Ethical AI Research Untangled: Mapping Interdisciplinary Perspectives for Information Systems Research

Manoj Kahdan
RWTH Aachen University, kahdan@time.rwth-aachen.de

Nicole Janine Hartwich
RWTH Aachen University, hartwich@time.rwth-aachen.de

Oliver Salge
RWTH Aachen University, salge@time.rwth-aachen.de

Ethical AI Research Untangled: Mapping Interdisciplinary Perspectives for Information Systems Research

Short Paper

Manoj Kahdan, Nicole Hartwich, Torsten-Oliver Salge
Institute for Technology and Innovation Management, RWTH Aachen University
https://www.rwth-aachen.de/digital-responsibility-lab
{kahdan, hartwich, salge}@time.rwth-aachen.de

Abstract

We provide a systematic overview of the interdisciplinary discourse on ethical AI by combining bibliometric and text mining approaches on a corpus of 23,870 ethical AI publications from journals and conference proceedings. In our research in progress, we offer three contributions of interest to IS scholars: First, in our term analyses, we empirically delineate ethical AI and related terms such as responsible or trustworthy AI. Second, we unearth the intellectual structure of the field and identify five thematic clusters, some of which are directly relevant to IS scholars. Third, we identify that IS research on ethical AI should more intensely consider fairness and transparency as well as the link to explainability. Additionally, we suggest that IS scholars contribute towards policymakers' ethical AI guidelines by contributing their strong expertise in practical applications.

Keywords: Ethical AI, Responsible AI, Trustworthy AI, Explainable AI, Transparent AI

Introduction

Artificial intelligence (AI) is one of the crucial technologies of our century and is paving the way for a new age. It has become an irreplaceable part of current technological developments and profoundly impacts our daily lives (Guo et al., 2020). Machine learning in its manifold and rapidly evolving forms changes how we handle data and generate knowledge (Jordan & Mitchell, 2015). AI offers vast economic potential, ranging from simple process optimization and automation to new AI-based products, new predictive services and entirely new business areas. Moreover, AI gives rise to new applications and associated new business models like autonomous driving or (semi-)autonomous systems in production systems (Jobin et al., 2019). In addition to businesses, AI applications also profoundly impact our social environment.

However, with such fundamental changes, Information System (IS) research needs to go beyond the technological and economic opportunities and threats. It is not only the optimization and cost savings when weighing up the costs and benefits of AI application, but also the implications for people and society, e.g., gender fairness (Teodorescu et al., 2021), privacy issues relating to DeepFakes (Thiebes et al., 2021), unfair decision-making (Mariani et al., 2022), or mail fraud, cyber-crime, and cyber-attacks (Taddeo, 2018). Overall, AI applications face technological, ethical, legal, and social challenges (Thiebes et al., 2021). Therefore, a reflection on AI's societal implications and possible moral and ethical implications from a holistic perspective is essential. Since AI applications are often based on particularly large amounts of data and the use of high-dimensional statistical models, it is difficult for users to verify to what extent the properties promised by the manufacturer or provider are fulfilled. A discussion on the responsible use of AI and related topics have become increasingly prevalent in recent years and is currently part of cross-thematic debates within research and society (Breidbach & Maglio, 2020; Floridi et al., 2018; Guidotti et al., 2019). Current literature indicates a shift within the ethical AI discussion from principles-based ethics to the practical implementation of ethical principles (Seppälä et al., 2021).

The actual role and impact of ethical AI within the IS discipline remains, however, unclear, as the existing AI ethics literature dilute the overall picture of this essential topic (Mariani et al., 2022; Trocin et al., 2021; Wamba et al., 2021; Wamba & Queiroz, 2021; Zhang et al., 2021). For instance, Wamba and Queiroz (2021) and Trocin et al. (2021) focus on ethical AI in digital health, Mariani et al. (2022) examine the field of ethical AI intersection of marketing and psychology and Zhang et al. (2021) analyzes ethical AI relevant publication by means of text mining techniques in three journals. However, these aforementioned studies investigate ethical AI with a narrow focus on specific disciplines rather than providing a holistic overview. In contrast, Wamba et al. (2021) present a comprehensive overview on AI research, evaluating 40,147 articles. While well-suited for understanding research on AI in general, they do not provide a detailed mapping of ethical AI. The latter is particularly important for several reasons: First, the key terms such as "ethical AI", "trustworthy AI", "moral AI", "responsible AI", and "explainable AI" used in the literature are not always clearly distinguishable and are used interchangeably across and even within disciplines. The interchangeability of the terms leaves researchers and practitioners alike confused concerning whether or not these terms are similar or relay to different subtopics. Second, due to the current relevance of the topic, there is a large and rapidly growing number of existing publications on AI ethics. It appears vital at this point to take stock and portray how the interdisciplinary debate on ethical AI is being conducted in the academic world, has evolved over time, and is likely to impact future ethical AI research in IS. Such synthesis and consolidation effort is much needed to enhance definitional clarity, structure the debate and delineate recommendations for future research on ethical AI as well as for managerial practice.

We provide an explorative study of the extensive body of ethical AI publications from an interdisciplinary perspective. For this purpose, we apply state-of-the-art text mining approaches on 23,870 ethical AI paper published in journals and conferences. Text mining techniques follow a systematic and objective process and derive hidden structures from large amounts of data and facilitate the comprehension of complex thematic dynamics and interdependencies for scholars (Barth et al., 2020; Brust et al., 2017; Debortoli et al., 2016). With our approach, we explore the composition of current publications and to provide future researchers with a more transparent and precise overview of terminologies regarding the debate on ethical AI.

We thereby contribute to IS research in at least three meaningful ways: First, we contribute to IS research by empirically establishing a clear and simple terminology for the field of ethical AI. We highlight the need for IS researchers to consider morality and trust in ethical aspects of AI. Moreover, research on explainable AI should also include perspectives of transparency and fairness to provide holistic insights. Second, in our interdisciplinary analyses, we uncover a tripartite thematic structure in the ethical AI discourse: Based on the word co-occurrence analysis, technical implementation, the application of AI, especially in the medical and natural science areas, and the basic discourse on ethical AI can be derived. Third, we discover the enormous potential for IS scholars to contribute towards the discourse on ethical AI in two meaningful ways: IS research on ethical AI should consider fairness and transparency aspects and the link to explainability more intensely. Further, we propose IS scholars to contribute towards policymakers' ethical AI guidelines by contributing the strong expertise in practical application. Thus, there is great potential for development in the ethical discourse.

Perspectives on Ethical AI

The current discussion on ethical AI is rich in variety. Given the potential impact of AI at all levels, researchers, legislators, and regulators across countries and companies discuss the ethically responsible use of AI and establish framework structures (Hagendorff, 2020; Jobin et al., 2019). In the existing discourse, different terminologies such as "ethical AI" (Breidbach & Maglio, 2020; Mariani et al., 2022; Mittelstadt, 2019), "transparent AI" (Wachter et al., 2017), "trustworthy AI" (European Commission, 2019; Mökander & Floridi, 2021; Thiebes et al., 2021), "explainable AI" (Lundberg et al., 2020), or "responsible AI" (Trocin et al., 2021) are often used interchangeably without clear distinction. For example, "explainable AI" aims to improve trust and transparency of AI-enabled systems (Lundberg et al., 2020). Additional terminologies such as "beneficial AI" and "responsible AI" are also currently discussed (Trocin et al., 2021). Given the broad impact of ethical AI on humans and society, related themes within the ethical discourse like fairness, bias, security, and privacy also play an essential role (Mariani et al., 2022; Zhang et al., 2021) where security and privacy as well as fairness and bias share close links (Jobin et al., 2019).

Further, Breidbach and Maglio (2020) highlight the interchangeable use of moral and ethics and the need to include additional perspectives by assessing AI from an explainable and responsible view. They emphasize that questions regarding how decisions are made and who is responsible for the decision arise from, for example, the use of AI techniques like unsupervised deep learning algorithms. According to Thiebes et al. (2021), trust is a sustainable foundation for society, individual, and economics and therefore plays a crucial role in AI development. Additionally, Floridi (2019) highlights that AI systems should be developed responsibly. Mökander and Floridi (2021) propose the responsible development and quality improvement of automated decision making through an ethics-based auditing approach.

In addition to the extensive scientific discussion on trustworthy AI and related perspectives, the topic is also of high practical relevance for policymakers. For example, the European Commission emphasises that "trustworthiness is a prerequisite for people, and societies to develop, deploy and use AI systems" in their ethical guidelines for trustworthy AI, which where co-developed with an interdisciplinary group of 52 experts from business, science, and regulation (European Commission, 2019). According to these guidelines, trust in the design, development, deployment, and use of AI systems is about the inherent properties of the technology and the qualities of the socio-technical systems that comprise AI applications. They also identify a strong interconnection between ethical AI perspectives regarding transparent and explainable AI as an enabler for the successful implementation of trustworthy AI. Jobin et al. (2019) conducted a policy review, bringing together the broad discourse in the ethical AI field by analyzing 84 AI publications containing ethical principles and clustering them according to five key aspects: transparency, justice and fairness, non-maleficence, responsibility and privacy. However, only nine of these publications stem from academic research institutes.

Additionally, Seppälä et al. (2021) offer empirical evidence for a recent shift in attention from identifying and outlining ethical AI principles towards implementing them into practice. According to Hagendorff (2020) such implementation could be stifled by not providing detailed technical explanations in such policy guidelines. Floridi (2021) and Hagendorff (2020) point out that even if ethical AI principles are applied, doing so for the wrong reasons may undermine their intended propose. As an example, they name ethical AI implementation that is primarily driven by competitiveness and merely serve as a differentiation tool for marketing. Moreover, a lack of established methods to transform norms into practice and the missing legal accountability might be obstacles to transforming AI ethics principles into AI design (Mittelstadt, 2019). Further, principles on AI ethics may have a bounded effect on the design of explainable AI (Mittelstadt, 2019).

In this work, we present and examine the presented perspectives and highlight their similarities, differences, and relation to the related themes like privacy and fairness.

Research Method and Data

We used Web of Science's Core Collection to identify the interdisciplinary corpus of ethical AI research. This vast database is appropriate and widely accepted in the research community to design an analysis that covers as many research areas as possible (e.g., Belmonte et al., 2020; Cetindamar et al., 2022; Wamba & Queiroz, 2021; Zhang et al., 2021). The traditional limitation to individual research areas is deliberately omitted to ensure the completeness of the review and map the rich research landscape on ethical AI across scientific disciplines. We implemented the following search strategy from the article titles, abstracts, and keywords to select the relevant publications from the database. From the regulatory and scientific considerations, as well as existing fuzzy descriptions and interchangeably perspectives we summarize the presented ethical AI aspects into six core perspectives: "ethical AI", "moral AI", "responsible AI", "transparent AI", "trustworthy AI" and "explainable AI". The topic search string includes combinations of AI and AI-specific related terms and techniques with the identified perspectives on ethical AI: *TS = (("artificial intelligence" OR "machine learning" OR "neural network" OR "deep learning" OR "reinforcement learning" OR "supervised learning" OR "unsupervised learning" AND ("ethic*" OR "moral*" OR "responsi*" OR "transparen*" OR "trust*" OR "explainab*")).* We selected only English articles and conference proceedings published between 1990 and 2022, totaling 23,870 publications (*dataset* in the following). The integrity of our text corpus was evaluated by manually verifying that title, abstract, and keywords of a random sub-sample belong to the field of ethical

AI. Within this database, 3,937 publications (*subset* in the following) were categorized as IS Research by Web of Science's Core Collection. Text preprocessing consisted of stop word, punctuation, and number removal as well as lower-casing and lemmatizing the text corpus.

Methodologically, we followed an established two-step process. First, we performed a traditional bibliometric analysis, which is an appropriate and powerful tool for understanding the structure and evolution of one or more research streams (Mishra et al., 2018; Wamba & Queiroz, 2021; Zhang et al., 2021). This uncovered the most relevant authors, institutions, and topics in the ethical AI landscape. Second, we used state-of-the-art text mining approaches to identify networks and relations within the corpus (Zhang et al., 2021). In doing so, we employed the R package bibliometrix (Aria & Cuccurullo, 2017), VosViewer (Van Eck & Waltman, 2013), and Word2Vec (Mikolov et al., 2013) for further text analysis of titles, abstracts, and keywords as well as the application of word co-occurrence networks and word embeddings on our text corpus. Word2Vec enables word embedding using vector representations of words. The significant advantage here is the objective representation of words and terms that describe these words. The core idea is to represent words with the same meaning in similar vectors. The main benefit is to search the entire corpus for ethical and related terms to clearly describe them and distinguish them from each other or represent similarities.

Preliminary Results

Descriptive Analysis

The first analysis briefly describes the initial bibliometric results regarding the thematic field of ethical AI from an interdisciplinary perspective. As shown in Figure 1, only a slow increase in annual publications before can be observed up until 2015. Contrary, the exponential growth in the number of publications related to ethical AI since 2016 is particularly noteworthy. The exponential growth in number of annual ethical publications on AI since 2016 is particularly noteworthy and clearly illustrates the rapidly emerging discourse on ethical AI across disciplines.

Analyzing the top 30 affiliations and grouping them by countries, we observe: US-based affiliations generated 2,810 articles, UK-based affiliations 1,064 articles, French affiliations 594 articles, Chinese affiliations 489 articles, Singaporean affiliations 483 articles, Indian affiliations 387 articles, Canadian affiliations 212 articles, German affiliations 160 articles, and Egyptian affiliations generated 154 articles. From this, we observe an apparent North-American influence within the debate on ethical AI. Approximately half of the current publications belong to US or Canadian affiliations. The other half is approximately twofold between European and Asian institutions. Moreover, these top 30 affiliations represent in total 27% of the entire underlying publications within our dataset.

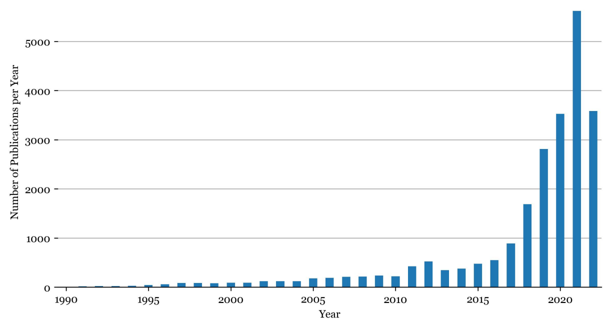

Figure 1. Annual Publications 1990 – 2022 (dataset)

Textual Analysis

The second section draws on text mining and provides greater analytical depth with regards to term and topic structure of the ethical AI research landscape. First, we use word embedding to identify relations between the different terminologies and perspectives on ethical AI. Second, our text mining approach maps the word clusters within the underlying text corpus based on the entire interdisciplinary corpus. Third, we focus on the textual analysis of only IS articles to enhance the ethical AI discussion with IS research.

Ethical AI: Terminology

First, we investigated similarities in the meaning of related terms associated to the different perspectives of ethical AI. We used word embedding to map the perspectives, providing objective empirical evidence for a more precise and explicit distinction between terms used in the academic discourse on ethical AI.

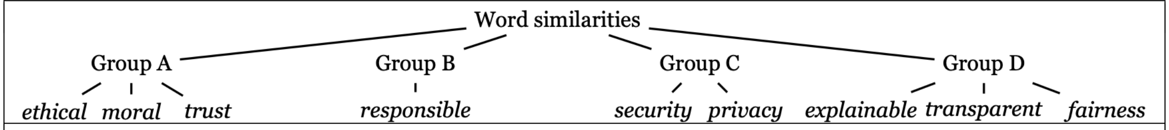

Figure 2. Word Similarities Using Word2vec

In Figure 2, the underlying multidimensional Word2Vec model was reduced to a two-dimensional representation using t-Distributed Stochastic Neighbor Embedding to enable visualization of term relations. Due to the spatial proximity of the points, a similar meaning of terms can be derived. Here, Group A represents the related perspectives of ethics, moral, and trust. Group D highlights the relations of explainable, transparent, and fair AI. In Group C, a strong relationship between security and privacy is provided. These related themes are highly within interconnected, as seen in Table 1, but still a part of the ethical, responsible, and explainable AI discourse. Group B is represented by the term responsible. A detailed overview of similarities and differences between the perspectives and related terms is provided in Table 1. Here, transparent and explainable terms have the descriptive word interpretable in common and so on. As a robustness check, we run the same analysis on a text corpus consisting only of journal articles, and observed highly similar results.

Perspectives	**Group of seven semantically most closely related words**						
ethical	ethic	normative	legal	ethically	moral	ethics	practice
moral	morality	morally	virtuous	kantian	ethical	semiotic	normative
trust	trustworthiness	reputation	credibility	confidence	reliability	trusting	trustee
responsible	involved	corresponds	affecting	influencing	underlie	activated	involves
security	cybersecurity	hacker	internet	cyber	privacy	attacks	securing
privacy	confidentiality	microaggregation	anonymity	personal	protection	hipaa	security
explainable	interpretable	explainability	explaining	models	explanations	xai	interpretability
transparent	understandable	opaque	comprehensible	auditable	interpretable	accurate	simple
fairness	bias	fair	algorithmic	unfairness	discrimination	transparency	accountability

Table 1. Different Perspectives on Ethical AI and Relevant Themes Using Word2vec

Ethical AI: Interdisciplinary Perspectives

We generated a co-occurrence word map based on titles and abstracts to better understand the network structure of key terms coined within the interdisciplinary field of ethical AI, as depicted in Figure 3 (interdisciplinary dataset). In the visualization, nodes represent individual terms, with their size scaled by occurrence frequency. Links between nodes indicate co-occurrence of the corresponding terms within publications. Thus, larger and more densely interconnected nodes represent more relevant terms in the text corpus. Clustering according to Waltman et al. (2010) yielded the following five thematic clusters:

Blue Cluster (Discourse on Ethical AI): Represents terms related to the core of the ethical discourse. The most frequent and connected terms are ethics, AI, acceptance, guidelines, and morality, as well as terms related to the medical sector like healthcare, radiology, and care. From an interdisciplinary perspective, the identified ethical AI discourse is strongly linked to the fields of application in healthcare and privacy/security. However, the link to technical implementation and application in the natural sciences is sparse.

Green Cluster (Application (Privacy/Security)): Representing the field of AI application in the field of privacy and security, this cluster includes terms like IoT, attack, blockchain, node, privacy, and phishing. Characterized by relatively small and densely interconnected nodes, this cluster has strong links only to the discourse on ethical AI.

Red Cluster (Application (Healthcare)): The terms patient, brain, neuron, disease, disorder, and covid describe this cluster. It is characterized by relatively large and dense nodes with strong links to the discourse on ethical AI, technical implementation, and application in natural science.

Violet Cluster (Application (Natural Science)): Terms like artificial neural network, concentration, forecasting, substrate, and chemical characterize this cluster. Despite strong links to AI application in healthcare, links to the discourse on ethical AI are almost non existent.

Yellow Cluster (Technical Implementation): This cluster is described by cnn (convolutional neural network), segmentation, detection method, xai model (explainable AI), and training sample function. Further, it includes healthcare terms like diagnosis, imaging, and radiologist, indicating an overlap with application in healthcare. Despite sparse intraconncetedness, direct links exists to all other clusters.

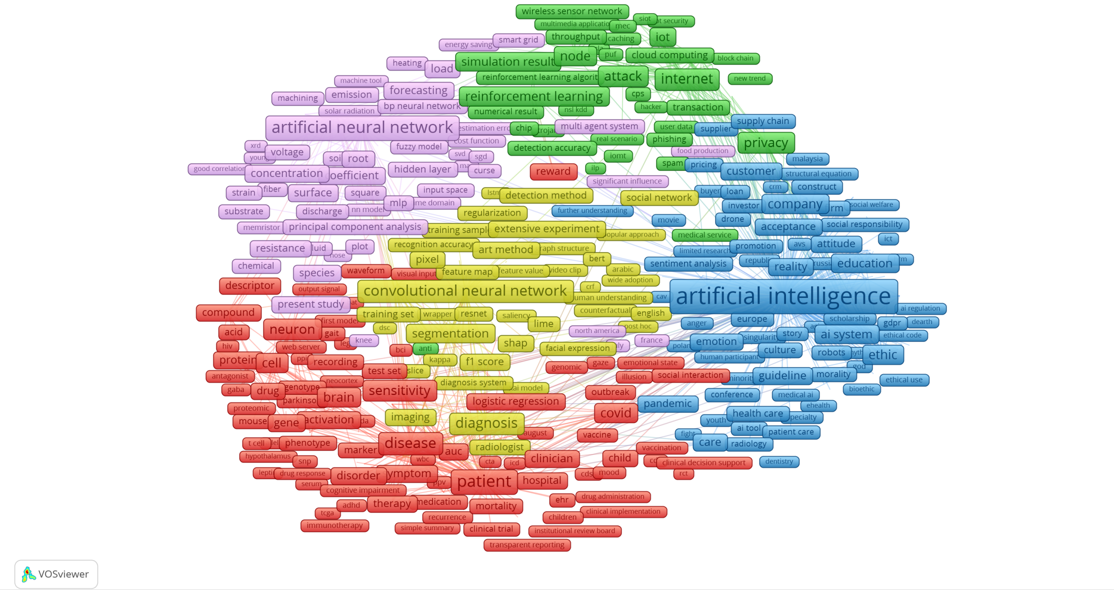

Figure 3. Interdisciplinary Co-occurrence Map Network With Five Clusters

Ethical AI: IS Perspectives

Figure 4 visualizes the co-occurrence map within the IS corpus (subset), where clustering yielded only four clusters. Similarities and differences between ethical AI in IS research and the interdisciplinary perspective on ethical AI in the interdisciplinary dataset can be observed through comparison with Figure 3. Here, the previously described violet cluster is not present within the subset.

Blue IS Cluster (Discourse on Ethical AI): Ethic, fairness, government, and explainable are relevant terms. We observe a sparse network and fewer intra-cluster links compared to the interdisciplinary dataset. This indicates comparatively low dissemination of discourse on ethical AI within IS research. Notably, this cluster is relatively well connected with the application in privacy/security and healthcare for both networks. Noteworthy is the vital connection regarding the “explainable AI” perspective between the clusters discourse on ethical AI, technical implementation and application in healthcare.

Green IS Cluster (Application (Privacy/Security)): Privacy, attack, device, and threat are relevant terms. This network is relatively extensive and highly interconnected. Similar to the multidisciplinary discourse, it is well connected to the blue and red clusters, indicating a strong relationship towards healthcare application and ethical discourse. Moreover, compared to the interdisciplinary network, the IS application on privacy and security is thematically broader.

Red IS Cluster (Application (Healthcare)): Classification, patient, disease, explanation, image, and cnn represent mainly this cluster. Compared to the interdisciplinary cluster, this IS cluster strongly focuses on the implementation and explainability within the healthcare case, indicating a more practical direction. It is well connected to the ethical AI and privacy cluster. However, this link towards the privacy cluster is not visible in the multidisciplinary network.

Yellow IS Cluster (Technical Implementation): Explainability, representation, recommendation, social network, nlp (natural language processing), and tweet indicate a strong focus on the technical implementation of AI within this cluster. Similar to the interdisciplinary cluster, this cluster is well linked to the healthcare application cluster. The presence of nodes close to all other clusters in the IS network indicates a solid relevance to technical implementation with the ethical discourse.

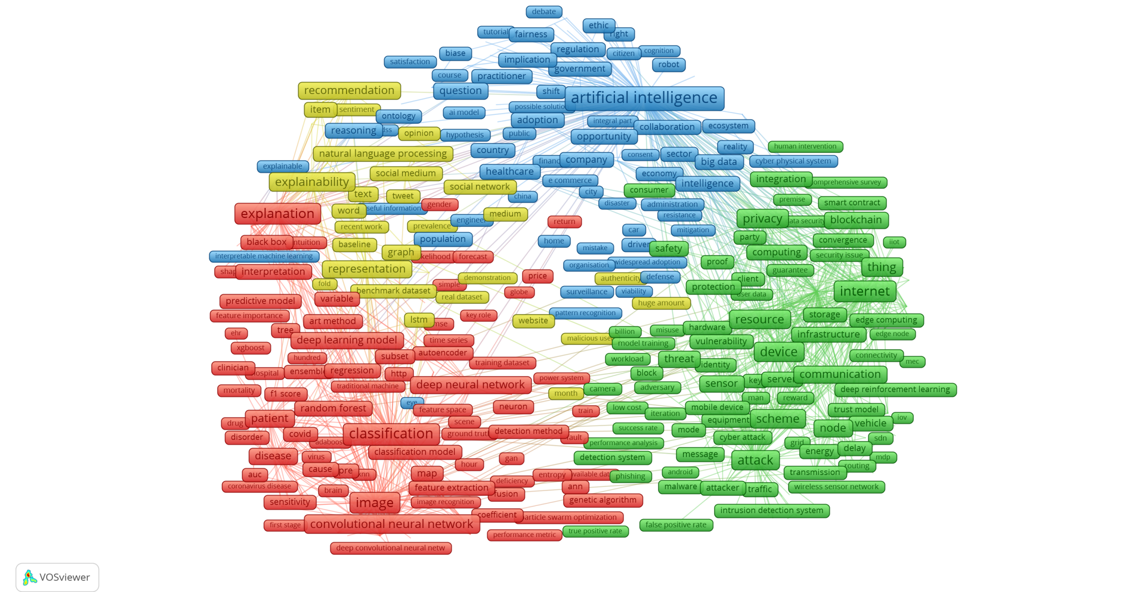

Figure 4. IS Co-occurrence Map Network With Four Clusters

Discussion

In this work, we contribute to IS research by empirically establishing a clear and simple terminology for the field of ethical AI, thus untangling semantically overlapping and interchangeably used terms for scholars. First, based on our empirical evidence, we highlight the need for IS researchers to consider morality and trust in ethical aspects of AI. Moreover, research on explainable AI should also include perspectives of transparency and fairness to provide holistic insights and not accidentally disregard research because certain terms may not be searched for.

Second, in our interdisciplinary analysis we uncover a tripartite thematic structure in the ethical AI discourse: Based on the word co-occurrence analysis, technical AI implementation, the application of AI (especially in the healthcare and privacy/security domains) and the scientific discourse on ethical AI can be derived. With this thematic structure, we provide an overview for non-experts in the ethical AI domain.

Third, we discover the unrealised potential for IS scholars to contribute to ethical AI research in two meaningful ways: On the one hand, our study shows a strong relation between explainability and fairness / transparency in ethical AI, whereas the analysis of our IS research corpus identified a strong focus on explainability but not on fairness nor transparency. Future IS research on ethical AI should hence fill this research gap by considering fairness and transparency aspects and the link to explainability more intensely. With its knowledge of explainability, IS can contribute to the fairness / transparency discussion on an interdisciplinary level. On the other hand, IS scholars could help overcome the lag of technical explanations in policymakers' ethical AI guidelines (Hagendorff, 2020) by contributing the strong expertise in practical application, especially in the healthcare and privacy/security domains, that is prevalent in IS literature. Through such a translation from theoretical and practical research experience into operationalizable guidelines, IS can make a meaningful contribution to the practical adoption of ethical AI by society.

Limitations and Future Work

Our research provides a broad interdisciplinary perspective on ethical AI. It explores the composition of current publications and provides future researchers with a transparent and precise overview of terminologies in ethical AI. Specifically, we identified and visualized an interdisciplinary landscape of extant research on ethical AI. We thereby provide the groundwork to the still young discussion of ethical AI in IS Research. However, as for this present research-in-progress, we do see potential to expand our analyses in several ways. Our results rely only on a single database and are thus limited. Yet, we are confident of overcoming this limitation by extending our database and including the Scopus database in the further course of our work. The current research progress in word embedding analysis only allows the ethical debate to be considered at the term level. To overcome this limitation, we will further analyse the ethical AI discourse on full text level by using Latent Dirichlet Allocation (LDA) topic modelling.

We also expect an interconnection of ethical AI topics with, for example, moral and responsible AI topics. However, we suspect connections between the analyzed terms at the topic level. Of interest for IS research would be the networks between the three identified groupings. The aim here is to create a granular, in-depth overview of ethical discourse in general and IS research. Further, applying our methodology to text corpus snapshots over time and comparing the outcomes, thus mapping the evolution of ethical AI could also contribute to the discourse on ethical AI. Moreover, the method can be used to identify similarities and differences between ethical AI perspectives within the identified clusters. Future research could contribute to the discourse by studying the perspectives in different domains such as healthcare and natural science.

References

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Barth, J. R., Herath, H. S., Herath, T. C., & Xu, P. (2020). Cryptocurrency valuation and ethics: A text analytic approach. *Journal of Management Analytics*, *7*(3), 367–388. https://doi.org/10.1080/23270012.2020.1790046

Belmonte, J. L., Segura-Robles, A., Moreno-Guerrero, A.-J., & Elena Parra-Gonzalez, M. (2020). Machine Learning and Big Data in the Impact Literature. A Bibliometric Review with Scientific Mapping in Web of Science. *Symmetry-Basel*, *12*(4). https://doi.org/10.3390/sym12040495

Breidbach, C. F., & Maglio, P. (2020). Accountable algorithms? The ethical implications of data-driven business models. *Journal of Service Management*, *31*(2), 163–185. https://doi.org/10.1108/JOSM-03-2019-0073

Brust, L., Breidbach, C. F., Antons, D., & Salge, T.-O. (2017). Service-Dominant Logic and Information Systems Research: A Review and Analysis Using Topic Modeling. *International Conference on Information Systems 2017 Proceedings*. https://aisel.aisnet.org/icis2017/ServiceScience/Presentations/7

Cetindamar, D., Kitto, K., Wu, M., Zhang, Y., Abedin, B., & Knight, S. (2022). Explicating AI Literacy of Employees at Digital Workplaces. *IEEE Transactions On Engineering Management*. https://doi.org/10.1109/TEM.2021.3138503

Debortoli, S., Müller, O., Junglas, I., & Vom Brocke, J. (2016). Text mining for information systems researchers: An annotated topic modeling tutorial. *Communications of the Association for Information Systems*, *39*(1), 7. https://doi.org/10.17705/1CAIS.03907

European Commission. (2019). *Ethics Guidelines for Trustworthy AI*. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

Floridi, L. (2019). Establishing the rules for building Trustworthy AI. *Nature Machine Intelligence*, *1*(6), 261–262. https://doi.org/10.1038/s42256-019-0055-y

Floridi, L. (2021). Translating principles into practices of digital ethics: Five risks of being unethical. In L. Floridi (Ed.), *Ethics, governance, and policies in artificial intelligence* (pp. 81–90). Springer International Publishing. https://doi.org/10.1007/978-3-030-81907-1_6

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Guidotti, R., Monreale, A., Ruggieri, S., Turin, F., Giannotti, F., & Pedreschi, D. (2019). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, *51*(5). https://doi.org/10.1145/3236009

Guo, Y., Hao, Z., Zhao, S., Gong, J., & Yang, F. (2020). Artificial Intelligence in Health Care: Bibliometric Analysis. *Journal Of Medical Internet Research*, *22*(7). https://doi.org/10.2196/18228

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *SCIENCE*, *349*(6245, SI), 255–260. https://doi.org/10.1126/science.aaa8415

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Mariani, M. M., Perez-Vega, R., & Wirtz, J. (2022). AI in marketing, consumer research and psychology: A systematic literature review and research agenda. *Psychology & Marketing*, *39*(4), 755–776. https://doi.org/10.1002/mar.21619

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. https://doi.org/10.48550/arXiv.1301.3781

Mishra, D., Gunasekaran, A., Papadopoulos, T., & Childe, S. J. (2018). Big Data and supply chain management: a review and bibliometric analysis. *Annals of operations research*, *270*(1-2, SI), 313–336. https://doi.org/10.1007/s10479-016-2236-y

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. *Minds and Machines*, *31*(2), 323–327. https://doi.org/10.1007/s11023-021-09557-8

Seppälä, A., Birkstedt, T., & Mäntymäki, M. (2021). From Ethical AI Principles to Governed AI. *International Conference on Information Systems 2021 Proceedings*. https://aisel.aisnet.org/icis2021/ai_business/ai_business/10

Taddeo, M. (2018). The limits of deterrence theory in cyberspace. *Philosophy & Technology*, *31*(3), 339–355. https://doi.org/10.1007/s13347-017-0290-2

Teodorescu, M. H., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of Fairness in Automation Require a Deeper Understanding of Human-Ml Augmentation. *MIS Quarterly*, *45*(3). https://doi.org/10.25300/MISQ/2021/16535

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

Trocin, C., Mikalef, P., Papamitsiou, Z., & Conboy, K. (2021). Responsible AI for Digital Health: a Synthesis and a Research Agenda. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10146-4

Van Eck, N. J., & Waltman, L. (2013). Vosviewer manual. *Leiden: Universiteit Leiden*, *1*(1), 1–53.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. *Science robotics*, *2*(6), eaan6080. https://doi.org/10.1126/scirobotics.aan6080

Waltman, L., van Eck, N. J., & Noyons, E. C. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*(4), 629–635. https://doi.org/10.1016/j.joi.2010.07.002

Wamba, S. F., Bawack, R. E., Guthrie, C., Queiroz, M. M., & Carillo, K. D. A. (2021). Are we preparing for a good AI society? A bibliometric review and research agenda. *Technological Forecasting and Social Change*, *164*, 120482. https://doi.org/10.1016/j.techfore.2020.120482

Wamba, S. F., & Queiroz, M. M. (2021). Responsible Artificial Intelligence as a Secret Ingredient for Digital Health: Bibliometric Analysis, Insights, and Research Directions. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10142-8

Zhang, Y., Wu, M., Tian, G. Y., Zhang, G., & Lu, J. (2021). Ethics and privacy of artificial intelligence: Understandings from bibliometrics. *Knowledge-Based Systems*, *222*. https://doi.org/10.1016/j.knosys.2021.106994