AMCIS 2022 Proceedings

SIG ODIS - Artificial Intelligence and Semantic Technologies for Intelligent Systems

Aug 10th, 12:00 AM

Let´s Do Our Bit: How Information Systems Research Can Contribute to Ethical Artificial Intelligence

Aycan Aslan
Chair of Information Management, aycan.aslan@uni-goettingen.de

Maike Greve
University of Goettingen, maike.greve@uni-goettingen.de

Tim-Benjamin Lembcke
University of Goettingen, tim-benjamin.lembcke@uni-goettingen.de

Recommended Citation

Aslan, Aycan; Greve, Maike; and Lembcke, Tim-Benjamin, "Let´s Do Our Bit: How Information Systems Research Can Contribute to Ethical Artificial Intelligence" (2022). *AMCIS 2022 Proceedings*. 10.
https://aisel.aisnet.org/amcis2022/sig_odis/sig_odis/10

Let´s Do Our Bit: How Information Systems Research Can Contribute to Ethical Artificial Intelligence

Completed Research

Aycan Aslan
University of Goettingen
aycan.aslan@uni-goettingen.de

Maike Greve
University of Goettingen
maike.greve@uni-goettingen.de

Tim-Benjamin Lembcke
University of Goettingen
tim-benjamin.lembcke@uni-goettingen.de

Abstract

With the growing diffusion of Artificial Intelligence (AI), a variety of novel ethical challenges become apparent. As a possible solution, the field of Ethical AI (EAI), the conviction, that to build trust in AI, ethical guidelines must be enforced, is gaining popularity. The importance of such guidelines has been recognized by regulatory institutions like the European Commission and led to the 'Artificial Intelligence Act', a legal framework to achieve EAI. In view of the growing importance of EAI and its implications for how societies interact with AI, it remains unclear to which extent information systems (IS) literature, socio-technical by its nature, can contribute to EAI. This article's contribution is threefold: It provides an explanation of the AI Act, introduces a systematic analysis of current IS literature on EAI, and derives future research directions for aspects of EAI which are currently insufficiently covered by IS scholars.

Keywords

Ethical Artificial Intelligence, Artificial Intelligence Act, Future Research Directions.

Introduction

Artificial Intelligence (AI) is among the most influential technology trends of our time. As frequently discussed, AI offers a range of new promising opportunities like improving the safety of medical care, predictive maintenance in manufacturing, or sustainable and precision agriculture (Bates et al. 2021; Mayr et al. 2019; P. Zhang et al. 2021). While the potentials are promising, recently researchers raised a variety of ethical and social challenges related to implementing and using AI-systems (Floridi et al. 2018; Thiebes et al. 2021). Examples of unethical AI are omnipresent: In 2018, researchers found that Amazons recruiting tool discriminated female coders (Dastin 2018). Further, Facebook´s ad targeting algorithms have been found to perpetuate gender disparities, such that female users were less likely to see ads from companies that predominantly hire male employees (Imana et al. 2021).

These examples show that ethical guidelines are needed to enforce trust in using AI, to maximize the potentials of AI while mitigating potential risks. These principles are reflected in the field of Ethical AI (EAI). In essence, EAI describes enforcing common ethical and moral principles when designing and developing AI, such as respecting fairness measures or human control over AI. The idea here being, that ethical considerations, such as algorithmic biases based on gender or race, are so essential for how we as individuals and societies are interacting with AI, that the enforcement of sufficient ethical guidelines are one of the most important prerequisites to realize the full potential of AI (Thiebes et al. 2021). This need for enforcing ethical guidelines have been recognized by regulatory bodies as well. Here, the effort of the European Union (EU) was noticed particularly. Working on ethical guidelines for AI for many years now, in 2021, the European Commission (EC) presented the ‘Artificial Intelligence Act’ (AI Act), a draft for the first holistic legal framework on AI. In the AI Act, the EC proclaims that the regulation is supposed to support the EU as the global leader in the development of EAI (AI Act 2021). However, the impact of such regulation will likely extend the scope of the EU. With the relevance of the European market in mind, one could expect that the enactment of the AI Act could further strengthen the so-called ‘Brussels effect’, which describes the unilateral impact of EU regulation (Bradford 2020). One example in this context is the General Data Protection Regulation (GDPR), which was enacted by the EU in 2016 and quickly become a model for many privacy laws, e.g., the state of California and the respective California Consumer Privacy Act (Lovejoy 2019).

To date, achieving EAI has mainly been discussed from a technical perspective by the Computer Science community, for example, by implementing mathematical representations of privacy in AI-systems (Papernot et al. 2018) or fairness metrics for AI-systems (Hardt et al. 2016). However, enforcing ethical guidelines to strengthen and achieve true trustworthiness of AI will affect how everyone one of us develop, manage, or use AI-systems. For example, the absence of appropriate transparency and interpretability measures for AI could hinder the adaption of it in domains such as healthcare (Park et al. 2019). Hence, we argue that, while a technical understanding is important, the purely technical perspective on EAI does not do justice to such a societally crucial topic. In this regard, the IS community – probably better than any other field – is predestined to address the complex challenges at the intersection of technology, organizational research, and behavioral research, by providing a technical understanding while also being able to understand complex behavioral and socio-technical problems. Despite this fit and potential for IS research, we note that to date EAI research in IS outlets is still limited (Thiebes et al. 2021).

Therefore, the goal of this paper is twofold: First, to assess the current status quo of EAI in IS research. Second, to derive future directions for IS research and to provide indications how IS research could cater to the emerging field of EAI. To address these objectives, this paper aims to answer the following guiding research question:

RQ: *Whether and to what extent are the facets of EAI defined by the AI Act addressed in IS research?*

Our work follows a threefold procedural approach to answer this question: First, we introduce the AI Act, common frameworks of EAI, and point out the differences among them. Second, we review and present a comprehensive overview of the current status quo of EAI research in IS related outlets, based on the EU’s conceptualization of EAI. Finally, we discuss future research directions how scholars can further enrich the EAI literature, particularly from an IS perspective.

Theoretical and Conceptual Background

In this section, we will explain and analyze the field of EAI. The term EAI can be understood as an umbrella term and is used for different concepts (Thiebes et al. 2021). In this chapter, we review different frameworks for EAI to provide an understanding how these definitions and frameworks differ in terminology and content. Based on these explanations, we introduce the AI Act and analyze the facets of EAI named in the Act.

What is Ethical Artificial Intelligence?

In terms of both wording and content, EAI is yet to become a clearly defined field. In this context, the most common definitions and frameworks are AI4People (Floridi et al. 2018), the OECD Principles of AI (OECD 2019), and the Ethics Guidelines for Trustworthy AI (‘High-level expert group on Artificial Intelligence’ (HLEG 2019)). In terms of wording, Ethical AI has been used interchangeably with ‘trustworthy AI’, ‘beneficial AI’, and ‘responsible AI’ (Thiebes et al. 2021).

However, despite these differences in terminology, we note a high degree of overlap in terms of content which we summarize in four major principles for EAI: First, the *beneficence of AI to humans and the environment*. All frameworks point out that AI should promote the well-beings of humans and the environment. Therefore, positive, and negative impacts of AI should be evaluated to permit well-being of all sentient beings and to ensure maintaining solidarity among people. Second, that *AI should not harm humans*. All frameworks agree that AI should not bring harm to people and that possible adverse consequences should be considered in the development of AI. A special focus here is the protection of privacy, where all frameworks state the importance of privacy and intimacy. Third, that *human autonomy, human agency, and oversight* should be enforced. Here, the frameworks state that AI should respect human autonomy and not lessen human responsibility but complement it. Further, humans should stay in control over the AI and retain the right to decide at any given time in the development and use of AI-systems. Fourth, that *AI should be fair and just*. The frameworks agree that AI should be fair and just in the sense that it respects democratic rights. AI should abandon discrimination, prevent the emergence of new inequities, and promote diversity inclusion.

Nevertheless, there are differences among the frameworks: First, we identify different extents to which the principle of technical robustness and resilience is mentioned. Here, the Ethics Guidelines for Trustworthy AI state that AI quality assurance should be embedded, and the technical robustness and resilience should be exhibited frequently (HLEG 2019). Similarly, the OECD Principles of AI highlight that AI must function in a robust and secure way (OECD 2019). However, such technical quality requirements are not referred to in the framework AI4People (Floridi et al. 2018).

The EU Artificial Intelligence Act

In April 2021, the EU Commission presented the proposal for the AI Act. The proposal's objective is to lay down harmonized rules on AI, by following a risk-based approach. It aims at providing clear requirements for the use of AI and categorizes AI's usage into four different classes, associated with different level of risks (AI Act 2021). The first two classes are 'minimal risk' and 'limited risk'. While AI-systems classified as bearing 'minimal risk' are free to use, AI-systems that show 'limited risk' are obligated to light transparency rules. However, the main goal of categorizing AI use-cases into risk classes is to identify AI-systems with high and unacceptable risk, for which defined guidelines will apply (AI Act 2021). According to the EU, AI-systems are categorized as bearing 'unacceptable risk' if they pose a threat to safety, livelihoods, and human rights. According to the AI Act, in the future such systems will be banned in the EU. Examples for such systems are voice assistance that encourages dangerous behavior or social scoring initiatives by governments. High-risk AI-systems are systems that are associated with critical areas of life. These areas include critical infrastructure that could put health of people at risk, educational training that may determine access to education, or safety components of products. Such systems classified as high-risk are subject to guidelines that shall ensure trustworthiness in AI-systems. These guidelines are based on the 'Ethics Guidelines for Trustworthy AI' by the HLEG stated in the previous sub-section, whereby the principle of 'Societal and environmental well-being' was removed. Hence, the respective similarities between the Ethics Guidelines for Trustworthy AI and the other frameworks apply to the AI Act as well.

According to the AI Act (AI Act 2021), high-risk AI-systems are subject to following six ethical requirements (see Table 1): The first requirement is '*Human agency and oversight*'. This requirement ensures that AI does not undermine human autonomy. On the contrary, AI-systems should empower human beings and allow them to make well-informed decisions. Additionally, proper AI oversight mechanisms must be ensured for example by human-in-the-loop (HITL) approaches that ensure human oversight over critical decision of the system. Second, '*Technical robustness and safety*', ensures that AI is secure and reliable. The AI Act requires AI-systems to be resilient and sage. This includes the existence of fallback plans in case something goes unexpected or wrong. Third, '*Privacy and data governance*', ensures that data is protected at all stages. Additional to appropriate data privacy rules, this requirement also describes adequate data governance mechanisms that consider the quality and integrity of the data. The fourth requirement is '*Transparency'*, which describes that AI-systems should be traceable and transparent in their decision-making. All components, the data, the system itself, and the AI business model should be made transparent. Traceability mechanisms can help to achieve this goal. Stakeholders, and particularly experts, should be able to understand and explain AI-systems and their decisions. Fifth, '*Diversity, non-discrimination and fairness*' describes the tackling of biases in developing AI-systems and AI-systems itself. Unfair bias must be avoided, since it could invoke negative consequences for marginalized groups to the exacerbation of prejudice and discrimination. Additionally, AI-systems should be as accessible as possible to promote more diversity in developing and using AI-systems. Sixth and last, '*Accountability'* indicates that appropriate mechanisms should be put in place to ensure responsibility and accountability for AI-systems. Consequently, it must be ensured that AI-systems can be assessed and consequently the accountability of the system is ensured.

EAI facet in the AI Act	Explanation
'Human agency and oversight'	AI-systems should incorporate human agency and oversight to guarantee that humans keep control over the systems and have oversight over AI-systems at any given time.
'Technical robustness and safety'	AI-systems should be technically robust and secure to minimize negative consequences in case of adverse events.
'Privacy and data governance'	AI-systems should respect the protection of personal data. Additionally, sufficient data governance methods to monitor the data use of AI-systems should be implemented.
'Transparency'	AI-systems should be transparent in the sense, that humans should be able to understand and challenge outcomes of the systems.
'Diversity, non-discrimination, and fairness'	AI-systems should be fair and respect human rights, such as non-discrimination and unbiased approach to the development of AI.
'Accountability'	AI-systems should be designed in ways that allows for sufficient accountability and assessment of the system and system outcomes.

Table 1. Explanations for the ethical guidelines stated in the AI Act

Methodology

To understand the status quo of EAI research in IS outlets and answer the RQ, we conducted a systematic literature review as suggested by vom Brocke et al. (2009). The procedure follows three phases: Literature search, literature evaluation and selection, and literature synthesis (vom Brocke et al., 2009).

Literature Search

The goal of our work is to understand the status quo of EAI in IS outlets, based on the definition in the AI Act (six facets for EAI). We examined the following databases: ProQuest, EBSCO Host, and Science Direct for the 'Basket of Eight' journals and the AIS eLibrary for the three leading IS conferences: International Conference on Information Systems (ICIS), the European Conference on Information Systems (ECIS), and the Pacific Asia Conference on Information Systems (PACIS). Besides journal articles, we deliberately included conference articles to recognize the topic's novelty. The language of the articles is limited to English and only peer-reviewed articles are included to validate the quality of research.

To reflect the multi-faceted character of the topic, we split our keyword search into two parts. The first represents the synonyms for 'ethical' and the principles for EAI reflected in the AI Act, the second part displays 'Artificial Intelligence' as the analyzed technology. The search query was as follow: *(ethical OR trustworthy OR responsible OR 'human agency' OR 'technical robustness' OR privacy OR transparency OR explainability OR diversity OR non-discrimination OR fairness OR accountability) AND (AI OR 'artificial intelligence' OR 'machine learning' OR 'deep learning').*

Literature Evaluation and Selection

Our search revealed an initial set of 64 studies. First, the research team analyzed the articles' titles and abstracts for their thematic fit. We excluded articles that mentioned the respective topics in the abstract, but did not discuss them as their focus in the article, leading to 39 remaining articles. With these articles, we conducted an in-depth full-text analysis, in which we checked the remaining articles based on the following inclusion criteria (all three had to be fulfilled): 1) The study should clearly state AI-systems as their technological focus, 2) The study should clearly state at least one of the EAI facets as their focus, 3) The study should make an original contribution to the existing IS literature in the respective EAI facet. The inclusion criteria limited the number of articles to 10. Eventually, a forward and backward search led to three additional articles and 13 relevant articles in total.

Results of the Analysis

This section discusses the results of the conducted literature analysis and synthesis (see Table 2), providing an overview of the 13 analyzed articles in terms of the respectively addressed EAI facets.

	Facets of EAI defined by the Artificial Intelligence Act					
Article	1) Human agency and oversight	2) Tech. robustness and safety	3) Privacy and data governance	4) Transparency	5) Diversity, non-discrimination, and fairness	6) Accountability
Asatiani et al. (2021)				X		
Van den Broek et al. (2021)	X					
Teodorescu et al. (2021)	X					
Thiebes et al. (2021)	O		O	O	O	
Meske et al. (2021)				X		
Trocin et al. (2021)	O			X	X	O
Zhang et al. (2021)			X			
Fu et al. (2021)					X	
Schneider and Handali (2019)				X		
Hemmer et al. (2021)	X					
Schemmer et al. (2021)		X				
Von Zah et al. (2021)					X	
Feuerriegel et al. (2020)					X	
Total No. of articles	5	1	2	5	5	1
***Legend**: X: Indicates that the concept is the focus of the work; O: Indicates that the concept is touched on.*						

Table 2. Overview of the results of our analysis

Human Agency and Oversight

Our literature review revealed five articles that address the facet 'Human agency and oversight' (van den Broek et al. 2021; Hemmer et al. 2021; Teodorescu et al. 2021; Thiebes et al. 2021; Trocin et al. 2021). These articles address issues regarding the design, characteristics, and success factors of Human-AI collaboration.

Hemmer et al. (2021) identify relevant success factors and group them into 'Collaboration characteristics', 'Task characteristics', 'AI characteristics', and 'Human characteristics'. Through this grouping, a better comprehension of successful Human-AI collaboration becomes possible by asking specific questions such as the order in which the AI´s predictions are made available (Collaboration characteristics) or importance of self-assessment capabilities of humans which interact with the AI-system (Human characteristics) (Hemmer et al. 2021). Additional research on Human-AI collaboration is provided by Teodorescu et al. (2021). They present a typology for Human-AI augmentation for achieving fairness and base their typology on the fairness difficulty and locus of decision. Four approaches distinguish the given context: Reactive oversight, proactive oversight, informed reliance, and supervised reliance (Teodorescu et al. 2021). This work contributes to the existing literature by considering the complexity in Human-AI augmentation. Similar suggestions of sharing responsibilities among humans and AI are provided by Trocin et al. (2021): They discuss the collaboration between humans and AI-systems for example through co-creation approaches and provide a future research agenda on Human-AI collaboration. Additional to this body of research on the characteristics and success factors of Human-AI collaboration, two studies analyze and present the benefits of HITL approaches. Here, van den Broek et al. (2021) conducted a two-year ethnography study focusing on the collaboration between domain experts and AI-systems. Their study finds that AI developers and human domain experts arrive at a hybrid, HITL practice, since a process of mutual learning between experts and AI developers showed the interdependence between the involved parties (van den Broek et al. 2021). A review on Human-AI interaction and support for implementing proper oversight mechanisms and keeping humans in the loop is provided by Thiebes et al. (2021).

Technical Robustness and Safety

Our analysis identified only one article that addresses the facet 'Technical robustness and safety'. In their work, Schemmer et al. (2021) deepen the understanding of digital resilience for AI-based information systems against external shocks. They break down AI-systems into three sub-systems which contribute to the overall resilience of AI-systems: the AI model itself, the humans' building and interacting with the AI, and the backend of the system (Schemmer et al. 2021). The authors argue that resilience in these sub-systems can increase the digital resilience of the overall AI-system (Schemmer et al. 2021).

Privacy and Data Governance

As depicted in Table 2, we identified two relevant studies in IS literature that address the facet 'Privacy and data governance' (Thiebes et al. 2021; L. Zhang et al. 2021). Both discuss techniques of implementing AI that respects privacy (i.e., privacy-preserving AI).

Here, besides trusted execution environments, the authors mainly discuss the techniques differential privacy and federated learning (Thiebes et al. 2021; L. Zhang et al. 2021). Differential privacy describes the process of noise-addition to a model or dataset, whereby the data privacy is increased (L. Zhang et al. 2021). Further, federated learning is a paradigm that describes the distributed learning of AI-systems, without sharing the original dataset (L. Zhang et al. 2021). Therefore, federated learning is a suitable technical tool to enhance the data governance related to AI-systems.

Transparency

We identified five relevant contributions that address the facet 'Transparency'. These studies address sources for a lack of AI transparency and solutions to tackle insufficient transparency (Asatiani et al. 2021; Meske et al. 2020; Schneider and Handali 2020; Thiebes et al. 2021; Trocin et al. 2021).

Trocin et al. (2021) touch on possible reasons for the lack of transparency in AI-systems and address the factor of inscrutable evidence. They argue that the lack of AI transparency arises through inscrutable evidence for human observers (Trocin et al. 2021). Meaning that especially for complex deep-learning models, it often remains unclear how single data-points led to different outcomes of AI algorithms. Further, two articles discuss possible solutions to tackle transparency and explainability challenges. Schneider et al. (2019) argue that personalized explanations are necessary for humans to comprehend and understand AI-systems. They propose three key explanation properties for personalized explanations: 1) Considering the complexity (e.g., the number of features of the AI model), 2) Decision information (e.g., prioritization of features that are presented to humans), and 3) Presentation (e.g., choice of visualization techniques) (Schneider and Handali 2020). Asatiani et al. (2021) propose to use the concept of envelopment to tackle the problem of AI explainability. Envelopment is a concept originally adopted from robotics that describes the process of containing AI within defined microenvironments in terms of information processing. In the context of AI-systems, defined microenvironments could be based on the task, for example that AI-systems should only be used for tasks they have been trained on but not on tasks they cannot master (Asatiani et al. 2021). Additionally, there are contributions on the interest of different stakeholders and future research opportunities in explainable AI (Meske et al. 2020; Thiebes et al. 2021).

Diversity, Non-Discrimination, and Fairness

We identified five relevant studies for the facet 'Diversity, non-discrimination, and fairness' in AI (Feuerriegel et al. 2020; Fu et al. 2021; Thiebes et al. 2021; Trocin et al. 2021; von Zahn et al. 2021). These articles address existing biases of AI-systems, cater to a deeper understanding of characteristics, and root causes of biases and possible debiasing mechanisms.

Thiebes et al. (2021) identify current biases in AI-systems, such as racial biases. In addition, Feuerriegel et al. (2020) explain sources of unfairness in AI by breaking down possible sources in three groups: Data, Modeling, and Inadequate applications. The authors argue that understanding unfairness in AI requires understanding possible biases in these three groups (Feuerriegel et al. 2020). Similarly, Trocin et al. (2021) point out that possible biases can occur at multiple stages in the development and deployment lifecycle of AI-systems: Biases might be present in the data itself, emerge at the design and implementation phase through designers and implementors values', and could emerge from technical constraints and challenges (Trocin et al. 2021). Three articles also address possible solutions for fair AI by debiasing mechanisms. Feuerriegel et al. (2020) provide an overview of current algorithms for measuring fairness levels, designing fair predictions, and modeling fair decisions. Additionally, Fu et al. (2021) propose a general debiasing framework through the removal of redundant encoding which leads to the input variables being independent from sensitive attributes. They point out that fair predictions are possible but come with the price of a loss in model performance (Fu et al. 2021). Research on the cost of achieving fair AI is also done by von Zah et al. (2021): They find that implementing fair and unbiased AI comes with additional financial costs, which must be considered. Understanding the added costs of fair AI is important for further operationalizing fair AI in organizations.

Accountability

Our review only identified one article that addresses the facet 'Accountability' of AI-systems: In their work, Trocin et al. (2021) highlight that accountability for AI-systems means to understand the rationale behind the processes that are followed during decision making. They also elaborate on stakeholders involved in the accountability of AI-systems, such as the companies that design and develop AI, the users of the systems, and the parties affected by the outcome of the AI-system (Trocin et al. 2021).

Discussion

In this section we discuss the findings of the conducted analysis. First, based on the identified foci in current IS literature, we highlight areas that haven't been covered sufficiently. Building on this synthesis, we state future directions for IS scholars to further advance the field of EAI. Eventually, we indicate and suggest potential trajectories to address the limitations of our work.

Synthesis

The results of our analysis can be summed up in three major findings: First, we find that all of the identified articles were published recently, with most of it being published in 2021. This strengthens our argument that work on EAI gained momentum in the last years, due to various ethical and social challenges regarding AI being raised. Second, we note that in total there is still little work on EAI in IS related outlets. As illustrated, despite being a highly relevant topic for individuals and the society in general, we could only identify 13 relevant papers in our literature review. This is evidence, that IS research haven't yet fully grasped the relevance of ethical aspects on AI and consequently scientific work on this topic. Based on the fit for IS scholars for working on EAI described, we can conclude that generally more work is needed on all aspects of EAI. Third, we find important differences in the respective areas which have been studied so far by IS scholars. While we identified generally more work on the facets '*Human agency and oversigh*t', '*Transparency*', and '*Diversity, non-discrimination, and fairness*', in comparison we found little work on '*Technical robustness and safety*', '*Privacy and data governance*', and '*Accountability*'.

These unequally distributed areas of focus show, that there is a particular need for future research in these less discussed fields. Hence, the future research directions stated in the next sub-section will address exactly these three areas.

Future Research Directions

Based on the identified facets with little prior work ('*Technical robustness and safety*', '*Privacy and data governance*', and '*Accountability'*), we derive future research directions for IS scholars.

Table 3 summarizes different research directions and the rationale for filling the stated research gaps. We also highlight related prior IS research to build on and the respective EAI facet addressed.

EAI facet addressed	**Related prior IS research**	**Future avenues for IS research**	**Rationale for filling the gap**
'Technical robustness and safety'	Resilience of AI-based information systems	To which extent are consumers appreciating technical safeguards of organizations for robustness and safety?	Knowledge for the organizational decision-making for implementing technical safeguards in consumer facing products and services.
		How to design fallback plans that are sensitive to the given business context?	Knowledge for organizations to successfully implement fallback plans in agile environments.
'Privacy and data governance'	Consequences of privacy-preserving AI	How does federated learning influence inter-organizational collaboration in ecosystems?	Knowledge for organizations how to utilize federated learning for data exchange in competitive markets.
		Which new opportunities emerge from distributed learning paradigms (e.g., federated learning)?	Knowledge on new emerging business models which are based on cross-company data exchange.
'Account-ability'	Root-causes for lack of accountability	What processes can organizations implement to achieve sufficient managerial auditability?	Knowledge on how managers can be efficiently involved in AI auditability, which is comprehensible for them.
		How to distribute accountability for AI between organizations and developers?	Knowledge on which forms of distribution are efficient but also accepted by the workforce.

Table 3. Identified past IS research, future directions, and the rationale for filling the gap

First, in terms of '*Technical robustness and safety*', while we note prior work on resilience, we see that the consumer perspective on technical safeguards and the design of such by organizations haven't been covered sufficiently yet. Hence, we argue that future research should explore if consumers are appreciating implemented technical safeguards when evaluating AI-based products and services, to derive knowledge and incentives for organizations to implement such safeguards. Further, future research should create a more thorough understanding of how to design fallback plans, to derive practical knowledge for organizations for implementing fallback plans in agile business environments.

Second, regarding the facet *'Privacy and data governance'* we see that current IS coverage is lacking an understanding of organizational implications of implementing privacy-preserving AI (e.g., Federated learning). In this context, we propose that future research should gain an understanding of how, for example federated learning, can be used for inter-organizational collaboration, to generate an understanding of how such techniques can be utilized by organizations to profit from positive effects such as increased data exchange. In this context, we also argue that future research should gain a more in-depth understanding of new opportunities offered by distributed learning paradigms, for example by generating knowledge on how organizations can build new digital business models that are based on cross-company data exchange to build their products and services.

Third and last, in terms of 'Accountability', we note that especially the organizational focus on AI auditability and accountability haven't been covered sufficiently. In this context, auditability describes the ability of organizations to audit AI. Hence, we propose that future research should build an understanding of managerial AI auditability, to design organizational processes which are efficient and understandable for managers. Further, we argue that IS scholars should address questions regarding the distribution of accountability between developers and organizations, to gain more knowledge on which concepts of shared accountability are efficient in everyday work but also accepted by the developers.

In summary, we see that there are multiple research gaps for which future research avenues are identified. Filling this research gaps respectively will lay an initial socio-technical understanding, which in turn can be used by companies to adapt to the regulatory actions initiated by the AI Act and thus yielding a high level of practical relevance.

Limitations

Besides our analysis and contribution, we note three areas in which future research could further strengthen our results: First, we recognize that we limited our search by only including articles in IS related outlets, hence did not include valuable work on EAI that exist outside of IS outlets. However, as illustrated in the introduction as well, we see a special fit for the domain of IS on working on EAI, by providing a holistic socio-technical perspective which goes beyond the current, mainly technical perspective of scholars from the computer science community. Based on this fit, we specifically wanted to understand how the state on EAI is in IS related outlets, to gain an appropriate understanding of current foci and blind spots on EAI in IS research. Nevertheless, we note that future research could extend our findings by incorporating work outside of IS related outlets. Second, we must note that we did not consider prior research on related topics that could be mapped to EAI: For example, there is a rich body of work in IS literature on data privacy in IS which could cater to the privacy of AI-systems. Similarly, there is considerable IS literature on managerial accountability for information systems, from which learnings could be transferred to the context of AI-systems. Third, regarding the execution of the literature review, we can add more keywords and synonyms (e.g., 'human oversight' or 'auditability') into the search query, include more databases, or use more sophisticated information retrieval tools such as topic modelling.

Conclusion

Due to the ethical and social challenges surrounding AI, we consider research on EAI as very important and necessary. As such, the goal of this work is to show the role that the IS community can play in understanding and evolving EAI. Therefore, it is essential to thoroughly understand the current research space of EAI in common IS outlets. Our analysis showed, that while there is some prior work on aspects of EAI in IS outlets, the current IS research space for EAI is still very limited. In particular, we find that studies regarding '*Technical robustness and safety*', *'Privacy and data governance',* and *'Accountability'* are scarce. Consequently, we formulated clear research directions for IS scholars to follow and also stated the respective rationale to show the motivation for future work in this direction. Taking these research directions seriously and advancing the field of EAI will become even more important when considering that the EU will foreseeably enact the AI Act in the upcoming year. The AI Act will likely set a new ethical benchmark that might serve as a blueprint for AI regulation in many other countries and regions.

In conclusion, we are confident that this paper provides a comprehensive understanding of the facets of EAI as stated in the AI Act and the current state of IS research based on these facets. We aspire this understanding and the stated research directions to stimulate future research and motivate IS scholars to engage in this emerging field to facilitate a more ethical use of AI.

REFERENCES

Act AI 2021. "LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS." Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., and Salovaara, A. 2021. "Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems," *Journal of the Association for Information Systems* (22:2), pp. 325–352.

Bates, D. W., Levine, D., Syrowatka, A., Kuznetsova, M., Craig, K. J. T., Rui, A., Jackson, G. P., and Rhee, K. 2021. "The Potential of Artificial Intelligence to Improve Patient Safety: A Scoping Review," *Npj Digital Medicine* (4:1), Springer US, pp. 1–8.

Bradford, A. 2020. *The Brussels Effect: How the European Union Rules the World*, Oxford University Press.

Brocke, J. vom, Simons, A., Niehaves, Bjoern, Niehaves, Bjorn, Reimer, K., Plattfaut, R., and Cleven, A.

2009. "Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process," *CIS 2009 Proceedings* (372).

Dastin, J. 2018. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women."

Feuerriegel, S., Dolata, M., and Schwabe, G. 2020. "Fair AI: Challenges and Opportunities," *Business and Information Systems Engineering* (62:4), pp. 379–384.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., and Vayena, E. 2018. "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and Machines* (28:4), Springer Netherlands, pp. 689–707.

Fu, R., Huang, Y., and Singh, P. V. 2021. "Crowds, Lending, Machine, and Bias," *Information Systems Research* (32:1), pp. 72–92.

Hardt, M., Price, E., and Srebro, N. 2016. "Equality of Opportunity in Supervised Learning," *Advances in Neural Information Processing Systems*, pp. 3323–3331.

Hemmer, P., Schemmer, M., and Kühl, N. 2021. "Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review," *Twenty-Fifth Pacific Asia Conference on Information Systems*.

HLEG 2019. *ETHICS GUIDELINES FOR TRUSTWORTHY AI*. European Commission.

Imana, B., Korolova, A., and Heidemann, J. 2021. "Auditing for Discrimination in Algorithms Delivering Job Ads," *Proceedings of the World Wide Web Conference 2021*, pp. 3767–3778.

Lovejoy, B. 2019. "GDPR-Style Privacy Law in California Acting as a 'Blueprint' for Other States." Retrieved from: https://9to5mac.com/2019/12/27/gdpr-style-privacy-law/.

Mayr, A., Kißkalt, D., Meiners, M., Lutz, B., Schäfer, F., Seidel, R., Selmaier, A., Fuchs, J., Metzner, M., Blank, A., and Franke, J. 2019. "Machine Learning in Production - Potentials, Challenges and Exemplary Applications," *7th CIRP Global Web Conference* (86), Elsevier B.V., pp. 49–54.

Meske, C., Bunde, E., Schneider, J., and Gersch, M. 2020. "Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities," *Information Systems Management*, Taylor & Francis, pp. 1–11.

OECD 2019. "OECD AI Principles Overview." Retrieved from: https://oecd.ai/en/ai-principles.

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., and Erlingsson, Ú. 2018. "Scalable Private Learning with Pate," *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, pp. 1–34.

Park, S. Y., Kuo, P. Y., Barbarin, A., Kaziunas, E., Chow, A., Singh, K., Wilcox, L., and Lasecki, W. S. 2019. "Identifying Challenges and Opportunities in Human-AI Collaboration in Healthcare," *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, pp. 506–510.

Schemmer, M., Heinz, D., Baier, L., Vössing, M., and Kühl, N. 2021. "Conceptualizing Digital Resilience for AI-Based Information Systems," *ECIS 2021 Research-in-Progress Papers*, pp. 6–14.

Schneider, J., and Handali, J. P. 2020. "Personalized Explanation for Machine Learning: A Conceptualization," *27th European Conference on Information Systems*.

Teodorescu, M. H. M., Morse, L., and Kane, G. C. 2021. "Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation," *MIS Quarterly* (45:3), pp. 1483–1499.

Thiebes, S., Lins, S., and Sunyaev, A. 2021. "Trustworthy Artificial Intelligence," *Electronic Markets* (31:2), pp. 447–464.

Trocin, C., Mikalef, P., Papamitsiou, Z., and Conboy, K. 2021. "Responsible AI for Digital Health: A Synthesis and a Research Agenda," *Information Systems Frontiers*, Information Systems Frontiers.

van den Broek, E., Sergeeva, A., and Huysman, M. 2021. "When the Machine Meets the Expert: An Ethnography of Developing Ai for Hiring," *MIS Quarterly* (45:3), pp. 1557–1580.

von Zahn, M., Feuerriegel, S., and Kuehl, N. 2021. "The Cost of Fairness in AI: Evidence from E-Commerce," *Business and Information Systems Engineering*, Springer Fachmedien Wiesbaden.

Zhang, L., Shen, B., Barnawi, A., Xi, S., Kumar, N., and Wu, Y. 2021. "FedDPGAN: Federated Differentially Private Generative Adversarial Networks Framework for the Detection of COVID-19 Pneumonia," *Information Systems Frontiers*.

Zhang, P., Guo, Z., Ullah, S., Melagraki, G., Afantitis, A., and Lynch, I. 2021. "Nanotechnology and Artificial Intelligence to Enable Sustainable and Precision Agriculture," *Nature Plants* (7:7), Springer US, pp. 864–876.