Intelética 2(3) (2025), 1-16
Revista de Inteligencia Artificial, Ética y Sociedad

INTELETICA

https://inteletica.iberamia.org/

Mapping the ethical landscape in AI: A bibliometric analysis with educational insights

Darío José Delgado-Quintero[1], Jheimer Julián Sepúlveda-López[2], Margarita María Pineda-Romero[3], Diego Julián Santos-Méndez[4]

[1] Universidad Nacional Abierta y a Distancia (UNAD), Colombia
[2] Universidad Nacional de Colombia, Colombia
[3] Universidad Nacional Abierta y a Distancia (UNAD), Colombia
[4] Universidad Nacional Abierta y a Distancia (UNAD), Colombia

[1] dario.delgado@unad.edu.co
[2] jjsepulvedal@unal.edu.co
[3] margarita.pineda@unad.edu.co
[4] diego.santos@unad.edu.co

"Artículo revisado y recomendada su publicación por el Dr. Jhonatan Camacho Navarro, profesional de innovación y tecnología ICPET ECOPETROL S.A, Colombia y por el Dr. Carlos Betancourt Correa, profesor titular de la Facultad de Ciencias e Ingeniería de la Universidad de Manizales, Colombia".

Abstract Artificial intelligence has become an increasingly used trend in different sectors; there are applications that affect how people work, study and relate. For this reason, it is increasingly necessary to identify and address the ethical implications that these types of tools bring with them. Based on the above, this document provides a bibliometric analysis that aims to comprehensively address the multidimensional aspects of ethics in Artificial Intelligence. For this, a search was carried out in the SCOPUS academic database in a period of time ranging from 2013 to 2023 in which thematic areas such as Engineering, Education, Law, Philosophy, Computer Science, Business and Sociology were addressed. The results of this process show that there has been exponential growth in scientific production in this field since 2020; Likewise, it is possible to identify nine clusters in which the documents in this area are grouped, among which Ethics and Principles, AI in Education and Explainability and Interpretability stand out. In practical terms, it is hoped that this document will allow interested persons addressing the ethical elements of artificial intelligence to identify relevant areas and locate their own research.

Resumen La inteligencia artificial se ha convertido en una tendencia cada vez más utilizada en diferentes sectores; existen aplicaciones que afectan a la forma de trabajar, estudiar y relacionarse de las personas. Por esta razón, es cada vez más necesario identificar y abordar las implicaciones éticas que este tipo de herramientas traen consigo. Con base en lo anterior, este documento ofrece un análisis bibliométrico que pretende abordar de manera integral los aspectos multidimensionales de la ética en la Inteligencia Artificial. Para ello, se realizó una búsqueda en la base de datos académica SCOPUS en un periodo de tiempo que va de 2013 a 2023 en la que se abordaron áreas temáticas como Ingeniería, Educación, Derecho, Filosofía, Informática, Negocios y Sociología. Los resultados de este proceso muestran que ha habido un crecimiento exponencial de la producción científica en este campo desde 2020; asimismo, es posible identificar nueve

ISSN: 3020-7444

clústeres en los que se agrupan los documentos de esta área, entre los que destacan Ética y Principios, IA en Educación y Explicabilidad e Interpretabilidad. En términos prácticos, se espera que este documento permita a las personas interesadas en abordar los elementos éticos de la inteligencia artificial identificar las áreas relevantes y localizar sus propias investigaciones.

Keywords: AI in Education, Artificial Intelligence, Bibliometric Analysis, Intelligent Tutoring Systems, Ethics and AI.
Palabras clave: IA en Educación, Inteligencia Artificial, Análisis Bibliométrico, Sistemas Tutores Inteligentes, Ética e IA.

1 Introduction

Currently, there is evidence of exponential growth in the use of artificial intelligence (AI) in various sectors, with the educational field being one of the most impacted. This emerging technology, with its vast potential, promises to revolutionize the way teaching and learning are conducted. However, the implementation of AI in education is not without significant challenges, especially regarding ethical and practical aspects.

AI offers the possibility of personalizing education, tailoring it to the individual needs of each student. However, its application can also present ethical risks and challenges. In this context, ethics in AI has become a crucial topic. The proper appropriation of ethical frameworks for the application of AI in higher education is essential to ensure that AI is used responsibly, fairly, and transparently, considering the particularities of educational processes. To understand the considerations stipulated in ethical frameworks, various ethical theories must be considered; one of the most relevant theories in this area is the ethics of technology [1], which focuses on reflecting on the nature and impact of technology on society and human life. Therefore, AI, being a technological development, must consider the ethical principles already established in other international ethical frameworks, such as other theories like the ethics of responsibility [2], which focuses on making ethical decisions based on the consequences of actions, and the ethics of justice [3], which focuses on fairness and the fair distribution of resources and benefits.

Based on its ethical and philosophical foundation, some of the most common principles for the application of artificial intelligence are documented [4]:

- Transparency: Decisions made by AI systems must be explainable and understandable to users and those affected by them.
- Responsibility: Developers, manufacturers, and users of AI systems must be accountable for their use and the effects they may have.
- Justice: AI systems must be designed and used fairly, without discrimination or unfair biases.
- Privacy: AI systems must protect the privacy and rights of individuals and respect data protection laws and regulations.
- Security: AI systems must be secure and protect people from physical or psychological harm.
- Confidentiality: Personal data of users and those affected by AI systems must be treated confidentially and protected from any unauthorized access.
- Sustainability: AI systems must be designed and used sustainably and respect the environment.

These general principles must be contextualized in their different fields of action. Based on the above, a search was carried out in the SCOPUS academic database in a period of time ranging from 2013 to 2023, addressing thematic areas such as Engineering, Education, Law, Philosophy, Computer Science, Business, and Sociology.

This document begins by making an approach to the method used for the bibliographic analysis process, subsequently an approach is made to the results obtained in different sections ranging from the analysis of the sources, the most influential authors on the subject, an analysis of keywords. and text. Finally, the clusters that emerge when reviewing the ethical aspects of AI are shown.

2 Method

[5] proposes systematic reviews as a type of scientific research that aims to objectively and systematically integrate the results of empirical studies on a specific research problem, in order to determine the "state of the art" in that field of research. study. When reviewing these topics, it is essential to define the search process and with it the source or sources that were consulted [6]; additionally, the other fundamental factor is the definition of the search equation.

To carry out the process of systematic literature search on ethical aspects in AI with Educational Insights, the three steps proposed by [7] were executed:

- Step 1 - Planning: the problem to be solved through the search was formulated and the search equation was established and the source of information where this process would be carried out was defined.
- Step 2 - Execution: The search equation was applied to the defined database and the quantity and specificity of the results were reviewed.
- Step 3 - Report: finally, the classification, bibliometric analysis and presentation of the results were carried out.

Based on the above, for this process, a series of thesauri were drawn upon to serve as a reference framework for this bibliometric research, aiming to comprehensively address the multi-dimensional aspects of ethics in Artificial Intelligence (AI). Included in these thesauri are terms like 'AI Ethics,' 'Artificial Intelligence Ethics,' 'Ethical Frameworks,' and 'Ethical Guidelines,' which focus on the ethical and moral principles involved in the design and implementation of AI. Additionally, specific areas such as 'Ethics in Education' and 'AI in Education' were considered, as they explore the impact of AI within the educational sector. The approach also incorporates terms that focus on attributes and principles like 'Transparency in AI,' 'Accountability in AI,' 'Fairness in AI,' 'Explainable AI' (also known as 'XAI'), 'Algorithmic Bias,' and 'Algorithmic Fairness.' These terms are crucial for understanding aspects like transparency, accountability, and fairness in AI systems.

Utilizing the selected thesauri, a search equation was formulated to identify academic articles that address these topics. The equation was applied to the SCOPUS academic database and encompasses a timespan ranging from 2013 to 2023. Thematic areas such as Engineering, Education, Law, Philosophy, Computer Science, Business, and Sociology were targeted. The developed equation is as follows, see equation 1.

(TITLE-ABS-KEY("AI Ethics" OR "Artificial Intelligence Ethics" OR "Ethical Frameworks" OR "Ethical Guidelines" OR "Ethical AI" OR "Ethics in AI" OR "Ethics in Education" OR "AI in Education" OR "Artificial Intelligence in Education" OR "Educational AI") OR TITLE-ABS-KEY("Transparency in AI" OR "Accountability in AI" OR "Fairness in AI" OR "Explainable AI" OR "XAI") OR TITLE-ABS-KEY("Algorithmic Bias" OR "Algorithmic Fairness")) AND PUBYEAR > 2012 AND PUBYEAR < 2024 AND (LIMIT-TO(DOCTYPE, "ar")) AND (LIMIT-TO(SUBJAREA, "ENGI") OR LIMIT-TO(SUBJAREA, "EDUC") OR LIMIT-TO(SUBJAREA, "LAW") OR LIMIT-TO(SUBJAREA, "PHIL") OR LIMIT-TO(SUBJAREA, "COMP") OR LIMIT-TO(SUBJAREA, "BUSI") OR LIMIT-TO(SUBJAREA, "SOCI")) (1)

The search equation serves as the backbone for this bibliometric analysis, allowing the ethical landscape and evolution of artificial intelligence applications to be comprehensively mapped.

The initial bibliometric data set, or Corpus, comprising 4462 documents, is subjected to Bradford Law analysis [8] to filter the bibliographic dataset, see TABLE I. For the purposes of this work, only the Core Zone and Zone 2 will be utilized. This approach serves to make the bibliometric analysis more manageable and efficient by reducing the Corpus from 4461 to 2991 documents, while minimizing bias. This reduction accounts for 72.5% of the original authors and includes 23.6% of the most representative sources.

TABLE I. Filter Sources by Bradford Law Zones

Filter by	Documents	Sources	Authors
All Sources	4461 – 100%	1660 – 100%	12386 – 100%
Core Sources	1474 – 33,04%	53 – 3,2%	4850 – 39,1%
Core + Zone 2 Sources	2991 – 67%	392 – 23,6%	8985 – 72,5%

Additionally, when further filtered to include only documents in English, a corpus is obtained consisting of 2948 documents, 384 sources, and 8905 authors. From a general perspective, the corpus is composed of relatively recent documents, with an average document age being recorded at 1.71 years, and significant growth having been observed since the year 2019, see FIGURE 1.

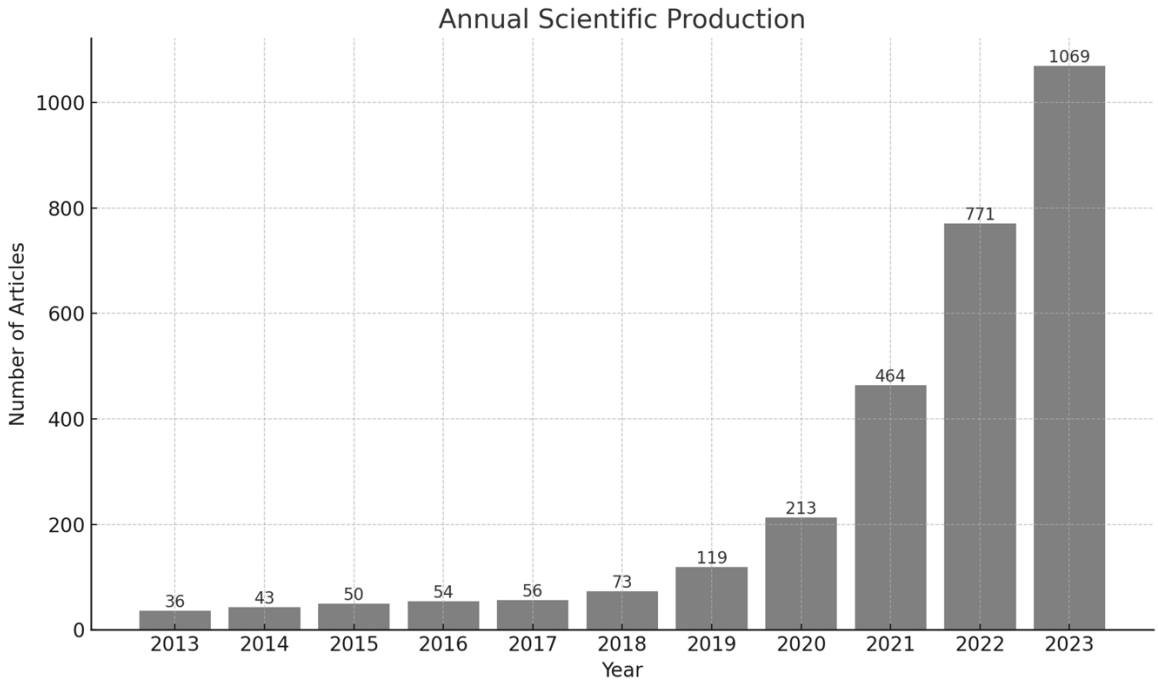

FIGURE 1. The Annual Growth Rate is Depicted at 40,3% for Annual Scientific Production

3 Sources

According to Bradford's Law, as depicted in FIGURE 2, a concentration of articles within a select group of journals that constitute the core zone (Zone 1) of the study is observed. 'IEEE Access' is identified as the most prolific source with 145 articles, closely followed by 'AI and Society' with 92 articles, and 'Applied Sciences (Switzerland)' with 78 articles. Significant contributions are also made by journals such as 'Sensors' and 'International Journal of Artificial Intelligence in Education,' which account for 54 and 48 articles, respectively. The core collection is further enriched by journals like 'Journal of Medical Ethics' and 'Science and Engineering Ethics,' which serve as primary reservoirs of high-quality literature in the field under investigation. A substantial portion of the articles in the corpus (approximately 33,7%) is found to be accounted for by these leading journals, emphasizing their pivotal role in shaping the ethical discourse around artificial intelligence.

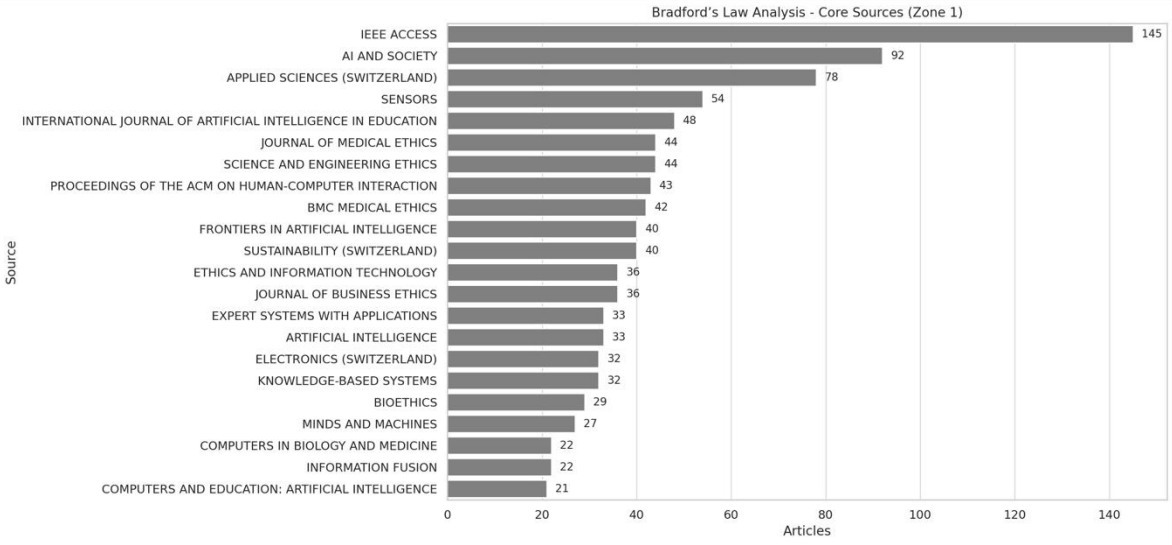

FIGURE 2. Core Sources by Bradford's Law, the corpus is found to be dominated by 22 most relevant sources. In descending order, the first 10 sources: IEEE Access, Ai And Society, Applied Sciences (Switzerland), Sensors, International Journal Of Artificial Intelligence In Education, Journal Of Medical Ethics, Science And Engineering Ethics, Bmc Medical Ethics, Sustainability (Switzerland),Frontiers In Artificial Intelligence.

4 Contributing authors

In the field of ethics in artificial intelligence, significant prominence has been observed among certain authors, such as LEE S, HOLZINGER A, and FLORIDI L, who have been featured across multiple categories, see FIGURE 3. This indicates their substantial influence and preeminence. For instance, LEE S is not only one of the most relevant authors in terms of the number of articles and their fractional contribution, but also possesses high H and G indices in the local impact analysis, see TABLE II. Additionally, a fascinating divergence is noted between the most locally cited authors, like HOWE B, JAGADISH H, STOYANOVICH J, and those with high local impact and global relevance. This divergence might suggest that some authors are highly valued in specific discussions within the field, while others hold broader, more general influence. The impact and relevance indicators, including H, G, and M indices, along with total citations (TC) and the number of publications (NP), provide a detailed view of each author's impact. Notably, HOLZINGER A and FLORIDI L exhibit high h and g indices, as well as a high number of total citations, suggesting that their works are quantitatively significant and widely recognized and cited within the academic community. The publication start date (PY_start) in the local impact analysis offers insights into the trajectory and currency of the authors in the field. Authors like SAMEK W and ALI S, with more recent publication starts (2021 and 2022, respectively) and high indices, are indicative of emerging figures and current trends in AI ethics.

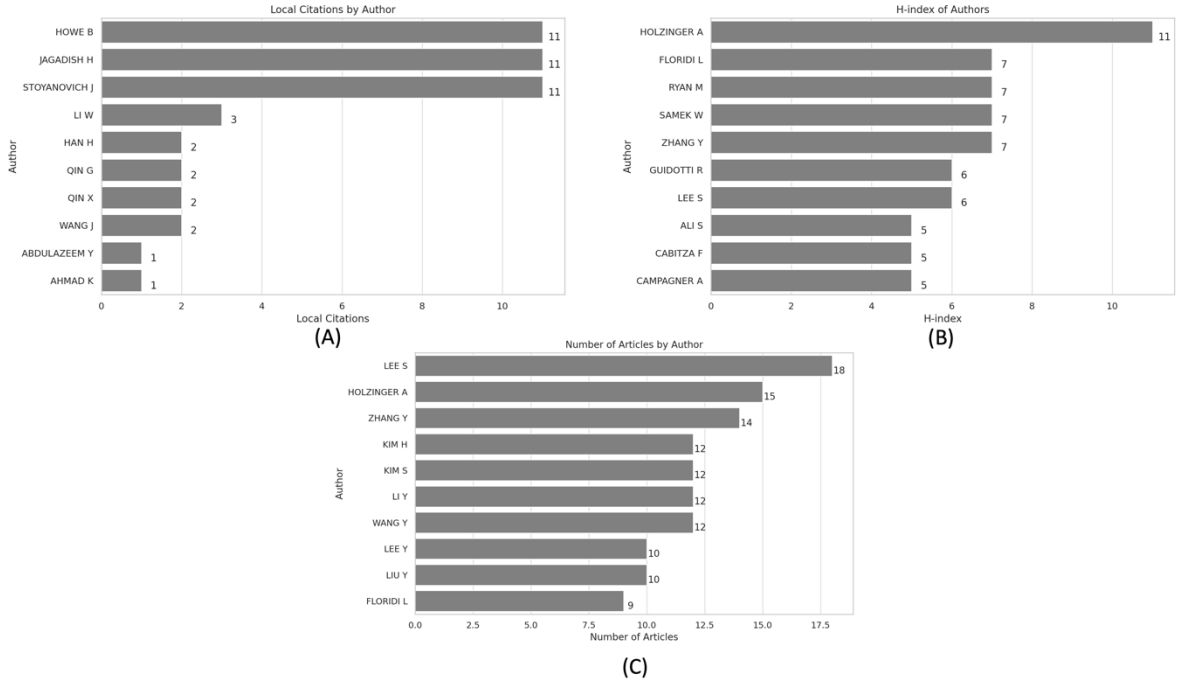

FIGURE 3. Featured in Image A is the analysis of 'Most Local Cited Authors, detailing the count of local citations per author. Image B presents the Authors' Local Impact analysis, indicating each author's h-index. Image C illustrates the Most Relevant Authors analysis, showing both the total number of articles by each author and the fractional contribution to the field. These visualizations collectively underscore the varied dimensions of authorial impact within the study of ethics in artificial intelligence.

The integration of these data fosters a richer and more nuanced understanding of the field. While some authors are key in terms of local citation influence, others make significant contributions through frequent and high-quality publications. This amalgamation of metrics assists in identifying both established and emerging figures in the ethics of AI.

TABLE II. Authors' Local Impact Indicators in Artificial Intelligence Ethics Research

Element	h_index	g_index	m_index	TC	NP	PY_start
HOLZINGER A	11	15	1.833	666	15	2018
FLORIDI L	7	9	0.875	1099	9	2016
RYAN M	7	8	1.750	276	8	2020
SAMEK W	7	8	2.333	641	8	2021
ZHANG Y	7	10	2.333	109	14	2021
GUIDOTTI R	6	9	1.200	368	9	2019
LEE S	6	14	1.200	218	18	2019

ALI S	5	8	2.500	86	8	2022
CABITZA F	5	7	1.250	109	7	2020
CAMPAGNER A	5	6	1.250	109	6	2020

5 Keyword and text analysis

In the examination of ethical considerations in artificial intelligence, See FIGURE 4, network centrality measures, such as degree centrality and link strength, have been utilized. Degree centrality, counting the number of direct connections of a keyword, and link strength, reflecting the total strength of a keyword's connections, were employed to identify the most significant keywords within the network.

The centralities of each keyword were calculated, revealing the following influential keywords in the context of ethics in artificial intelligence:

- *Explainable AI* was found to possess the highest link strength in the network, indicating a robust association with numerous other keywords.
- *Machine Learning* also exhibited high link strength, underscoring its significance in the realm of AI ethics.
- *Deep Learning* emerged as another considerable keyword, with a high weighted degree.
- *Ethics*, being directly relevant to the theme, manifested many connections within the network.
- *XAI (Explainable Artificial Intelligence)*, akin to "Explainable AI", emphasized the importance of explicability in AI.

These findings suggest that explicability and understanding of AI processes, particularly in deep learning and machine learning, are central themes in the discourse on ethics in artificial intelligence.

The aim of this analysis is to identify clusters of keywords that are closely related to each other, representing subthemes or focus areas within the broader field of ethics in artificial intelligence. The data from the Co-occurrence analysis, performed with VOSviewer software, which includes cluster information for each keyword, has been utilized. Keywords have been grouped by cluster, followed by an analysis of the most prominent clusters to identify the themes they represent. The grouping of keywords has been initiated, leading to the exploration of the largest clusters.

The largest clusters in the keyword co-occurrence network, along with some of their representative keywords, are identified as follows, see TABLE III:

In cluster I, ethics and principles, the sample of articles selected address fundamental aspects of ethics in artificial intelligence (AI), reflecting concerns about ethical dilemmas, privacy, and governance frameworks. They focus on creating consensus on ethical principles to guide the development and adoption of AI, evaluating existing ethical guidelines to identify omissions and overlaps [2][3]. Additionally, they explore ethical responsibility in the use of algorithms and data, pointing out the discontinuities between current practices and traditional ethical regulations [4][5]. The transition from ethical principles to concrete practices is discussed, highlighting the need for tools and methods that facilitate this application in the AI development cycle [6][7][8]. Specific ethical considerations in the use of social media data and ethical governance in robotics and AI systems are also addressed to foster public trust, among others. Finally, the issue of meaningful human control over autonomous systems is examined, proposing approaches to ensure human moral responsibility in both military and non-military operations [9]. Collectively, this cluster emphasizes the importance of a comprehensive and pragmatic ethical approach to the design, development, and application of AI technologies, underscoring the need for continuous dialogue between developers, regulators, and society to address emerging ethical challenges.

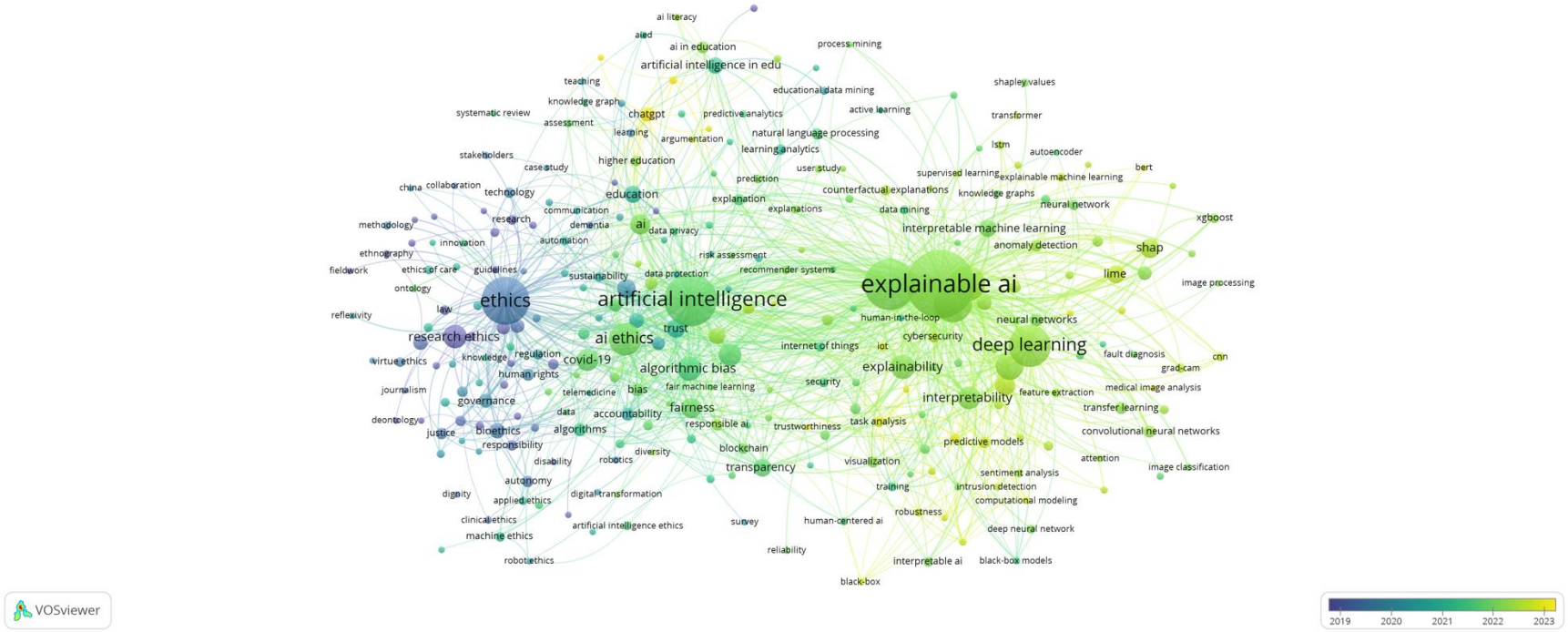

FIGURE 4. The co-occurrence of authors' keywords. presents the prevalence of 541 keywords appearing in our search results using SCOPUS. The thickness of lines is an indication of the strength of the relationship between keywords relative to the others. The strength of these relationships was determined by the frequency with which they appeared together in published articles. Their inclusion into specific thematic groups was based on their clustering with a certain constellation of terms. The position of a keyword within this constellation represents how interrelated and frequent its co-occurrence was with other terms.

TABLE III. Ethical themes in Artificial Intelligence Research

Cluster	Description	Principal Key Words	References
Ethics and Principles	This cluster is centered around the foundational ethical considerations and principles that guide the application of artificial intelligence. Ethical dilemmas, privacy concerns, and governance frameworks are prominently featured.	Ethics, Research Ethics, COVID-19, Privacy, Bioethics, Informed Consent, Governance, Ethical Guidelines, Ethical Framework, social media	[9], [10], [11], [12], [13], [14], [15], [16]
Explainable and transparency in AI	The importance of transparency and understandability in AI systems is highlighted in this cluster. Techniques and methodologies to make AI decisions explainable are explored extensively.	Explainable AI, Machine Learning, Deep Learning, Explainable Artificial Intelligence, XAI, Explainable AI (XAI), SHAP, LIME, Interpretable Machine Learning	[17], [18], [19], [20], [21], [22], [23], [24]
AI in Education	The application of AI within educational settings is examined, focusing on its potential to enhance learning outcomes and make educational resources more accessible. The need for AI literacy and trustworthy AI is also emphasized.	Artificial Intelligence, AI, Education, Artificial Intelligence in Education, Big Data, ChatGPT, AI in Education, Natural Language Processing, Trustworthy AI, Explanation	[25], [26], [27], [28], [29], [30], [31], [32]
Bias and Fairness in AI	Concerns regarding bias and fairness within AI algorithms are addressed. The cluster discusses efforts to mitigate algorithmic bias and promote fairness, equity, and ethical AI development.	AI Ethics, Algorithmic Bias, Algorithmic Fairness, Fairness, Artificial Intelligence (AI), Algorithms, Ethical AI, Bias, Reinforcement Learning, Responsible AI	[11], [16], [26], [33], [34], [35], [36], [37]
Computer Vision and Predictive Models	Innovations and challenges in computer vision and predictive modeling are explored, highlighting their applications and the ethical considerations they raise.	Computer Vision, Visualization, Predictive Models, Task Analysis, Deep Neural Networks, Visual Analytics, Data Models, Interpretable AI, Training, Sentiment Analysis	[38], [39], [40], [41], [42], [43], [44]
Neural Networks and Security	The development and deployment of neural networks in enhancing security measures are discussed, along with the ethical implications of automation and data science in security contexts.	Neural Networks, Feature Selection, Security, Data Science, Automation, IoT, Decision Trees, Intrusion Detection, Mental Health, Risk Assessment	[45], [46], [47], [48], [49], [50], [51], [52]
Explainability and	The necessity for AI systems to be explainable and interpretable is emphasized,	Explainability, Interpretability, Transparency, Accountability, Deep, Neural	[53], [54], [55], [56], [57],

Interpretability	exploring methods to achieve transparency and accountability in AI applications.	Network, Autonomous Vehicles, Black-box, Object Detection, Reliability	[58], [59], [60]
Human-AI Interaction	The dynamics of human interaction with AI systems are investigated, focusing on trust, design principles for human-AI interfaces, and the impact of AI on user experiences and usability.	Trust, Counterfactual Explanations, Explanations, Human-AI Interaction, User Study, Design, Human-Computer Interaction, Virtual Reality, Active Learning, Usability	[16], [20], [60], [61], [62], [63], [64], [65]
Internet of Things and Affective Computing	The integration of AI in the Internet of Things and its application in affective computing are reviewed, highlighting the technological advancements and ethical considerations of emotion recognition and personal data usage.	Internet of Things, Feature Importance, Clustering, Emotion Recognition, Shapley Additive Explanations, Affective Computing, Counterfactual Explanation	[66], [67], [68], [69], [70], [71], [72], [73]

In cluster II, Explainable and transparency in AI, the articles in this cluster focus on the importance of transparency and comprehensibility in AI systems, exploring techniques and methodologies to make AI decisions explainable [10-17]. They address the challenge of the "black box" in AI systems, proposing solutions to improve trust and transparency [11, 12]. The need for explainability in critical sectors such as healthcare is highlighted [12], and the impact of explainability on users' perception of trust and acceptance of AI is examined [13]. Additionally, the translation of ethical principles into practices through tools and methods that promote explainable AI is discussed [14]. The articles suggest a multidisciplinary approach to address the demands of different stakeholders in explainable AI, underscoring the need for interdisciplinary research in this field [17].

In cluster III, AI in Education, the sample of articles analyzed addresses the implementation of artificial intelligence in education, highlighting not only its potential benefits but also the related ethical concerns. Topics such as the need for equity in access to AI-powered education ([18]), the importance of avoiding algorithmic biases that can influence education ([19]), and the promotion of innovative and accessible educational environments through educational cobots and smart classrooms ([20]) are discussed. Additionally, the role of MOOCs in democratizing education is examined ([21]) and the impact of tools such as ChatGPT on academic assessment is analyzed ([22]). These studies emphasize the importance of developing and using AI technologies in a way that respects fundamental ethical principles such as transparency, fairness, and respect for student privacy.

In cluster IV, Bias and Fairness in AI, this sample of articles addresses concerns about bias and fairness within artificial intelligence algorithms. These efforts focus on mitigating algorithmic bias and promoting fairness, equity, and ethical development of AI. Topics covered range from the ethics of algorithms and meaningful human control over autonomous systems, to empirical studies on gender-based discrimination in STEM job advertisements, perceptions of automated decision-making, and the participatory design of algorithmic policies for governing with fairness. The need to align algorithmic fairness with legal frameworks such as the EU's right to non-discrimination is also discussed, highlighting the challenges of automating fairness in complex legal contexts and the urgency of using ethical principles responsibly in AI to prevent their manipulation. Articles [4], [9], [19], [26], [27], [28], [29], and [30] reflect a broad spectrum of concerns and proposed solutions for addressing bias and fairness in AI from ethical, technical, and legal perspectives.

In cluster V, Computer Vision and Predictive Models, the selected sample of articles, innovations and challenges in computer vision and predictive modeling are addressed by [31], [32], [34], and [35], with their applications and ethical considerations being highlighted. Innovations for visualizing discriminative image regions in Convolutional Neural Networks (CNNs) are examined in article [31], enhancing transparency in computer vision. A visual analysis system for comparing predictive models in clinical data is presented in [32], facilitating evidence-based medical decision-making. The use of twin systems to explain deep learning models through examples is explored in [34], contributing to explainable artificial intelligence (XAI). Lastly, deep learning models and XAI methods for sentiment analysis in food delivery service reviews are reviewed in [35], emphasizing the importance of interpretability. These works reflect an effort to make AI more transparent and ethical, tackling everything from the localization of discriminative features to the comparison and evaluation of predictive models in varied contexts.

In cluster VI, Neural Networks and Security, the collection of analyzed articles is highlighted for its critical intersection among neural networks, security, and ethics in artificial intelligence from a passive voice perspective. Topics such as explainable artificial intelligence (XAI) in cybersecurity, ethics, and privacy in AI and big data, applications of XAI in cyber security, and ethical and legal challenges in AI-driven cybersecurity are addressed. The opacity of AI systems and its negative impact on trust and security is explored, emphasizing the importance of transparency and explainability. The concept of Responsible Research and Innovation (RRI) is discussed as a framework to address these ethical challenges, ensuring the technology's benefits outweigh its drawbacks. Furthermore, the application of XAI across different cybersecurity domains, like intrusion and malware detection, is examined, highlighting the need for more interpretable AI models that can foster trust and be effectively managed by users. In summary, these works point to the need for a balance between advancing security technology through neural networks and addressing the ethical implications arising from their deployment. References [38], [39], [40], [41], [42], [43], [44], [45] provide a foundation for future research in this critical area of AI ethics.

In cluster VII, Explainability and Interpretability, in the selected sample of articles, [46] to [53] are addressed from various aspects of explainability and interpretability in artificial intelligence (AI), emphasizing the necessity for AI systems to be both explainable and interpretable to ensure transparency and accountability in AI applications. It is highlighted that transparent and accountable AI systems are crucial for decision support, object detection, and recognition in autonomous driving, with ethical and responsibility implications discussed for autonomous vehicles. Methods and applications for explaining deep neural networks are reviewed, proposing explainable deep learning architectures and evaluating the impact of post-hoc explanations on user perception and trust in AI systems. These studies underscore the ethical imperative to develop AI technologies that are not only advanced in performance but also accessible, understandable, and fair to users and society at large.

In cluster VIII, Human-AI Interaction, within the field of AI ethics is addressed through the investigation of the dynamics of human interaction with AI systems, focusing on trust, design principles for human-AI interfaces, and the impact of AI on user experiences and usability. It is covered by key articles that fundamental topics such as significant human control over autonomous systems [9], the importance of explainability and causability in the perception of trust and acceptance of AI [13], and the design of explainable interactions through virtual agents [56] are focused on. The necessity for autonomous systems to respond to human moral reasons and to allow the tracing of their operations' outcomes back to humans is highlighted [9], while explainability and causability are identified as crucial factors in fostering users' trust and understanding of AI algorithms [13, 53, 54, 55]. Furthermore, the enhancement of trust in explainable AI systems by virtual agents and the relevance of human-centered design for explanations in clinical decision support systems are examined [57, 58]. These studies underscore the importance of integrating ethical and humanistic considerations into AI development to ensure more transparent, comprehensible, and trustworthy systems.

Cluster IX, Internet of Things and Affective Computing, the integration of Artificial Intelligence (AI) into the Internet of Things (IoT) and its application in affective computing are highlighted for their technological advancements and ethical considerations concerning emotion recognition and personal data usage. It is discussed how articles range from explainable AI in Industry 4.0 to digital vulnerabilities, intrusion detection systems, AI differentiation of facial expressions, and the GDPR framework for IoT transparency. Emphasis is placed on the significance of explainable AI techniques for understanding complex model decisions, addressing cybersecurity, privacy, and ethics in data use for emotional recognition and experience personalization. Future research is directed towards responsible and human-centric AI, focusing on explainability and ethics in critical systems [66], [67], [68], [69], [70], [71], [72], [73].

6 Conclusions

This bibliometric analysis examined the ethical landscape and development of AI applications using the SCOPUS academic database and focusing on publications between 2013 and 2023. The analysis identified several key findings. The field of ethics in AI has seen significant growth in recent years, with a 40.3% annual increase in scientific output seen since 2019. Primary sources for this research include journals such as “IEEE Access” and “AI and Society". Some authors, such as Lee S, Holzinger A and Floridi L, have been particularly influential in this field and have made frequent and high-quality contributions. There is also a distinction between authors highly cited in specific debates and those with broader influence.

The authors' keyword co-occurrence revealed nine thematic clusters representing key areas of focus in AI ethics research; findings that offer valuable insights into the current state of AI ethics research. The emphasis on explainability, fairness, transparency and human-centered design principles highlights a growing recognition of the importance of ethical considerations in the development of AI. The analysis also identifies emerging areas of interest, such as the ethics of AI in education and the use of AI in security contexts. Overall, this bibliometric analysis performed provides a comprehensive overview of the key themes and trends in AI ethics research. As the field continues to evolve, addressing these ethical challenges will be crucial to ensure the responsible and beneficial development of AI technologies.

Finally, it is important to indicate that AI is experiencing exponential growth in various sectors, including education. However, its implementation presents significant challenges, especially in ethical and practical aspects. It is possible to identify common principles for the application of AI, such as transparency, responsibility, fairness, privacy, security, confidentiality and sustainability. This process then allows two final elements to be highlighted:

- Future Challenges:
    - Practical Application: The translation of ethical principles into concrete practices remains an area of active research.
    - Impact Assessment: A deeper assessment of the real impact of ethical policies on the implementation of AI systems is needed.

- Recommendations for the Scientific Community:
    - Promote interdisciplinary collaboration between experts in ethics, computer science and education.
    - Develop specific ethical frameworks for the application of AI in educational contexts.

References

[1] J. Núñez, "Ética, Ciencia y Tecnología: Sobre la función social de la tecnología," *Llull: Revista de la Sociedad Española de Historia de las Ciencias y de las Técnicas*, vol. 25, pp. 459–484, 2002.
[2] A. Mestre, "La ética de la responsabilidad según Robert Spaemann," *Universitas (Stuttg)*, vol. 10, p. 233259, 2008.
[3] A. Fascioli, "Ética del cuidado y ética de la justicia en la teoría moral de Carol Gilligan.," *Revista Actio*, vol. 12, pp. 41–57, 2010.
[4] OCDE, "OECD AI Principles overview," https://oecd.ai/en/ai-principles. Accessed: Apr. 04, 2024. [Online]. Available: https://oecd.ai/en/ai-principles
[5] J. Sanchez, "Cómo realizar una revisión sistemática y un meta-análisis," *Aula abierta*, vol. 38, pp. 53–64, 2010.
[6] E. Galvis and M. Sanchez, "Revisión Sistemática de literatura sobre procesos de gestión de conocimiento," *Revista Gerencia Tecnológica Informática*, vol. 13, pp. 45–67, 2014.
[7] J. D. Velásquez, "Una Guía Corta para Escribir Revisiones Sistemáticas de Literatura Parte 3," *Dyna (Medellin)*, vol. 82, no. 189, pp. 9–12, Feb. 2015, doi: 10.15446/dyna.v82n189.48931.
[8] B. C. Peritz, "A bradford distribution for bibliometrics," *Scientometrics*, vol. 18, no. 5–6, pp. 323–329, May 1990, doi: 10.1007/BF02020148.
[9] L. Floridi *et al.*, "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds Mach (Dordr)*, vol. 28, no. 4, pp. 689–707, Dec. 2018, doi: 10.1007/s11023-018-9482-5.
[10] T. Hagendorff, "The Ethics of AI Ethics: An Evaluation of Guidelines," *Minds Mach (Dordr)*, vol. 30, no. 1, pp. 99–120, Mar. 2020, doi: 10.1007/s11023-020-09517-8.
[11] M. Ananny, "Toward an Ethics of Algorithms," *Sci Technol Human Values*, vol. 41, no. 1, pp. 93–117, Jan. 2016, doi: 10.1177/0162243915606523.
[12] J. Metcalf and K. Crawford, "Where are human subjects in Big Data research? The emerging ethics divide," *Big Data Soc*, vol. 3, no. 1, p. 205395171665021, Jun. 2016, doi: 10.1177/2053951716650211.
[13] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," *Sci Eng Ethics*, vol. 26, no. 4, pp. 2141–2168, Aug. 2020, doi: 10.1007/s11948-019-00165-5.
[14] M. L. Williams, P. Burnap, and L. Sloan, "Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation," *Sociology*, vol. 51, no. 6, pp. 1149–1168, Dec. 2017, doi: 10.1177/0038038517708140.
[15] A. F. T. Winfield and M. Jirotka, "Ethical governance is essential to building trust in robotics and artificial intelligence systems," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2133, p. 20180085, Nov. 2018, doi: 10.1098/rsta.2018.0085.
[16] F. Santoni de Sio and J. van den Hoven, "Meaningful Human Control over Autonomous Systems: A Philosophical Account," *Front Robot AI*, vol. 5, Feb. 2018, doi: 10.3389/frobt.2018.00015.
[17] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.
[18] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[19] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Trans Neural Netw Learn Syst*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314.
[20] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI," *Int J Hum Comput Stud*, vol. 146, p. 102551, Feb. 2021, doi: 10.1016/j.ijhcs.2020.102551.
[21] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," *Sci Eng Ethics*, vol. 26, no. 4, pp. 2141–2168, Aug. 2020, doi: 10.1007/s11948-019-00165-5.
[22] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," *Information Fusion*, vol. 77, pp. 29–52, Jan. 2022, doi: 10.1016/j.inffus.2021.07.016.
[23] H. Hagras, "Toward Human-Understandable, Explainable AI," *Computer (Long Beach Calif)*, vol. 51, no. 9, pp. 28–36, Sep. 2018, doi: 10.1109/MC.2018.3620965.
[24] M. Langer *et al.*, "What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research," *Artif Intell*, vol. 296, p. 103473, Jul. 2021, doi: 10.1016/j.artint.2021.103473.
[25] L. Chen, P. Chen, and Z. Lin, "Artificial Intelligence in Education: A Review," *IEEE Access*, vol. 8, pp. 75264–75278, 2020, doi: 10.1109/ACCESS.2020.2988510.
[26] A. Lambrecht and C. Tucker, "Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads," *Manage Sci*, vol. 65, no. 7, pp. 2966–2981, Jul. 2019, doi: 10.1287/mnsc.2018.3093.
[27] M. J. Timms, "Letting Artificial Intelligence in Education Out of the Box: Educational Cobots and Smart Classrooms," *Int J Artif Intell Educ*, vol. 26, no. 2, pp. 701–712, Jun. 2016, doi: 10.1007/s40593-016-0095-y.
[28] J. Kay, P. Reimann, E. Diebold, and B. Kummerfeld, "MOOCs: So Many Learners, So Much Potential ...," *IEEE Intell Syst*, vol. 28, no. 3, pp. 70–77, May 2013, doi: 10.1109/MIS.2013.66.
[29] "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?," *Journal of Applied Learning & Teaching*, vol. 6, no. 1, Jan. 2023, doi: 10.37074/jalt.2023.6.1.9.
[30] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Comput Human Behav*, vol. 73, pp. 247–256, Aug. 2017, doi: 10.1016/j.chb.2017.01.047.
[31] I. Roll and R. Wylie, "Evolution and Revolution in Artificial Intelligence in Education," *Int J Artif Intell Educ*, vol. 26, no. 2, pp. 582–599, Jun. 2016, doi: 10.1007/s40593-016-0110-3.
[32] K. Mageira, D. Pittou, A. Papasalouros, K. Kotis, P. Zangogianni, and A. Daradoumis, "Educational AI Chatbots for Content and Language Integrated Learning," *Applied Sciences*, vol. 12, no. 7, p. 3239, Mar. 2022, doi: 10.3390/app12073239.
[33] T. Araujo, N. Helberger, S. Kruikemeier, and C. H. de Vreese, "In AI we trust? Perceptions about automated decision-making by artificial intelligence," *AI Soc*, vol. 35, no. 3, pp. 611–623, Sep. 2020, doi: 10.1007/s00146-019-00931-w.
[34] M. K. Lee *et al.*, "WeBuildAI: Participatory Framework for Algorithmic Governance," *Proc ACM Hum Comput Interact*, vol. 3, no. CSCW, pp. 1–35, Nov. 2019, doi: 10.1145/3359283.
[35] D. Pessach and E. Shmueli, "A Review on Fairness in Machine Learning," *ACM Comput Surv*, vol. 55, no. 3, pp. 1–44, Mar. 2023, doi: 10.1145/3494672.
[36] S. Wachter, B. Mittelstadt, and C. Russell, "Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI," *Computer Law & Security Review*, vol. 41, p. 105567, Jul. 2021, doi: 10.1016/j.clsr.2021.105567.
[37] A. Rességuier and R. Rodrigues, "AI ethics should not remain toothless! A call to bring back the teeth of ethics," *Big Data Soc*, vol. 7, no. 2, p. 205395172094254, Jul. 2020, doi: 10.1177/2053951720942541.
[38] K. R. Mopuri, U. Garg, and R. Venkatesh Babu, "CNN Fixations: An Unraveling Approach to Visualize the Discriminative Image Regions," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2116–2125, May 2019, doi: 10.1109/TIP.2018.2881920.
[39] Y. Li, T. Fujiwara, Y. K. Choi, K. K. Kim, and K.-L. Ma, "A visual analytics system for multi-model comparison on clinical data predictions," *Visual Informatics*, vol. 4, no. 2, pp. 122–131, Jun. 2020, doi: 10.1016/j.visinf.2020.04.005.
[40] T. Spinner, U. Schlegel, H. Schafer, and M. El-Assady, "explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning," *IEEE Trans Vis Comput Graph*, pp. 1–1, 2019, doi: 10.1109/TVCG.2019.2934629.

[41] E. M. Kenny and M. T. Keane, "Explaining Deep Learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in XAI," *Knowl Based Syst*, vol. 233, p. 107530, Dec. 2021, doi: 10.1016/j.knosys.2021.107530.
[42] A. Adak, B. Pradhan, and N. Shukla, "Sentiment Analysis of Customer Reviews of Food Delivery Services Using Deep Learning and Explainable Artificial Intelligence: Systematic Review," *Foods*, vol. 11, no. 10, p. 1500, May 2022, doi: 10.3390/foods11101500.
[43] H. Strobelt, S. Gehrmann, M. Behrisch, A. Perer, H. Pfister, and A. M. Rush, "Seq2seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models," *IEEE Trans Vis Comput Graph*, vol. 25, no. 1, pp. 353–363, Jan. 2019, doi: 10.1109/TVCG.2018.2865044.
[44] M. Carabantes, "Black-box artificial intelligence: an epistemological and critical analysis," *AI Soc*, vol. 35, no. 2, pp. 309–317, Jun. 2020, doi: 10.1007/s00146-019-00888-w.
[45] B. C. Stahl and D. Wright, "Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation," *IEEE Secur Priv*, vol. 16, no. 3, pp. 26–33, May 2018, doi: 10.1109/MSP.2018.2701164.
[46] N. Capuano, G. Fenza, V. Loia, and C. Stanzione, "Explainable Artificial Intelligence in CyberSecurity: A Survey," *IEEE Access*, vol. 10, pp. 93575–93600, 2022, doi: 10.1109/ACCESS.2022.3204171.
[47] Z. Zhang, H. Al Hamadi, E. Damiani, C. Y. Yeun, and F. Taher, "Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research," *IEEE Access*, vol. 10, pp. 93104–93139, 2022, doi: 10.1109/ACCESS.2022.3204051.
[48] K. Zarina I, B. Ildar R, and S. Elina L, "Artificial Intelligence and Problems of Ensuring Cyber Security," *nternational Journal of Cyber Criminology*, vol. 13, no. 2, 2019, doi: 10.5281/zenodo.3709267.
[49] M. Zolanvari, Z. Yang, K. Khan, R. Jain, and N. Meskin, "TRUST XAI: Model-Agnostic Explanations for AI With a Case Study on IIoT Security," *IEEE Internet Things J*, vol. 10, no. 4, pp. 2967–2978, Feb. 2023, doi: 10.1109/JIOT.2021.3122019.
[50] A. M. Oprescu *et al.*, "Towards a data collection methodology for Responsible Artificial Intelligence in health: A prospective and qualitative study in pregnancy," *Information Fusion*, vol. 83–84, pp. 53–78, Jul. 2022, doi: 10.1016/j.inffus.2022.03.011.
[51] A. Pnevmatikakis, S. Kanavos, G. Matikas, K. Kostopoulou, A. Cesario, and S. Kyriazakos, "Risk Assessment for Personalized Health Insurance Based on Real-World Data," *Risks*, vol. 9, no. 3, p. 46, Mar. 2021, doi: 10.3390/risks9030046.
[52] M. Sarhan, S. Layeghy, and M. Portmann, "Evaluating Standard Feature Sets Towards Increased Generalisability and Explainability of ML-Based Network Intrusion Detection," *Big Data Research*, vol. 30, p. 100359, Nov. 2022, doi: 10.1016/j.bdr.2022.100359.
[53] S. R. Hong, J. Hullman, and E. Bertini, "Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs," *Proc ACM Hum Comput Interact*, vol. 4, no. CSCW1, pp. 1–26, May 2020, doi: 10.1145/3392878.
[54] S. Larsson and F. Heintz, "Transparency in artificial intelligence," *Internet Policy Review*, vol. 9, no. 2, May 2020, doi: 10.14763/2020.2.1469.
[55] B. Kim, J. Park, and J. Suh, "Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information," *Decis Support Syst*, vol. 134, p. 113302, Jul. 2020, doi: 10.1016/j.dss.2020.113302.
[56] Y. Li *et al.*, "A Deep Learning-Based Hybrid Framework for Object Detection and Recognition in Autonomous Driving," *IEEE Access*, vol. 8, pp. 194228–194239, 2020, doi: 10.1109/ACCESS.2020.3033289.
[57] H.-Y. Liu, "Irresponsibilities, inequalities and injustice for autonomous vehicles," *Ethics Inf Technol*, vol. 19, no. 3, pp. 193–207, Sep. 2017, doi: 10.1007/s10676-017-9436-2.
[58] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Muller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, Mar. 2021, doi: 10.1109/JPROC.2021.3060483.
[59] P. Angelov and E. Soares, "Towards explainable deep neural networks (xDNN)," *Neural Networks*, vol. 130, pp. 185–194, Oct. 2020, doi: 10.1016/j.neunet.2020.07.010.
[60] E. M. Kenny, C. Ford, M. Quinn, and M. T. Keane, "Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies," *Artif Intell*, vol. 294, p. 103459, May 2021, doi: 10.1016/j.artint.2021.103459.
[61] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Min Knowl Discov*, Apr. 2022, doi: 10.1007/s10618-022-00831-6.
[62] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, "Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications," *Information Fusion*, vol. 81, pp. 59–83, May 2022, doi: 10.1016/j.inffus.2021.11.003.

[63] K. Weitz, D. Schiller, R. Schlagowski, T. Huber, and E. André, "'Let me explain!': exploring the potential of virtual agents in explainable AI interaction design," *Journal on Multimodal User Interfaces*, vol. 15, no. 2, pp. 87–98, Jun. 2021, doi: 10.1007/s12193-020-00332-0.
[64] T. A. J. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, and K. van den Bosch, "Human-centered XAI: Developing design patterns for explanations of clinical decision support systems," *Int J Hum Comput Stud*, vol. 154, p. 102684, Oct. 2021, doi: 10.1016/j.ijhcs.2021.102684.
[65] J. Dieber and S. Kirrane, "A novel model usability evaluation framework (MUsE) for explainable artificial intelligence," *Information Fusion*, vol. 81, pp. 143–153, May 2022, doi: 10.1016/j.inffus.2021.11.017.
[66] I. Ahmed, G. Jeon, and F. Piccialli, "From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where," *IEEE Trans Industr Inform*, vol. 18, no. 8, pp. 5031–5042, Aug. 2022, doi: 10.1109/TII.2022.3146552.
[67] S. Ransbotham, R. G. Fichman, R. Gopal, and A. Gupta, "Special Section Introduction—Ubiquitous IT and Digital Vulnerabilities," *Information Systems Research*, vol. 27, no. 4, pp. 834–847, Dec. 2016, doi: 10.1287/isre.2016.0683.
[68] Z. A. El Houda, B. Brik, and L. Khoukhi, "'Why Should I Trust Your IDS?': An Explainable Deep Learning Framework for Intrusion Detection Systems in Internet of Things Networks," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1164–1176, 2022, doi: 10.1109/OJCOMS.2022.3188750.
[69] K. Weitz, T. Hassan, U. Schmid, and J.-U. Garbas, "Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods," *tm - Technisches Messen*, vol. 86, no. 7–8, pp. 404–412, Jul. 2019, doi: 10.1515/teme-2019-0024.
[70] S. Wachter, "The GDPR and the Internet of Things: a three-step transparency model," *Law Innov Technol*, vol. 10, no. 2, pp. 266–294, Jul. 2018, doi: 10.1080/17579961.2018.1527479.
[71] S. Arisdakessian, O. A. Wahab, A. Mourad, H. Otrok, and M. Guizani, "A Survey on IoT Intrusion Detection: Federated Learning, Game Theory, Social Psychology, and Explainable AI as Future Directions," *IEEE Internet Things J*, vol. 10, no. 5, pp. 4059–4092, Mar. 2023, doi: 10.1109/JIOT.2022.3203249.
[72] J. Białek, W. Bujalski, K. Wojdan, M. Guzek, and T. Kurek, "Dataset level explanation of heat demand forecasting ANN with SHAP," *Energy*, vol. 261, p. 125075, Dec. 2022, doi: 10.1016/j.energy.2022.125075.
[73] T. Freiesleben, "The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples," *Minds Mach (Dordr)*, vol. 32, no. 1, pp. 77–109, Mar. 2022, doi: 10.1007/s11023-021-09580-9.