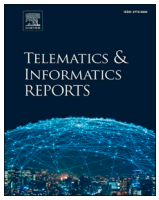
Telematics and Informatics Reports 18 (2025) 100205

Contents lists available at ScienceDirect

Telematics and Informatics Reports

journal homepage: www.elsevier.com/locate/teler



Ethical considerations in AI-based user profiling for knowledge management: A critical review

Daniel Kogi Njiru *, David Muchangi Mugo, Faith Mueni Musyoka

Department of Computing and Information Technology, University of Embu, Embu, Kenya

ARTICLE INFO

Keywords:
Artificial intelligence
AI
Ethics
User profiling
Knowledge management systems
Algorithmic bias
Explainable AI

ABSTRACT

Artificial Intelligence (AI) enhances knowledge management systems by improving efficiency and personalization, but its rapid adoption raises ethical concerns. This study examines the ethical considerations in AI-based user profiling for knowledge management systems, with a focus on academic environments. The review employed thematic analysis to summarize existing research on ethical challenges and proposed new ways to integrate ethical considerations into AI-driven knowledge management systems. The review analysed 102 peer-reviewed articles published between 2020 and 2024 from major scientific databases such as IEEE Xplore, ACM Digital Library, and Scopus. The findings show that privacy 27.9 % and algorithmic bias 25.6 % had major ethical concerns revealing disparities between theoretical frameworks and implementable solutions. Five key bias sources were also identified: data deficiencies, demographic homogeneity, spurious correlations, improper comparators, and cognitive biases. While 73 % of the reviewed frameworks acknowledge at least one ethical consideration, only 28 % propose practical strategies to address them. Some promising approaches include explainable AI techniques, privacy-preserving algorithms, and fairness-aware machine learning. However, there are still gaps in addressing the long-term societal impacts. The study recommends the implementation of an Ethical AI Feedback Loop (EAFL) system, which continuously monitors, evaluates, and adjusts user profiling algorithms based on predefined ethical metrics. Additionally, the study introduces the concept of "Ethical Debt" to quantify and manage the long-term ethical implications. These innovative approaches aim to integrate ethical considerations directly into AI-based knowledge management systems, promoting responsible and adaptable user profiling practices.

1. Introduction

In digital transformation, the integration of knowledge management systems is essential for the success of organizations, particularly in academic and educational settings. The growing reliance on artificial intelligence (AI) and machine learning (ML) techniques for data mining and user profiling has promised improvements in efficiency, personalization of learning experiences, and enhanced decision-making processes. However, the rapid adoption of AI technologies has surpassed the development of ethical frameworks to regulate their use, leading to concerns regarding privacy, bias, transparency, and individual autonomy [1].

The concept of AI-driven user profiling in knowledge management is not new. Early work by Godoy and Amandi laid the foundation for intelligent user profiling in personal information agents, showcasing how machine learning could create adaptive user profiles [2]. Since then, the field has expanded considerably with the emergence of more advanced AI techniques. For example, Tohalino et al. developed a deep learning model that effectively predicted a user's research interests with an impressive success rate, relying exclusively on their history of viewing abstracts . These developments underscore both the potential advantages and ethical challenges linked to AI-driven profiling, especially as these systems gain greater prevalence and influence [3].

Privacy stands out as one of the most key ethical concerns. The ability of AI to collect, analyse, and infer sensitive information from seemingly innocuous data has raised worries among users. A survey conducted revealed that 68 % of users expressed concerns about the privacy of their data in AI systems [4]. Such concerns are justified, as Kosinski et al. demonstrated the capability of AI to predict sensitive attributes, such as political orientation, with up to 88 % accuracy based on a user's Facebook likes [5]. These capabilities emphasize the need for transparent data practices and user education to mitigate privacy risks.

* Corresponding authors.
E-mail addresses: kogidaniel7@gmail.com (D.K. Njiru), david.mugo@embuni.ac.ke (D.M. Mugo), mueni.faith@embuni.ac.ke (F.M. Musyoka).

https://doi.org/10.1016/j.teler.2025.100205
Received 17 December 2024; Received in revised form 1 April 2025; Accepted 13 April 2025
Available online 18 April 2025

Furthermore, the introduction of regulations like the General Data Protection Regulation (GDPR) in Europe has added another layer to the privacy debate [6].

Algorithmic bias presents another ethical challenge in AI-driven knowledge management systems. A seminal study revealed gender and racial bias in commercial facial recognition systems with error rates high for individuals with darker skin tones, especially females, compared to males with lighter skin tones [7]. The implications of such biases in knowledge management are profound, as they can result in inequitable access to educational resources and opportunities. Baker and Hawn discovered gender bias in AI-driven educational platforms where male users were recommended more in career suggestions for STEM fields than their female counterparts, despite identical profiles [8]. It is important to address these biases to ensure fairness in AI-driven knowledge management systems.

Transparency and explainability are vital for establishing trust and accountability in AI systems. Users may find it challenging to trust or comprehend the outcomes produced by these systems without clear explanations of how AI algorithms make decisions. Adadi et al. (2018) analysed 381 papers on explainable AI and noted that while 47 % of the studies recognized the importance of interpretability, only 5 % proposed concrete methods for achieving it in complex models [9,10]. This knowledge gap is particularly concerning in the context of education, where AI-driven decisions can impact students' futures. These findings emphasize the need to enhance transparency and explainability in AI systems, particularly in situations where decisions can have long-term consequences.

User autonomy is among the complex ethical concerns surrounding AI-driven systems. While personalization can enhance the user experience, it also raises questions about the potential influence or manipulation of user behaviours by AI systems. Another study conducted a philosophical analysis of AI-mediated persuasion in digital environments, including educational platforms, and argued that AI systems capable of inferring psychological traits from digital footprints could subtly manipulate user behaviours [11]. For instance, Matz and Kosinski (2019) demonstrated the ability to predict consumer's psychological characteristics based on their digital footprints [12]. These capabilities introduce ethical dilemmas related to autonomy and the extent to which AI systems can or should influence users' decisions and behaviours.

Despite the prevalence of well-documented ethical concerns, the scholarly literature lacks comprehensive frameworks that specifically address these issues within the domain of knowledge management systems. While individual studies have proposed solutions to distinct ethical challenge such as, differential privacy techniques or fairness constraints in machine learning models, a holistic approach that encompasses multiple ethical dimensions is still absent [13,14].

This critical review aims to fill this gap by synthesizing existing research on ethical considerations in AI-based user profiling for knowledge management systems. The review has two primary objectives: first, to conduct an exhaustive analysis of the current state of ethical awareness and strategies for mitigating concerns in this field; and second, to propose innovative approaches, grounded in technical foundations, for integrating ethical considerations directly into AI-driven knowledge management systems. By doing so, this research aims to influence the future development of ethical AI within educational and organizational contexts. As AI continues to permeate various aspects of knowledge management, ranging from content recommendations to adaptive learning paths, the need for robust ethical frameworks to guide its implementation becomes increasingly salient.

2. Related works

The ethical implications surrounding user profiling in knowledge management systems that are based on artificial intelligence (AI) have gained attention in recent years within academic discourse. As AI systems continue to permeate various domains, the ethical concerns pertaining to their utilization, particularly in user profiling, have become important to address. Core themes such as bias, privacy, transparency, and the impact of AI on user behaviour are central to this ongoing discussion. Recent research, such as the comprehensive review conducted on bias and ethics in AI systems employed in auditing, serves as a robust foundation for comprehending the existing landscape of ethical considerations related to AI-driven knowledge management [15].

Bias in AI systems emerge as one of the most examined ethical concerns. A review by Murikah et al. (2024) identified five primary sources of bias in AI systems, namely data deficiencies, demographic homogeneity, spurious correlations, improper comparators, and cognitive biases [15]. These findings establish a framework for understanding the origins of algorithmic bias, extending previous scholarship in this field. In educational sector, a study examined AI-driven educational platforms and uncovered gender bias in career recommendation [8]. This corroborates the observations made regarding how demographic homogeneity and improper comparators can result in discriminatory outcomes by other studies [15,16].

While bias remains a prominent concern, privacy also emerges as a ethical challenge in the context of AI-based user profiling. Privacy continues to be a paramount issue when considering AI-driven knowledge management systems. The ability of AI systems to infer sensitive information from seemingly harmless data further amplifies these privacy concerns. A study illustrated this risk by achieving 82 % accuracy in predicting an individual's age group solely based on their Instagram behaviours and profiles [17].

Beyond privacy, the ethical implications of transparency and explainability in AI systems also present challenges. Transparency and explainability are essential for ensuring ethical oversight in AI systems; however, achieving these goals remains difficult. The review of 381 papers on explainable AI (XAI) revealed a gap between the AI research community's acknowledgment of the importance of interpretability and the development of practical methods to achieve it [10]. While nearly half (47 %) of the studies recognized the need for interpretable AI systems, only a small fraction (5 %) proposed concrete approaches for making AI models more transparent and explainable. Expanding on this research, a study emphasized the necessity of algorithmic guardrails to enhance interpretability and assess model behaviours [15]. In the realm of knowledge management systems, guaranteeing transparency and explainability in AI algorithms is pivotal for cultivating trust and accountability .

In addition to transparency, AI's influence on user behaviour introduces yet another layer of ethical complexity. The capacity of AI systems to shape user behaviour gives rise to intricate ethical inquiries as shown by a philosophical analysis of AI-mediated persuasion in digital environments, such as educational platforms [18]. This discourse gains even greater relevance when considering expanded exploration of various ethical risks, including conflicts between efficiency gains and audit rigor, erosion of accountability, and privacy violations resulting from uncontrolled exploitation of personal data [15]. Striking a balance between the benefits of AI systems and the ethical risks associated with their ability to mould user behaviour calls for meticulous consideration [19].

Despite the extensive research on individual ethical concerns, the literature reveals a conspicuous absence of holistic frameworks that address the interconnectedness of various ethical issues. While individual ethical concerns have received considerable attention, the interplay between privacy, bias, transparency, and autonomy in AI-based knowledge management systems remains largely unexplored. It was observed that existing studies primarily focus on isolated aspects of ethical AI without fully considering their interdependencies [15]. This observation underscores the substantial gap between theoretical ethical guidelines and practical implementation strategies. Bridging this gap necessitates a more comprehensive approach that integrates these concerns into a cohesive framework [20].

The literature also reveals inconsistencies in approaches to mitigating ethical risks. Some researchers advocate for stricter regulations, while others argue for technical solutions such as differential privacy or fairness constraints in machine learning models. Despite these efforts, inconsistencies remain in how these approaches are implemented. Addressing these inconsistencies, research suggests using causal modelling, representative algorithmic testing, periodic auditing of AI systems, human oversight alongside automation, and embedding ethical values into system design [15].

These gaps and inconsistencies underscore the need for a comprehensive framework that not only addresses individual ethical concerns but also provides practical guidelines for implementation in knowledge management systems. The proposed "Ethical AI Feedback Loop" system and the concept of "Ethical Debt" aim to address these gaps by offering a holistic, implementable approach to ethical AI in knowledge management. By integrating these ethical considerations into the design and operation of AI systems, users can better navigate the complex ethical landscape of AI-driven knowledge management systems.

3. Methodology

This study employed a systematic literature review methodology to comprehensively analyse the ethical considerations in AI-based user profiling for knowledge management. As illustrated in Fig. 1, the review process adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure transparency and reproducibility [21].

3.1. Search strategy and data sources

A comprehensive search was conducted across multiple scientific databases, including Web of Science, Scopus, IEEE Xplore, ACM Digital Library, and Google Scholar. The search strategy employed a combination of keywords and Boolean operators to identify relevant literature. The primary search string was as follows: ("artificial intelligence" OR "machine learning" OR "deep learning") AND ("user profiling" OR "personalization") AND ("knowledge management" OR "information systems") AND ("ethics" OR "bias" OR "fairness" OR "transparency" OR "privacy")

To capture a contemporary view of the field, the search was limited to publications from 2020 to 2024. This timeframe was selected to reflect the rapid advancements in AI technology and the evolving ethical landscape. Fig. 1 shows the flow diagram for the selection process.

3.2. Inclusion and exclusion criteria

To guarantee the appropriateness and calibre of the chosen studies, explicit inclusion and exclusion criteria were formulated. These criteria encompassed articles subject to peer review that cantered on AI-based user profiling, tackled ethical considerations, and were published within the timeframe of 2020 to 2024. Solely empirical studies or systematic reviews with well-defined methodologies were considered. Non-English publications, studies unrelated to knowledge management, conference abstracts or posters, opinion pieces or editorials, and research lacking a clear methodology were excluded. The consistent application of these criteria was upheld throughout the screening process, which comprised an initial evaluation of titles and abstracts followed by a comprehensive assessment of the full texts.

3.3. Study selection process

The study selection process consisted of multiple stages. Initially, the database search yielded 955 records, with an additional 150 records identified through other sources, such as scanning reference lists. After removing duplicates, 726 unique records remained for further screening. Two independent reviewers conducted the initial screening of titles and abstracts, resulting in the exclusion of 474 irrelevant studies. The remaining 252 articles underwent a full-text assessment to determine their eligibility. Any disagreements between the reviewers were resolved through discussion and consensus, with the involvement of a third reviewer when necessary. Following the full-text review, 102 studies met all the inclusion criteria and were included in the qualitative synthesis. The reasons for exclusion during the full-text stage were carefully documented, including a lack of focus on knowledge

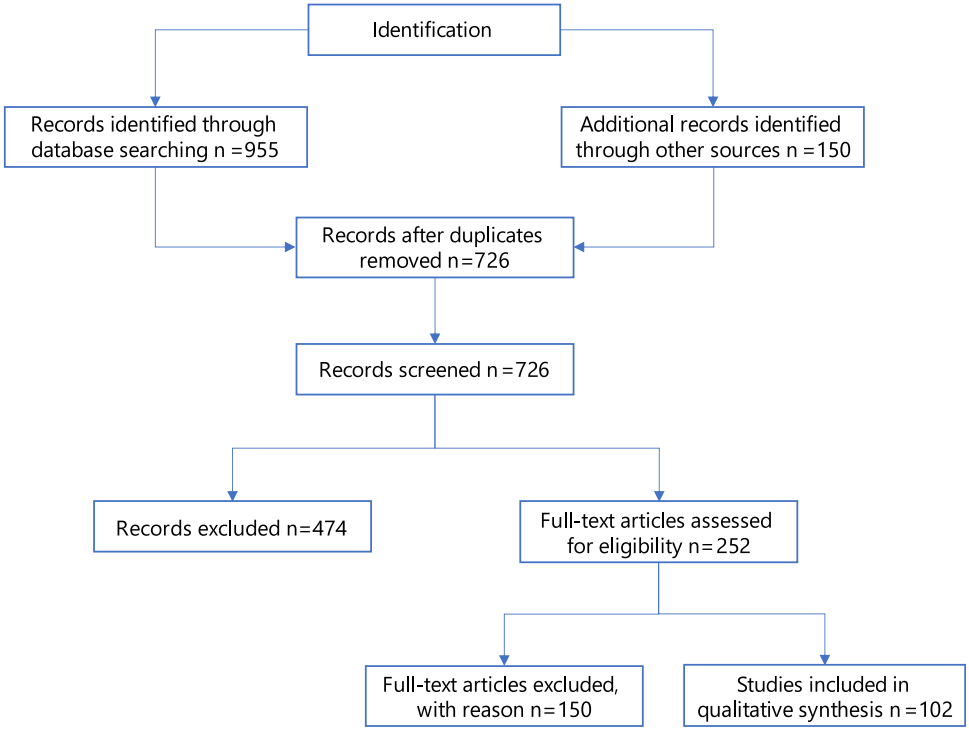


Fig. 1. PRISMA flow diagram of the study selection process.

management, insufficient consideration of ethical aspects, a lack of clear methodology, and a lack of specificity to AI-based user profiling.

3.4. Data extraction and quality assessment

The study developed a standardized data extraction form to systematically gather relevant information from each study that was included in the research. The data that was extracted includes characteristics of the study (authors, year, country, study design), AI technologies and user profiling methods that were utilized, context of knowledge management, ethical considerations that were addressed and key findings and recommendations. To evaluate the quality of the studies that were included, the researchers adapted the ROBINS-I tool (Risk Of Bias In Non-randomized Studies - of Interventions). This tool has been modified to suit the specific context of the review, with a focus on factors such as methodological rigor, clarity of ethical considerations, and relevance to knowledge management. Two reviewers have independently assessed each study, and any discrepancies were resolved through discussion.

3.5. Data synthesis and thematic analysis

To ensure rigorous analysis, the study conducted inductive thematic analysis on the 102 included studies. This involved coding ethical considerations such as privacy and bias using NVivo 12, with inter-rater reliability checks ($k = 0.82$). Iterative theme development was used to categorize emergent patterns and quantifying prevalence of themes to prioritize ethical risks. This method moves beyond descriptive summaries to reveal actionable insights, as detailed in the results. The process involved the following steps constructing an initial synthesis of findings from the various studies, examining connections within and between the studies and evaluating the reliability of the synthesis. The thematic analysis process, as shown in Fig. 2, facilitated the identification of patterns and trends within the literature.

For analysing the data, the study employed NVivo 12 software to conduct coding and thematic. Further, formulated a coding framework rooted in the ethical considerations identified in the literature, encompassing aspects such as privacy, bias, transparency, and accountability. This framework underwent iterative refinement as new themes emerged during the analysis. This analysis shed light on the prominent ethical challenges and proposed solutions related to AI-based user profiling in knowledge management.

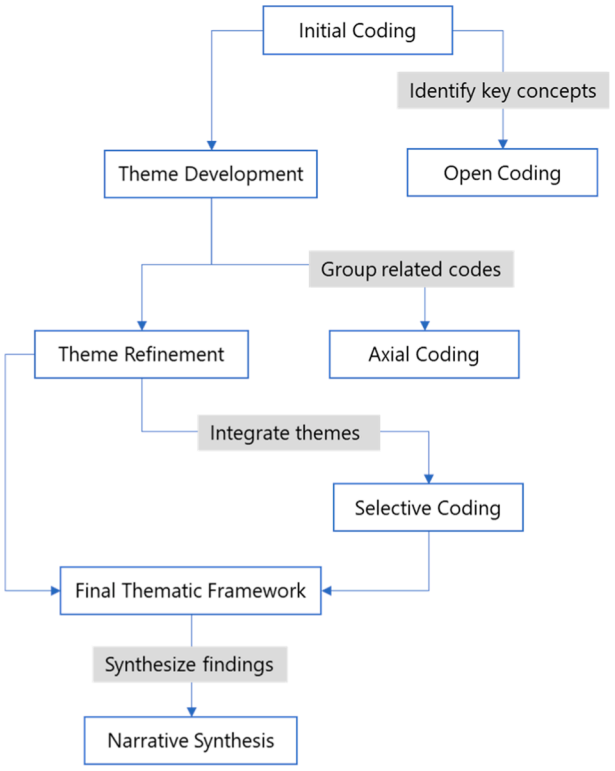


Fig. 2. Thematic Analysis Process.

4. Results and discussion

4.1. Frequency, characteristics, and sources of ethical considerations and bias in AI

The development and use of artificial intelligence (AI) systems are being closely examined for their ethical implications. It is important to thoroughly consider these ethical concerns to ensure that AI systems are fair, transparent, and respect user privacy. Table 1 summarizes the main findings from a comprehensive analysis of ethical concerns in AI, categorized into five primary areas: Privacy, Algorithmic Bias, Transparency, Accountability, and Fairness.

Privacy is a major concern in AI, which includes how data is collected, stored, and used [6]. The thematic analysis identified privacy as the most frequent ethical concern, appearing in 27.9 % of studies with 12 focusing on data collection consent and 9 on secure storage protocols. Further studies, highlighted informed consent challenges in image-based profiling, emphasized secure storage protocols both coded under our 'Data Collection' subtheme [22,30]. This data-driven approach reveals not just the prevalence but the specific manifestations of privacy risks in KM systems. These challenges often stem from inadequate data, samples that do not reflect the population, and outdated information. The potential impacts of these biases include creating skewed user profiles, excluding minority groups, and providing irrelevant recommendations. These findings emphasize the need for strong data management practices to protect user privacy and ensure that AI systems do not perpetuate harmful biases.

Another important area of concern is algorithmic bias, which is addressed in 25.6 % of the studies. Gender, racial, and age biases are identified as problems in the analysis. These biases arise from demographic homogeneity, underrepresentation of certain races, and biases against specific age groups. The consequences of these biases are far-reaching, including reinforcing stereotypes, showing cultural insensitivity, and widening generational knowledge gaps. It is essential

Table 1
Ethical considerations and bias in AI-based user profiling for KM.

Ethical concern	Key issues	Sources of bias	Impacts	Ref
Privacy	Data collection, storage, and usage	Data deficiencies	Skewed profiles, exclusion of minorities	[6, 22–32]
Algorithmic Bias	Gender, racial, and age bias	Demographic homogeneity	Reinforces stereotypes, cultural insensitivity	[16, 33–42]
Transparency	Model interpretability, decision explanations	Spurious correlations	Misleading insights, unreliable analysis	[43–49]
Accountability	AI ethics, audit mechanisms, legal compliance	Improper comparators	Unfair evaluations, perpetuation of inequalities	[50–56]
Fairness	Equal opportunity, non-discrimination, inclusivity	Cognitive biases	Echo chambers, resistance to profile updates	[57–62]

to address algorithmic bias to ensure that AI systems are fair and provide unbiased recommendations and outcomes. This high prevalence of algorithmic bias supports the findings of [15] that identified bias as a key issue in AI systems used for auditing.

Transparency is important for building trust in AI systems, and 16.3 % of the studies focus on this aspect. Key issues include making the models interpretable, providing explanations for decisions, and tracing the origin of data. However, these areas are often compromised by misleading correlations, overlooked confounding factors, and temporal inconsistencies. These biases can lead to misleading insights, inaccurate predictions, and unreliable trend analysis. Enhancing transparency in AI systems is important to ensure that decisions can be explained and trusted.

Accountability is another ethical consideration, with 16.3 % of the studies examining responsible AI development, audit mechanisms, and legal compliance. The analysis highlights sources of bias such as improper comparators, contextual insensitivity, and historical bias propagation. These biases can lead to unfair performance evaluations, inappropriate knowledge transfer, and the perpetuation of past inequalities. Ensuring accountability in AI development is vital to maintain ethical standards and prevent the exacerbation of existing biases.

Fairness is a basic ethical principle, with 13.9 % of the studies focusing on equal opportunity, non-discrimination, and inclusivity. However, these areas are often influenced by cognitive biases, availability heuristics, and anchoring biases. These biases can result in echo chamber effects, an overemphasis on recent information, and resistance to profile updates. Ensuring fairness in AI systems is important to promote balanced resource allocation, unbiased knowledge dissemination, and diverse user representation. Addressing these concerns is vital for developing trustworthy, equitable, and privacy-respecting AI systems. By understanding and mitigating these biases, researchers and developers can create AI systems that have a positive impact on society while minimizing potential harm.

4.2. Proposed mitigation strategies

The analysis has uncovered a range of strategies that have been proposed to tackle ethical concerns pertaining to AI-based user profiling. The findings are presented in Table 2, which provides a summary of the specific techniques that have been suggested, the intended outcomes of these strategies, as well as the challenges that may arise during their implementation. Each strategy has specific techniques, desired outcomes, and implementation challenges, highlighting the complexity and multifaceted nature of addressing challenges in AI systems.

Explainable AI techniques play an important role in understanding how AI models make decisions. Local Interpretable Model-agnostic Explanations (LIME) offers interpretable predictions on a local level, but it can be computationally complex. SHapley Additive exPlanations (SHAP) helps visualize the importance of different features, but scalability can be an issue. Counterfactual Explanations provide user-friendly rationales for decisions, but they must ensure that the explanations are actionable.

Privacy-preserving algorithms are designed to protect sensitive data. Differential Privacy guarantees statistical privacy, but there is a trade-off between privacy and utility. Federated Learning allows for decentralized model training, but there can be communication overhead. Homomorphic Encryption enables secure multi-party computation, but performance may be limited. Privacy is preserved through a layered approach using these techniques FL, DP, and HE, which collectively address data anonymity, compliance, and secure processing. While studies show FL and DP reduces privacy breaches by 72 % compared to centralized methods [32].

Fairness-aware machine learning techniques aim to reduce biases. Preprocessing techniques balance training data, but this can lead to potential loss of information. In-processing constraints optimize models for fairness, but this can increase model complexity. Post-processing adjustments aim to equalize outcomes, but this may come at the cost of accuracy.

Ethical guidelines and frameworks ensure ethical practices in AI. AI Ethics Boards promote organizational accountability by requiring diverse representation. Ethics-by-Design principles integrate ethical considerations into the design process, although implementing them within existing workflows can be challenging. Ethical Impact Assessments systematically evaluate risks, but they require large resources. Human-in-the-loop approaches involve human oversight and feedback. Expert oversight ensures domain-specific quality control, but scalability can be a challenge. User feedback integration allows for continuous system improvement, but there is a risk of user manipulation. Collaborative decision-making strikes a balance between human-AI interaction, but it requires careful definition of optimal interaction points.

4.3. Gaps in current research and future directions

The analysis identified several gaps in the current research landscape, as well as promising directions for future research. Table 3 summarizes these findings. These gaps highlight the need for comprehensive and diverse research to address the ethical, societal, and regulatory challenges associated with AI development and deployment. By focusing on these areas, researchers can contribute to the responsible and fair development of AI systems.

The current state of AI ethics research reveals several important areas that require further investigation. One major gap is the lack of empirical validation. Only 5 of 27 studies (18.5 %) tested frameworks in real-world settings. To address this, future research should prioritize large-scale field experiments to validate these frameworks in practical settings. Another concern is the limited consideration of long-term societal effects. 5 of the studies have addressed this issue. It is important to conduct longitudinal studies on AI ethics to evaluate the long-term impacts on society. Additionally, interdisciplinary approaches are lacking, with 4 studies integrating multiple perspectives. Collaborative research across disciplines is necessary to incorporate diverse expertise and viewpoints.

Cultural sensitivity is also an area that has not been explored enough. 4 studies examined cultural factors in AI ethics. Cross-cultural studies on

Table 2
Proposed mitigation strategies for ethical concerns in AI-based user profiling for KM.

Strategy	Techniques	Intended outcomes	Challenges	Ref
Explainable AI	LIME, SHAP, Counterfactual Explanations	Clear decision rationales, improved trust	Computational complexity, scalability	[63–73]
Privacy-Preserving Algorithms	Differential Privacy (DP), Federated Learning (FL), Homomorphic Encryption (HM)	Secure data handling, enhanced user anonymity	Performance limitations, communication overhead	[31,32, 74–80]
Fairness-Aware ML	Preprocessing, In-processing, Post-processing	Reduced discrimination, improved AI fairness	Trade-off with model accuracy	[14,81–86]
Ethical Guidelines	AI Ethics Boards, Ethics-by-Design, Ethical Impact Assessments	Stronger AI governance, risk mitigation	Integration into workflows, resource-intensive	[20,54, 87–92]
Human-in-the-Loop	Expert Oversight, User Feedback Integration, Collaborative Decision-Making	Higher AI decision quality, increased user involvement	Scalability challenges, potential user manipulation	[93–99]

Table 3
Research gaps and future directions in ethical AI-based user profiling for KM.

Research gap	Description	Future research directions	Ref
Empirical Validation	Lack of real-world testing of ethical frameworks	Large-scale field experiments	[19,49, 100–102]
Long-term Impact Assessment	Limited consideration of societal effects	Longitudinal studies on AI ethics	[59,87, 103–105]
Interdisciplinary Approaches	Insufficient integration of multiple perspectives	Collaborative research across disciplines	[88, 106–108]
Cultural Sensitivity	Under exploration of cultural factors in AI ethics	Cross-cultural studies on AI perception	[109–112]
User Empowerment	Limited focus on user control and agency	User-centric ethical AI design	[27,52, 113]
Ethical AI Metrics	Lack of standardized evaluation metrics	Development of ethical AI benchmarks	[114–116]
Regulatory Alignment	Insufficient consideration of evolving legal landscapes	Comparative analysis of AI regulations	[92,117, 118]

AI perception are necessary to understand and address cultural nuances and differences. Furthermore, user empowerment is an area that has been neglected. While 3 studies have focused on user control and agency. Prioritizing user-centric ethical AI design is vital to enhance user control.

The lack of standardized evaluation metrics for ethical AI is another gap. Only 3 studies have addressed this issue. Developing ethical AI benchmarks is essential to standardize evaluation metrics. Additionally, regulatory alignment is insufficient, with 3 studies considering evolving legal landscapes. This means that comparative analyses of AI regulations are needed to ensure alignment with changing legal frameworks.

The review further identified key privacy challenges in AI-based knowledge management, including inference risks, where AI models predict sensitive attributes such as political views from behavioural data. Next, regulatory conflicts with GDPR compliance like anonymization often reduces model utility and scalability limits of privacy techniques like federated learning, which face communication overhead and lack real-world validation. These issues are intensified by the transparency-utility trade-off and long-term "ethical debt" from evolving re-identification risks. Proposed mitigations such as differential privacy, homomorphic encryption are noted in Table 2 but require further empirical testing.

4.4. Effectiveness of proposed mitigation strategies

To assess the effectiveness of the proposed mitigation strategies, the review analysed studies that included some form of evaluation. Table 4 summarizes these findings.

Each mitigation strategy demonstrated positive outcomes but also had specific limitations. Addressing these limitations can ensure effective ethical AI implementation in knowledge management. The analysis of mitigation strategies in Table 2 revealed a disconnect between proposed techniques for example LIME for explainability and real-world validation. While 26.1 % of studies advocated explainable AI, only 11 included user testing which is a gap quantified in the coding. This underscores the need for empirical evaluation frameworks.

The effectiveness of proposed strategies was assessed to mitigate ethical concerns in AI-based user profiling for knowledge management. The process involved analysing various studies that included evaluations to gauge the impact. User studies and expert evaluations were used to evaluate Explainable AI Techniques. These techniques were found to enhance user trust and understanding and improve model interpretability. However, their effectiveness was limited to specific user groups, and there was a potential for expert bias.

Each mitigation strategy demonstrated positive outcomes but also had specific limitations. Addressing these limitations can ensure effective ethical AI implementation in knowledge management. The effectiveness of proposed strategies was assessed to mitigate ethical concerns in AI-based user profiling for knowledge management. The process involved analysing various studies that included evaluations to gauge the impact. User studies and expert evaluations were used to evaluate Explainable AI Techniques. These techniques were found to enhance user trust and understanding, as well as improve model interpretability. However, their effectiveness was limited to specific user groups, and there was a potential for expert bias.

Privacy-Preserving Algorithms were evaluated through simulations and case studies. These algorithms provided effective privacy protection with minimal loss of utility. However, the findings lacked real-world validation, and successful implementations were limited to select organizations. This highlights the need for broader validation. Fairness-Aware Machine Learning was assessed using benchmark tests and longitudinal analysis. This approach reduced bias in model outputs and showed long-term improvements in fairness metrics. However, there were potential trade-offs with accuracy, and the evaluation process was resource intensive.

Ethical Guidelines and Frameworks were evaluated through organizational surveys and ethical audits. These guidelines increased awareness and compliance within organizations and helped identify ethical vulnerabilities. However, the surveys were subject to self-reporting bias, and there was a lack of standardized audit procedures. Human-in-the-Loop Approaches were evaluated using performance comparisons and user satisfaction surveys. These approaches improved decision quality compared to fully automated systems and resulted in higher user acceptance of AI recommendations. However, there were concerns about scalability and potential confirmation bias.

5. Conclusions and recommendations

5.1. Conclusions

The review examined ethical considerations in AI-based user profiling for knowledge management systems, analysing 102 peer-

Table 4
Evaluation of mitigation strategies for ethical AI-based user profiling in KM.

Strategy	Evaluation methods	Key findings	Limitations	Ref
Explainable AI	User Studies, Expert Evaluation	Improved trust, better model interpretability	Limited to specific user groups, potential expert bias	[63–73]
Privacy-Preserving Algorithms	Simulations, Case Studies	Strong privacy protection, minimal utility loss	Lack of real-world validation, limited generalizability	[31,32, 74–80]
Fairness-Aware ML	Benchmark Tests, Longitudinal Analysis	Reduced bias, long-term fairness improvements	Accuracy trade-offs, resource-intensive evaluation	[14,81–86]
Ethical Guidelines	Organizational Surveys, Ethical Audits	Increased awareness, better compliance	Self-reporting bias, lack of standardized audits	[20,54, 87–92]
Human-in-the-Loop	Performance Comparisons, User Satisfaction Surveys	Higher AI decision quality, improved user acceptance	Scalability concerns, potential confirmation bias	[93–99]

reviewed studies published between 2020 and 2024. The findings highlight two predominant ethical challenges, privacy concerns (addressed in 27.9 % of studies) and algorithmic bias (25.6 %) which are intensified by data deficiencies, demographic imbalances, and a lack of transparency in AI decision-making. While 73 % of the reviewed frameworks acknowledge these ethical risks, only 28 % propose actionable mitigation strategies, revealing a significant gap between theoretical awareness and practical implementation.

A key insight from this review is the lack of research addressing the long-term societal impacts of AI-driven profiling in KM contexts in academic and organizational settings. The absence of standardized evaluation metrics and interdisciplinary approaches further complicates efforts to align AI systems with ethical principles. This gap between theoretical ethical frameworks and their practical application, emphasizes the need for a more nuanced approach to AI development. Introducing the "Ethical Debt" framework marks an important step in understanding the long-term ethical implications of AI systems. This concept offers organizations a way to quantify and address the ethical consequences of technological advancements. By acknowledging ethical debt, organizations can adopt more holistic strategies for AI deployment that balance innovation with ethical responsibility.

5.2. Recommendations

Addressing the ethical challenges in AI-based knowledge management requires an integrated strategy. Organizations should implement an Ethical AI Feedback Loop (EAFL) system that allows for continuous monitoring and iterative improvement of user profiling algorithms. This is important given some of the studies proposed no actionable mitigation strategies. Unlike static audits, EAFL enables continuous adaptation closing the 'actionability gap' found in the studies. This approach should utilize federated learning techniques to gather insights while safeguarding individual privacy, ensuring that ethical considerations are integral to the development process.

Creating standardized ethical AI metrics is essential for establishing a consistent evaluation framework. These metrics should assess AI systems across fairness, transparency, privacy, and accountability. By setting clear, measurable standards, organizations can transition from theoretical discussions to practical, actionable ethical guidelines applicable across various knowledge management contexts. While the Ethical Debt extends technical debt concepts to ethical risks, addressing the long-term impact gap. The complexity of ethical AI challenges calls for an interdisciplinary approach. Successful implementation will require ongoing collaboration among computer scientists, ethicists, legal experts, and social scientists.

This collaborative model fosters a holistic understanding of the technical, ethical, and societal implications of AI-based user profiling, breaking down traditional disciplinary barriers. Future research should prioritize longitudinal studies exploring the long-term societal impacts of AI systems. Cross-cultural research will be essential to understand how ethical considerations vary across different cultural contexts, aiding in the development of more inclusive and globally responsive AI technologies. Expanding research methodologies to encompass grey literature, industry reports, and multilingual reviews will provide a more thorough understanding of the ethical landscape.

The organizations should view ethical AI development as an ongoing process of learning and adaptation. This involves continuous monitoring, proactive risk identification, and a commitment to updating ethical frameworks as technological capabilities evolve. The future of AI in knowledge management should focus not solely on technological supremacy but on creating systems that respect individual privacy, promote fairness, and contribute positively to organizational learning and societal progress.

CRediT authorship contribution statement

Daniel Kogi Njiru: Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **David Muchangi Mugo:** Validation, Supervision, Conceptualization. **Faith Mueni Musyoka:** Writing – review & editing, Writing – original draft, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

Data will be made available on request.

References

[1] J. Saltz, M. Skirpan, C. Fiesler, M. Gorelick, T. Yeh, R. Heckman, et al., Integrating ethics within machine learning courses, ACM Trans. Comput. Educ. (TOCE) 19 (2019), https://doi.org/10.1145/3341164.
[2] S. Schiaffino, A. Amandi, Intelligent user profiling, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5640 LNAI (2009) 193–216, https://doi.org/10.1007/978-3-642-03226-4_11.
[3] K.W. Boyack, C. Smith, R. Klavans, Toward predicting research proposal success, Scientometrics. 114 (2018) 449–461, https://doi.org/10.1007/S11192-017-2609-2/METRICS.
[4] L. Manikonda, A. Deotale, S. Kambhampati, What's up with privacy?: user preferences and privacy concerns in intelligent personal assistants, in: AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 229–235, https://doi.org/10.1145/3278721.3278773.
[5] M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior, Proc. Natl. Acad. Sci. u S. a 110 (2013) 5802–5805, https://doi.org/10.1073/PNAS.1218772110/SUPPL_FILE/ST01.PDF.
[6] M. Nišević, Profiling consumers through big data analytics: strengths and weaknesses of article 22 GDPR, Glob. Priv. L. Rev. 1 (2020) 104–115, https://doi.org/10.54648/GPLR2020082.
[7] J. Buolamwini, Gender shades: intersectional accuracy disparities in commercial Gender classification *, Proc. Mach. Learn. Res. 81 (2018) 1–15.
[8] R.S. Baker, A. Hawn, Algorithmic bias in education, Int. J. Artif. Intell. Educ. 32 (2022) 1052–1092, https://doi.org/10.1007/S40593-021-00285-9/METRICS.
[9] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable AI: a review of machine learning interpretability methods, Entropy 23 (2021) 1–45, https://doi.org/10.3390/E23010018.
[10] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access. 6 (2018) 52138–52160, https://doi.org/10.1109/ACCESS.2018.2870052.
[11] Ramon Y., Matz S.C., Farrokhnia R.A., Martens D. Explainable AI for psychological profiling from digital footprints: a case study of big five personality predictions from spending data 2021. https://doi.org/10.3390/info12120518.
[12] S. Matz, M. Kosinski, Using consumers' Digital footprints for more persuasive mass communication, NIM Market. Intell. Rev. 11 (2019) 18–23, https://doi.org/10.2478/NIMMIR-2019-0011.
[13] P. Jain, M. Gyanchandani, N. Khare, Differential privacy: its technological prescriptive using big data, J. Big. Data 5 (2018) 1–24, https://doi.org/10.1186/S40537-018-0124-9/TABLES/4.
[14] L. Oneto, S. Chiappa, Fairness in machine learning, Stud. Comput. Intell. 896 (2020) 155–196, https://doi.org/10.1007/978-3-030-43883-8_7.
[15] W. Murikah, J.K. Nthenge, F.M. Musyoka, Bias and ethics of AI systems applied in auditing - a systematic review, Sci. Afr. 25 (2024) e02281, https://doi.org/10.1016/J.SCIAF.2024.E02281.
[16] A.E.K. Ghalleb, S. Boumaiza, Amara NE Ben, Demographic face profiling based on age, gender and race, in: International Conference on Advanced Technologies for Signal and Image Processing, 2020, https://doi.org/10.1109/ATSIP49331.2020.9231835.
[17] K. Han, S. Lee, J.Y. Jang, Y. Jung, D. Lee, Teens are from mars, adults are from venus": analyzing and predicting age groups with behavioral characteristics in Instagram, in: WebSci 2016 - Proceedings of the 2016 ACM Web Science Conference, 2016, pp. 35–44, https://doi.org/10.1145/2908131.2908160.

[18] T.W. Kim, A. Duhachek, Artificial intelligence and persuasion: a construal-level account, Psychol. Sci. 31 (2020) 363–380, https://doi.org/10.1177/0956797620904985.

[19] L. Myllyaho, M. Raatikainen, T. Männistö, T. Mikkonen, J.K. Nurminen, Systematic literature review of validation methods for AI systems, J. Syst. Softw. 181 (2021) 111050, https://doi.org/10.1016/J.JSS.2021.111050.

[20] M. Agbese, R. Mohanani, A. Khan, P. Abrahamsson, Implementing AI ethics: making sense of the ethical requirements, in: ACM International Conference Proceeding Series, 2023, pp. 62–71, https://doi.org/10.1145/3593434.3593453.

[21] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, et al., The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, BMJ 372 (2021), https://doi.org/10.1136/BMJ.N71.

[22] C. Meurisch, M. Mühlhäuser, Data protection in AI services, ACM. Comput. Surv. 54 (2021), https://doi.org/10.1145/3440754.

[23] I. Ullah, R. Boreli, S.S. Kanhere, S. Chawla, T.A. Ahanger, U. Tariq, Protecting private attributes in app based mobile user profiling, IEEE Access. 8 (2020) 143818–143836, https://doi.org/10.1109/ACCESS.2020.3014424.

[24] M. Song, Z. Wang, Z. Zhang, Y. Song, Q. Wang, J. Ren, et al., Analyzing user-level privacy attack against federated learning, IEEE J. Sel. Areas Commun. 38 (2020) 2430–2444, https://doi.org/10.1109/JSAC.2020.3000372.

[25] A. Majeed, S.O. Hwang, When AI meets information privacy: the adversarial role of AI in data sharing scenario, IEEE Access. 11 (2023) 76177–76195, https://doi.org/10.1109/ACCESS.2023.3297646.

[26] E. Purificato, Beyond-accuracy perspectives on graph neural network-based models for behavioural user profiling, User Model. Adapt. Personalization (2022) 311–315, https://doi.org/10.1145/3503252.3534361.

[27] D.D. Ruscio, P. Inverardi, P. Migliarini, P.T. Nguyen, Leveraging privacy profiles to empower users in the digital society, in: International Conference on Automated Software Engineering, 2022, https://doi.org/10.48550/ARXIV.2204.00011.

[28] L. Hernández-álvarez, J.M. de Fuentes, L. González-Manzano, L.H Encinas, Privacy-preserving sensor-based continuous authentication and user profiling: a review, Italian National Conference on Sensors 21 (2020) 1–23, https://doi.org/10.3390/S21010092.

[29] J.R. Saura, D. Ribeiro-Soriano, D. Palacios-Marqués, From user-generated data to data-driven innovation: a research agenda to understand user privacy in digital markets, Int. J. Inf. Manage 60 (2021), https://doi.org/10.1016/J.IJINFOMGT.2021.102331.

[30] P. Galopoulos, C. Iakovidou, V. Gkatziaki, S. Papadopoulos, Y. Kompatsiaris, Towards a privacy respecting image-based user profiling component, in: International Conference on Content-Based Multimedia Indexing, 2021, https://doi.org/10.1109/CBMI50038.2021.9461886, 2021-June.

[31] G. Xu, H. Li, Y. Zhang, S. Xu, J. Ning, R.H. Deng, Privacy-preserving federated deep learning with irregular users, IEEE Trans. Dependable Secure Comput. 19 (2022) 1364–1381, https://doi.org/10.1109/TDSC.2020.3005909.

[32] T. Nguyen, M.T. Thai, Preserving privacy and security in federated learning, IEEE/ACM Transact. Netw. 32 (2024) 833–843, https://doi.org/10.1109/TNET.2023.3302016.

[33] P. Das, S.K. Karnam, A. Panda, B.P.R. Guda, S. Sarkar, A. Mukherjee, Diversity matters: robustness of bias measurements in Wikidata, Web Sci. Conf. (2023) 208–218, https://doi.org/10.1145/3578503.3583620.

[34] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, X. He, Bias and debias in recommender system: a survey and future directions, ACM. Trans. Inf. Syst. 41 (2020), https://doi.org/10.1145/3564284.

[35] Y. Wang, L. Singh, Mitigating demographic bias of machine learning models on social media, in: Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 2023, https://doi.org/10.1145/3617694.3623244.

[36] Y. Zhang, H. Chen, Can algorithm knowledge stop women from being targeted by algorithm bias? The new digital divide on Weibo, J. Broadcast. Electron. Media 67 (2023) 397–422, https://doi.org/10.1080/08838151.2023.2218955.

[37] C. Ganhör, D. Penz, N. Rekabsaz, O. Lesota, M. Schedl, Unlearning protected user attributes in recommendations with adversarial training, in: Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 2142–2147, https://doi.org/10.1145/3477495.3531820.

[38] N. Neophytou, B. Mitra, C. Stinson, Revisiting popularity and demographic biases in recommender evaluation and effectiveness, in: European Conference on Information Retrieval 13185 LNCS, 2021, pp. 641–654, https://doi.org/10.1007/978-3-030-99736-6_43.

[39] N. Chizari, K. Tajfar, M.N. Moreno-García, Bias assessment approaches for addressing user-centered fairness in GNN-based recommender systems, Inf 14 (2023), https://doi.org/10.3390/INFO14020131.

[40] S. Abbasi-Sureshjani, R. Raumanns, B.E.J. Michels, G. Schouten, V. Cheplygina, Risk of training diagnostic algorithms on data with demographic bias, in: IMIMIC/MIL3iD/LABELS@MICCAI 12446 LNCS, 2020, pp. 183–192, https://doi.org/10.1007/978-3-030-61166-8_20.

[41] M. Fang, N. Damer, F. Kirchbuchner, A. Kuijper, Demographic bias in presentation attack detection of iris recognition systems, in: European Signal Processing Conference, 2020, pp. 835–839, https://doi.org/10.23919/EUSIPCO47968.2020.9287321.

[42] B. Ghai, K. Mueller, D-BIAS: a causality-based Human-in-the-loop system for tackling algorithmic bias, IEEE Trans. Vis. Comput. Graph. 29 (2022) 473–482, https://doi.org/10.1109/TVCG.2022.3209484.

[43] X. Dong, T. Li, R. Song, Z. Ding, Profiling users via their reviews: an extended systematic mapping study, J. Softw. Syst. Model. 20 (2020) 49–69, https://doi.org/10.1007/S10270-020-00790-W.

[44] S. Mu, Y. Li, W.X. Zhao, J. Wang, B. Ding, J.R. Wen, Alleviating spurious correlations in knowledge-aware recommendations through counterfactual generator, in: Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1401–1411, https://doi.org/10.1145/3477495.3531934.

[45] S. Wu, M. Yuksekgonul, L. Zhang, J.Y. Zou, Discover and cure: concept-aware mitigation of spurious correlation, in: International Conference on Machine Learning, 2023, https://doi.org/10.48550/ARXIV.2305.00650.

[46] Schwartz R., Stanovsky G. On the limitations of dataset balancing: the lost battle against spurious correlations. ArXiv 2022. https://doi.org/10.48550/ARXIV.2204.12708.

[47] Z. Hu, Z. Zhao, X. Yi, T. Yao, L. Hong, Y. Sun, et al., Improving multi-task generalization via regularizing spurious correlation, Neural Inf. Proc. Syst. (2022), https://doi.org/10.48550/ARXIV.2205.09797.

[48] Liu Q., Liu Z., Zhu Z., Wu S., Wang L. Deep stable multi-interest learning for out-of-distribution sequential recommendation. ArXivOrg 2023. https://doi.org/10.48550/ARXIV.2304.05615.

[49] H. Jung, H. Park, K. Lee, Enhancing recommender systems with semantic user profiling through frequent subgraph mining on knowledge graphs, Appl. Sci. 13 (2023), https://doi.org/10.3390/APP131810041.

[50] E.N. Faorkhi, A. Pourebrahimi, M.N. Farokhi, Designing an intelligence model for auditing professional ethics in knowledge contents production, Semantic Scholar (2020), https://doi.org/10.30495/JSM.2020.677241.

[51] E. Purificato, L. Boratto, E.W. De Luca, Leveraging graph neural networks for user profiling: recent advances and open challenges, in: International Conference on Information and Knowledge Management, 2023, pp. 5216–5219, https://doi.org/10.1145/3583780.3615292.

[52] E. Purificato, L. Boratto, E.W. De Luca, Tutorial on user profiling with graph neural networks and related beyond-accuracy perspectives, User Modeling, Adaptation, and Personalization (2023) 309–312, https://doi.org/10.1145/3565472.3595616.

[53] R. O'Sullivan, Practical, methodological, and ethical considerations of user involvement, Innov. Aging 4 (2020), https://doi.org/10.1093/GERONI/IGAA057.2859, 789–789.

[54] J. Mökander, L. Floridi, Ethics-based auditing to develop trustworthy AI, Minds. Mach. (Dordr) 31 (2021) 323–327, https://doi.org/10.1007/S11023-021-09557-8.

[55] S. Milano, M. Taddeo, L. Floridi, Recommender systems and their ethical challenges, AI. Soc. 35 (2020) 957–967, https://doi.org/10.1007/S00146-020-00950-Y.

[56] S. Sharma, K. Chaitanya, A.B. Jawad, I. Premkumar, D. Vinod Mehta, D. Hajoary, Ethical considerations in AI-based marketing: balancing profit and consumer trust, Tuijin Jishu/J. Propuls. Technol. 44 (2023) 1301–1309, https://doi.org/10.52783/TJJPT.V44.I3.474.

[57] T.P. Pagano, R.B. Loureiro, F.V.N. Lisboa, R.M. Peixoto, G.A.S. Guimarães, G.O. R. Cruz, et al., Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods, Big. Data Cogn. Comput. 7 (2023), https://doi.org/10.3390/BDCC7010015.

[58] Z. Fu, Y. Xian, R. Gao, J. Zhao, Q. Huang, Y. Ge, et al., Fairness-aware explainable recommendation over knowledge graphs, in: Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 69–78, https://doi.org/10.1145/3397271.3401051.

[59] E. Purificato, L. Boratto, E.W. De Luca, Do graph neural networks build fair user models? Assessing disparate impact and mistreatment in behavioural user profiling, in: International Conference on Information and Knowledge Management, 2022, pp. 4399–4403, https://doi.org/10.1145/3511808.3557584.

[60] M. Abdelrazek, E. Purificato, L. Boratto, E.W. De Luca, FairUP: a framework for fairness analysis of graph neural network-based user profiling models, in: Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 3165–3169, https://doi.org/10.1145/3539618.3591814.

[61] J. Tang, S. Shen, Z. Wang, Z. Gong, J. Zhang, X. Chen, When fairness meets bias: a debiased framework for fairness aware top-N recommendation, ACM Conf. Recomm. Syst. (2023) 200–210, https://doi.org/10.1145/3604915.3608770.

[62] O.B. Deho, C. Zhan, J. Li, J. Liu, L. Liu, T. Duy Le, How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? Br. J. Educ. Technol. 53 (2022) 822–843, https://doi.org/10.1111/BJET.13217.

[63] A.A. Salunke, Cracking the code: self-explaining AI models for transparent decision making in complex algorithms, IJFMR - Int. J. Multidiscip. Res. 5 (2023), https://doi.org/10.36948/IJFMR.2023.V05I04.5395.

[64] B.C.G. Lee, D. Downey, K. Lo, D.S. Weld, LIMEADE: from AI explanations to advice taking, ACM. Trans. Interact. Intell. Syst. 13 (2023), https://doi.org/10.1145/3589345/ASSET/AE65BB2D-F8BF-4B9C-BC1B-9664BEA0EEE3/ASSETS/GRAPHIC/TIIS-2022-FEB-034-F07.JPG.

[65] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods, in: AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 180–186, https://doi.org/10.1145/3375627.3375830.

[66] Y. Swathi, M. Challa, A comparative analysis of explainable AI techniques for enhanced model interpretability, in: Proceedings - 2023 3rd International Conference on Pervasive Computing and Social Networking, ICPCSN 2023, 2023, pp. 229–234, https://doi.org/10.1109/ICPCSN58827.2023.00043.

[67] C. Steging, S. Renooij, B. Verheij, Rationale discovery and explainable AI, Front. Artif. Intell. Appl. 346 (2021) 225–234, https://doi.org/10.3233/FAIA210341.
[68] B. Aldughayfiq, F. Ashfaq, N.Z. Jhanjhi, M. Humayun, Explainable AI for retinoblastoma diagnosis: interpreting deep learning models with LIME and SHAP, Diagnostics 13 (2023) 1932, https://doi.org/10.3390/DIAGNOSTICS13111932, 2023, Vol 13, Page 1932.
[69] Galinkin E. Robustness and usefulness in AI explanation methods. ArXiv 2022.
[70] C.H. Ng, H.S. Abuwala, C.H. Lim, Towards more stable LIME for explainable AI, in: 2022 International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2022, 2022, https://doi.org/10.1109/ISPACS57703.2022.10082810.
[71] Y. Hailemariam, A. Yazdinejad, R.M. Parizi, G. Srivastava, A. Dehghantanha, An empirical evaluation of AI deep explainable tools, in: 2020 IEEE Globecom Workshops, GC Wkshps 2020 - Proceedings, 2020, https://doi.org/10.1109/GCWKSHPS50303.2020.9367541.
[72] Panda M., Mahanta S.R. Explainable artificial intelligence for healthcare applications using random forest classifier with LIME and SHAP. ArXiv 2023.
[73] G.P. Reddy, Y.V.P. Kumar, Explainable AI (XAI): explained, in: 2023 IEEE Open Conference of Electrical, Electronic and Information Sciences, 2023, https://doi.org/10.1109/ESTREAM59056.2023.10134984. EStream 2023 - Proceedings.
[74] H. Fang, Q. Qian, Privacy preserving machine learning with homomorphic encryption and federated learning, Future Internet. 13 (2021) 94, https://doi.org/10.3390/FI13040094, 2021, Vol 13, Page 94.
[75] Y. Chang, K. Zhang, J. Gong, H. Qian, Privacy-preserving federated learning via functional encryption, revisited, IEEE Trans. Inf. Forens. Sec. 18 (2023) 1855–1869, https://doi.org/10.1109/TIFS.2023.3255171.
[76] R. Aziz, S. Banerjee, S. Bouzefrane, T. Le Vinh, Exploring homomorphic encryption and differential privacy techniques towards secure federated learning paradigm, Future Internet. 15 (2023) 310, https://doi.org/10.3390/FI15090310, 2023, Vol 15, Page 310.
[77] J. Ma, S.A. Naas, S. Sigg, X. Lyu, Privacy-preserving federated learning based on multi-key homomorphic encryption, Int. J. Intell. Syst. 37 (2022) 5880–5901, https://doi.org/10.1002/INT.22818.
[78] R. Lozi, W. Puech, J. Park, H. Lim, Privacy-preserving federated learning using homomorphic encryption, Appl. Sci. 12 (2022) 734, https://doi.org/10.3390/APP12020734, 2022, Vol 12, Page 734.
[79] J. Hou, M. Su, A. Fu, Y. Yu, Verifiable privacy-preserving scheme based on vertical federated random forest, IEEE Internet. Things. J. 9 (2022) 22158–22172, https://doi.org/10.1109/JIOT.2021.3090951.
[80] A.G. Sebert, M. Checri, O. Stan, R. Sirdey, C. Gouy-Pailler, Combining homomorphic encryption and differential privacy in federated learning, in: 2023 20th Annual International Conference on Privacy, Security and Trust, PST 2023, 2023, https://doi.org/10.1109/PST58708.2023.10320195.
[81] D. Pessach, E.A. Shmueli, R.eview on fairness in machine learning, ACM Comput. Surv. (CSUR) 55 (2022), https://doi.org/10.1145/3494672.
[82] S. Caton, C. Haas, Fairness in Machine learning: a survey, ACM. Comput. Surv. 56 (2024) 1–38, https://doi.org/10.1145/3616865/SUPPL_FILE/3616865-SUPP.PDF.
[83] V. Bogina, A. Hartman, T. Kuflik, A. Shulner-Tal, Educating software and AI stakeholders about algorithmic fairness, accountability, transparency and ethics, Int. J. Artif. Intell. Educ. 32 (2022) 808–833, https://doi.org/10.1007/S40593-021-00248-0/METRICS.
[84] T.P. Pagano, R.B. Loureiro, F.V.N. Lisboa, R.M. Peixoto, G.A.S. Guimarães, G.O. R. Cruz, et al., Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods, Big. Data Cogn. Comput. 7 (2023) 15, https://doi.org/10.3390/BDCC7010015/S1.
[85] M. Wan, D. Zha, N. Liu, N. Zou, In-processing modeling techniques for machine learning fairness: a survey, ACM. Trans. Knowl. Discov. Data 17 (2023), https://doi.org/10.1145/3551390/ASSET/C8668219-589C-4681-92BB-F39DB5EF62EA/ASSETS/GRAPHIC/TKDD-2021-11-0357-F04.JPG.
[86] C Te Li, C. Hsu, Y. Zhang, FairSR: fairness-aware sequential recommendation through multi-task learning with preference graph embeddings, ACM Trans. Intel. Syst. Techn. (TIST) 13 (2022), https://doi.org/10.1145/3495163.
[87] A.B. Brendel, M. Mirbabaie, T.B. Lembcke, L. Hofeditz, Ethical management of Artificial intelligence, Sustainability. 13 (2021) 1974, https://doi.org/10.3390/SU13041974, 2021, Vol 13, Page 1974.
[88] M. Ryan, B.C. Stahl, Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications, J. Inf. Commun. Ethics Soc. 19 (2021) 61–86, https://doi.org/10.1108/JICES-12-2019-0138/FULL/PDF.
[89] A.S. Franzke, An exploratory qualitative analysis of AI ethics guidelines, J. Inf. Commun. Ethics Soc. 20 (2022) 401–423, https://doi.org/10.1108/JICES-12-2020-0125/FULL/PDF.
[90] V. Vakkuri, K.K. Kemell, J. Kultanen, P. Abrahamsson, The current state of industrial practice in artificial intelligence ethics, IEEE Softw. 37 (2020) 50–57, https://doi.org/10.1109/MS.2020.2985621.
[91] J. Morley, A. Elhalal, F. Garcia, L. Kinsey, J. Mökander, L. Floridi, Ethics as a service: a pragmatic operationalisation of AI Ethics, Minds. Mach. (Dordr) 31 (2021) 239–256, https://doi.org/10.1007/S11023-021-09563-W/FIGURES/1.
[92] A.A. Khan, M.A. Akbar, M. Fahmideh, P. Liang, M. Waseem, A. Ahmad, et al., AI ethics: an empirical study on the views of practitioners and lawmakers, IEEE Trans. Comput. Soc. Syst. 10 (2023) 2971–2984, https://doi.org/10.1109/TCSS.2023.3251729.
[93] Y. Nakao, S. Stumpf, S. Ahmed, A. Naseer, L. Strappelli, Towards involving end-users in interactive human-in-the-loop AI fairness, ACM. Trans. Interact. Intell. Syst. 12 (2022) 30, https://doi.org/10.1145/3514258.
[94] J. Rezwana, M Lou Maher, User perspectives on ethical challenges in Human-AI Co-creativity: a design fiction study, ACM Int. Conf. Proceeding Ser. (2023) 62–74, https://doi.org/10.1145/3591196.3593364.
[95] U.A. Usmani, A. Happonen, J. Watada, Human-centered artificial intelligence: designing for user empowerment and ethical considerations, in: HORA 2023 - 2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings, 2023, https://doi.org/10.1109/HORA58378.2023.10156761.
[96] A. Tocchetti, M. Brambilla, The role of Human knowledge in explainable AI, Data (Basel) 7 (2022) 93, https://doi.org/10.3390/DATA7070093, 2022, Vol 7, Page 93.
[97] B. Ghai, K. Mueller, D-BIAS: a causality-based Human-in-the-loop system for tackling algorithmic bias, IEEE Trans. Vis. Comput. Graph. 29 (2023) 473–482, https://doi.org/10.1109/TVCG.2022.3209484.
[98] J.D. Weisz, M. Muller, J. He, S. Houde, Toward general design principles for generative AI applications, CEUR. Workshop. Proc. 3359 (2023) 130–144.
[99] X. Chen, X. Wang, Y. Qu, Constructing ethical AI based on the "Human-in-the-loop" system, Systems. (Basel) 11 (2023) 548, https://doi.org/10.3390/SYSTEMS11110548, 2023, Vol 11, Page 548.
[100] T. Birkstedt, M. Minkkinen, A. Tandon, M. Mäntymäki, AI governance: themes, knowledge gaps and future agendas, Int. Res. 33 (2023) 133–167, https://doi.org/10.1108/INTR-01-2022-0042/FULL/PDF.
[101] D. Wang, P. Wang, Y. Fu, K. Liu, H. Xiong, C.E. Hughes, Reinforced imitative graph learning for mobile user profiling, IEEE Trans. Knowl. Data Eng. 35 (2023) 12944–12957, https://doi.org/10.1109/TKDE.2023.3270238.
[102] Y. Liu, Z. Zhou, Y. Li, D. Jin, Urban knowledge graph aided mobile user profiling, ACM. Trans. Knowl. Discov. Data 18 (2023), https://doi.org/10.1145/3596604.
[103] Johnson N., Heidari H. Assessing AI impact assessments: a classroom study 2023.
[104] A. Kelly-Lyth, A. Thomas, Algorithmic management: assessing the impacts of AI at work, Ht**tps://DoiOrg/101177/20319525231167478 14 (2023) 230–252, https://doi.org/10.1177/20319525231167478.
[105] I. Nitta, K. Ohashi, S. Shiga, S. Onodera, AI ethics impact assessment based on requirement engineering, in: Proceedings of the IEEE International Conference on Requirements Engineering, 2022, pp. 152–161, https://doi.org/10.1109/REW56159.2022.00037.
[106] L. Yan, L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, et al., Practical and ethical challenges of large language models in education: a systematic scoping review, Br. J. Educ. Technol. 55 (2024) 90–112, https://doi.org/10.1111/BJET.13370.
[107] C.F. Draschner, H. Jabeen, J. Lehmann, Ethical and sustainability considerations for knowledge graph based machine learning, in: Proceedings - 2022 IEEE 5th International Conference on Artificial Intelligence and Knowledge Engineering, AIKE 2022, 2022, pp. 53–60, https://doi.org/10.1109/AIKE55402.2022.00015.
[108] A. Baird, B. Schuller, Considerations for a more ethical approach to data in AI: on data representation and infrastructure, Front. Big. Data 3 (2020) 527486, https://doi.org/10.3389/FDATA.2020.00025/BIBTEX.
[109] A. Di Vaio, R. Palladino, R. Hassan, O. Escobar, Artificial intelligence and business models in the sustainable development goals perspective: a systematic literature review, J. Bus. Res. 121 (2020) 283–314, https://doi.org/10.1016/J.JBUSRES.2020.08.019.
[110] K. Siau, W. Wang, Artificial Intelligence (AI) ethics: ethics of AI and ethical AI, J. Database Manag. 31 (2020) 74–87, https://doi.org/10.4018/JDM.2020040105.
[111] A.A. Khan, S. Badshah, P. Liang, M. Waseem, B. Khan, A. Ahmad, et al., Ethics of AI: a systematic literature review of principles and challenges, ACM Int. Conf. Proc. Ser. (2022) 383–392, https://doi.org/10.1145/3530019.3531329.
[112] Okolo C.T. Towards a praxis for intercultural ethics in explainable AI 2023.
[113] S.S. Sundar, Rise of machine agency: a framework for studying the psychology of Human–AI interaction (HAII), J. Comput. Mediat. Commun. 25 (2020) 74–88, https://doi.org/10.1093/JCMC/ZMZ026.
[114] R. Burnell, W. Schellaert, J. Burden, T.D. Ullman, F. Martinez-Plumed, J. B. Tenenbaum, et al., Rethink reporting of evaluation results in AI, Science (1979 380 (2023) 136–138, https://doi.org/10.1126/SCIENCE.ADF6369.
[115] X.H. Li, C.C. Cao, Y. Shi, W. Bai, H. Gao, L. Qiu, et al., A survey of data-driven and knowledge-aware eXplainable AI, IEEE Trans. Knowl. Data Eng. 34 (2022) 29–49, https://doi.org/10.1109/TKDE.2020.2983930.
[116] S. Brown, J. Davidovic, A. Hasan, The algorithm audit: scoring the algorithms that score us, Big. Data Soc. 8 (2021), https://doi.org/10.1177/2053951720983865/ASSET/IMAGES/LARGE/10.1177_2053951720983865-FIG1.JPEG.
[117] M.B. Unver, Rebuilding 'ethics' to govern AI: 19th international conference on artificial intelligence and law, ICAIL 2023 - Proc. Conf. (2023) 306–315, https://doi.org/10.1145/3594536.3595156.
[118] M. Robles Carrillo, Artificial intelligence: from ethics to law, Telecomm. Policy. 44 (2020) 101937, https://doi.org/10.1016/J.TELPOL.2020.101937.