder HR, Förster M, Klier M, Sigler I (2023) Explainable artificial intelligence in information systems: a review of the status quo and future research directions. Electron Mark 33(1):1–30

Butler T, Gozman D, Lyytinen K (2023) The regulation of and through information technology: towards a conceptual ontology for IS research. J Inf Technol 38(2):86–107

Caton S, Haas C (2024) Fairness in machine learning: a survey. ACM Comput Surv 56(7):1–38

Corbett-Davies S, Gaebler JD, Nilforoshan H, Shroff R, Goel S (2023) The measure and mismeasure of fairness. J Mach Learn Res 24(1):14730–14846

Cram W, D'Arcy J, Benlian A (2024) Time will tell: a case for an idiographic approach for behavioral cybersecurity research. MIS Q 48(1):95–136

de Nobrega KM, Rutkowski A-F, Saunders C (2024) The whole of cyber defense: syncing practice and theory. J Strateg Inf Syst 33(4):101861

Dehling T, Lins S, Sunyaev A (2019) Security of critical information infrastructures. In: Reuter C (ed) Information technology for peace and security: It applications and infrastructures in conflicts, crises, war, and peace. Springer Vieweg, Wiesbaden, pp 319–339

Dolata M, Feuerriegel S, Schwabe G (2022) A sociotechnical view of algorithmic fairness. Inf Syst J 32(4):754–818

Edwards L (2022) Expert explainer: The EU AI Act proposal. Ada Lovelace Institute. https://www.adalovelaceinstitute.org/resource/eu-ai-act-explainer/. Accessed 9 Nov 2024

European Commission (2021) Proposal for a regulation laying down harmonised rules on artificial intelligence. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence. Accessed 9 Nov 2024

European Union (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (2024). https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Fekete A (2011) Common criteria for the assessment of critical infrastructures. Int J Disaster Risk Sci 2:15–24

Franz M, Popp K, Schaefer R, Sitte W, Schneider C, Hardt J, Decker O, Braehler E (2008) Alexithymia in the German general population. Soc Psychiatry Psychiatr Epidemiol 43:54–62

Fraser AG, Biasin E, Bijnens B, Bruining N, Caiani EG, Cobbaert K, Davies RH, Gilbert SH, Hovestadt L, Kamenjasevic E (2023) Artificial intelligence in medical device software and high-risk medical devices—a review of definitions, expert recommendations and regulatory initiatives. Exp Rev Med Devices 20(6):467–491

Frazer HML, Peña-Solorzano CA, Kwok CF, Elliott MS, Chen Y, Wang C, Team TB, Lippey JF, Hopper JL, Brotchie P, Carneiro G, McCarthy DJ (2024) Comparison of AI-integrated pathways

with human-AI interaction in population mammographic screening for breast cancer. Nat Commun 15(1):7525
Fügener A, Grahl J, Gupta A, Ketter W (2021) Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. MIS Q 45(3):1527–1556
Fügener A, Grahl J, Gupta A, Ketter W (2022) Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. Inf Syst Res 33(2):678–696
Future of Life Institute (2024) Annex III: High-risk AI systems referred to in article 6(2). https://artificialintelligenceact.eu/annex/3/. Accessed 9 Nov 2024
García-Ros R, Pérez-González F, Tomás JM, Sancho P (2023) Effects of self-regulated learning and procrastination on academic stress, subjective well-being, and academic achievement in secondary education. Curr Psychol 42(30):26602–26616
Gierlich-Joas M, Baiyere A, Hess T (2024) Inverse transparency and the quest for empowerment through the design of digital workplace technologies. J Assoc Inf Syst 25(5):1212–1239
Goddard K, Roudsari A, Wyatt JC (2012) Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 19(1):121–127
Gross JJ (1998) The emerging field of emotion regulation: An integrative review. Rev Gen Psychol 2(3):271–299
Harford T (2012) High-frequency trading and the $440m mistake. BBC. https://www.bbc.com/news/magazine-19214294. Accessed 9 Nov 2024
Hernandez J, Lovejoy J, McDuff D, Suh J, O'Brien T, Sethumadhavan A, Greene G, Picard R, Czerwinski M (2021) Guidelines for assessing and minimizing risks of emotion recognition applications. In: 9th International conference on affective computing and intelligent interaction, Virtual Conference
Hollick M, Katzenbeisser S (2019) Resilient critical infrastructures. In: Reuter C (ed) Information technology for peace and security: It applications and infrastructures in conflicts, crises, war, and peace. Springer Vieweg, Wiesbaden, pp 305–318
Honold D, Reiche E, Wacker G (2023) Venture capital market study 2023. https://www.bvkap.de/files/content/Studien/Venture%20capital%20market%20study_23_Honold%20et.%20al.pdf. Accessed 9 Nov 2024
Hupont I, Micheli M, Delipetrev B, Gómez E, Garrido JS (2023) Documenting high-risk AI: a European regulatory perspective. Comput 56(5):18–27. https://doi.org/10.1109/MC.2023.3235712
Jussupow E, Spohrer K, Heinzl A (2022) Radiologists' usage of diagnostic AI systems: The role of diagnostic self-efficacy for sensemaking from confirmation and disconfirmation. Bus Inf Syst Eng 64(3):293–309
Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. Inf Syst Res 32(3):713–735. https://doi.org/10.1287/isre.2020.0980
Kaissis GA, Makowski MR, Rückert D, Braren RF (2020) Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell 2(6):305–311
Karim MM, Li Y, Qin R (2022) Toward explainable artificial intelligence for early anticipation of traffic accidents. Transp Res Rec 2676(6):743–755
Kemper J, Kolkman D (2019) Transparent to whom? No algorithmic accountability without a critical audience. Inf Commun Soc 22(14):2081–2096
Ketter W, Schroer K, Valogianni K (2023) Information systems research for smart sustainable mobility: a framework and call for action. Inf Syst Res 34(3):1045–1065
Khan MA, Saleh AM, Waseem M, Sajjad IA (2022) Artificial intelligence enabled demand response: prospects and challenges in smart grid environment. IEEE Access 11:1477–1505
Kordzadeh N, Ghasemaghaei M (2022) Algorithmic bias: review, synthesis, and future research directions. Eur J Inf Syst 31(3):388–409
Kulicke M (2023) Spin-offs aus Hochschulen und Forschungseinrichtungen in Deutschland und weiteren Ländern. https://www.stifterverband.org/sites/default/files/2023-11/spin-offs_aus_hochschulen_und_forschungseinrichtungen_in_deutschland_und_weiteren_laendern.pdf. Accessed 9 Nov 2024
Kumar P, Chauhan S, Awasthi LK (2023) Artificial intelligence in healthcare: review, ethics, trust challenges and future research directions. Eng Appl Artif Intell 120:105894
Lebovitz S, Lifshitz-Assaf H, Levina N (2022) To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. Organ Sci 33(1):126–148
Long D, Magerko B (2020) What is AI literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI conference on human factors in computing systems, Yokohama
Maedche A, Legner C, Benlian A, Berger B, Gimpel H, Hess T, Hinz O, Morana S, Söllner M (2019) AI-based digital assistants. Bus Inf Syst Eng 61(4):535–544
MedDev Compliance Ltd (2024) How long does it take to CE-mark a medical device? https://mdrc-services.com/ce-marking-timelines/. Accessed 17 Nov 2024
Münsterland e.V. (2024) Project DeepEye: Protecting against age blindness with AI. https://www.muensterland.com/en/economy/business-location/innovations/innovation-stories/deep-eye/. Accessed 13 Nov 2024
Nagbøl PR, Müller O, Krancher O (2021) Designing a risk assessment tool for artificial intelligence systems. In: 16th International conference on design science research in information systems and technology, Kristiansand
National Institute of Standards and Technologies (NIST) (2023) AI risk management framework. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf. Accessed 12 Nov 2024
Nelson TO, Narens L (1994) Why investigate metacognition? In: Metcalfe J, Shimamura AP (eds) Metacognition: knowing about knowing. MIT Press, Cambridge, pp 1–25
Pfeiffer J, Gutschow J, Haas C, Möslein F, Maspfuhl O, Borgers F, Alpsancar S (2023) Algorithmic fairness in AI. Bus Inf Syst Eng 65(2):209–222. https://doi.org/10.1007/s12599-023-00787-x
Pfeiffer J, Lachenmaier JF, Hinz O, van der Aalst W (2024) New laws and regulation. Bus Inf Syst Eng 66:653–666
Pinski M, Benlian A (2024) AI literacy for users – a comprehensive review and future research directions of learning methods, components, and effects. Comput Hum Behav Artif Hum 2(1):100062
Piraianu A-I, Fulga A, Musat CL, Ciobotaru O-R, Poalelungi DG, Stamate E, Ciobotaru O, Fulga I (2023) Enhancing the evidence with algorithms: how artificial intelligence is transforming forensic medicine. Diagn 13(18):2992
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215
Sánchez PMS, Celdrán AH, Xie N, Bovet G, Pérez GM, Stiller B (2024) FederatedTrust: a solution for trustworthy federated learning. Futur Gen Comput Syst 152:83–98
Sangers TE, Wakkee M, Moolenburgh FJ, Nijsten T, Lugtenberg M (2023) Towards successful implementation of artificial intelligence in skin cancer care: a qualitative study exploring the views of dermatologists and general practitioners. Arch Dermatol Res 315(5):1187–1195
Schoormann T, Strobel G, Möller F, Petrik D, Zschech P (2023) Artificial intelligence for sustainability – a systematic review of information systems literature. Commun Assoc Inf Syst 52(1):556–592

Schuetz S, Venkatesh V (2020) The rise of human machines: How cognitive computing systems challenge assumptions of user-system interaction. J Assoc Inf Syst 21(2):460–482

Schulte-Althoff M, Fürstenau D, Lee GM (2021) A scaling perspective on AI startups. In: 54th Annual Hawaii International Conference on System Sciences, Hawaii

Schultz T, Maedche A (2023) Biosignals meet adaptive systems. SN Appl Sci 5(9):234

Seitz J, Benke I, Heinzl A, Maedche A (2024) The impact of video meeting systems on psychological user states: a state-of-the-art review. Int J Hum-Comput Stud 182:103178

Shirouzu N, Kirkham C (2024) Tesla's robotaxi push hinges on 'black box' AI gamble. https://www.reuters.com/technology/tesla-gambles-black-box-ai-tech-robotaxis-2024-10-10/. Accessed 12 Jan 2025

Simbeck K (2024) They shall be fair, transparent, and robust: auditing learning analytics systems. AI Ethics 4(2):555–571

Slovak P, Antle A, Theofanopoulou N, Daudén Roquet C, Gross J, Isbister K (2023) Designing for emotion regulation interventions: an agenda for HCI theory and research. ACM Trans Comput-Hum Interact 30(1):1–51

Spiekermann S, Krasnova H, Hinz O, Baumann A, Benlian A, Gimpel H, Heimbach I, Köster A, Maedche A, Niehaves B (2022) Values and ethics in information systems: a state-of-the-art analysis and avenues for future research. Bus Inf Syst Eng 64(2):247–264

Stacey K (2023) UK risks scandal over 'bias' in AI tools in use across public sector. BBC. https://www.theguardian.com/technology/2023/oct/23/uk-risks-scandal-over-bias-in-ai-tools-in-use-across-public-sector. Accessed 9 Nov 2024

Taudien A, Walzner DD, Fuegener A, Gupta A, Ketter W (2024) Know thyself: the relationship between metacognition and human-AI collaboration. In: International conference on information systems, Bangkok

Teodorescu MH, Morse L, Awwad Y, Kane GC (2021) Failures of fairness in automation require a deeper understanding of human-ML augmentation. MIS Q 45(3):1483–1500

Thiebes S, Lins S, Sunyaev A (2021) Trustworthy artificial intelligence. Electron Mark 31(2):447–464. https://doi.org/10.1007/s12525-020-00441-4

Tjoa S, Temper PKM, Temper M, Zanol J, Wagner M, Holzinger A (2022) AIRMan: An artificial intelligence (AI) risk management system. In: 2022 International Conference on Advanced Enterprise Information System, London

Tranberg P (2023) Corporate Europe: US big tech lobbied against generative AI as high risk in the AI act. DataEthics. https://dataethics.eu/corporate-europe-us-big-tech-lobbied-against-generative-ai-as-high-risk-in-the-ai-act/. Accessed 9 Nov 2024

Trier M, Kundisch D, Beverungen D, Müller O, Schryen G, Mirbabaie M, Trang S (2023) Digital responsibility: a multilevel framework for responsible digitalization. Bus Inf Syst Eng 65(4):463–474

Ungoed-Thomas J, Abdulahi Y (2024) Warnings AI tools used by government on UK public are 'racist and biased'. The Guardian. https://www.theguardian.com/technology/article/2024/aug/25/register-aims-to-quash-fears-over-racist-and-biased-ai-tools-used-on-uk-public. Accessed 10 Jan 2025

Vaccaro M, Almaatouq A, Malone T (2024) When combinations of humans and AI are useful: a systematic review and meta-analysis. Nat Hum Behav 8:1–11

Vanta (2024) How much does ISO 27001 certification cost? https://www.vanta.com/collection/iso-27001/iso-27001-certification-cost. Accessed 16 Nov 2024

Wang Y, Chung SH (2022) Artificial intelligence in safety-critical systems: a systematic review. Ind Manag Data Syst 122(2):442–470

Wiener M, Cram WA, Benlian A (2023) Algorithmic control and gig workers: a legitimacy perspective of Uber drivers. Eur J Inf Syst 32(3):485–507

Wilson C, Ghosh A, Jiang S, Mislove A, Baker L, Szary J, Trindel K, Polli F (2021) Building and auditing fair algorithms: a case study in candidate screening. In: 2021 ACM conference on fairness, accountability, and transparency, virtual conference

Yang J, Amrollahi A, Marrone M (2024) Harnessing the potential of artificial intelligence: affordances, constraints, and strategic implications for professional services. J Strateg Inf Syst 33(4):101864