Review

Toward Fairness, Accountability, Transparency, and Ethics in AI for Social Media and Health Care: Scoping Review

Aditya Singhal[1], MSc; Nikita Neveditsin[2], MSc; Hasnaat Tanveer[3], BSc; Vijay Mago[4], PhD

[1]Department of Computer Science, Lakehead University, Thunder Bay, ON, Canada

[2]Department of Mathematics and Computing Science, Saint Mary's University, Halifax, NS, Canada

[3]Faculty of Mathematics, University of Waterloo, Waterloo, ON, Canada

[4]School of Health Policy and Management, York University, Toronto, ON, Canada

Corresponding Author:
Nikita Neveditsin, MSc
Department of Mathematics and Computing Science
Saint Mary's University
923 Robie Street
Halifax, NS, B3H 3C3
Canada
Phone: 1 902 420 5893
Email: Nikita.Neveditsin@smu.ca

Abstract

Background: The use of social media for disseminating health care information has become increasingly prevalent, making the expanding role of artificial intelligence (AI) and machine learning in this process both significant and inevitable. This development raises numerous ethical concerns. This study explored the ethical use of AI and machine learning in the context of health care information on social media platforms (SMPs). It critically examined these technologies from the perspectives of fairness, accountability, transparency, and ethics (FATE), emphasizing computational and methodological approaches that ensure their responsible application.

Objective: This study aims to identify, compare, and synthesize existing solutions that address the components of FATE in AI applications in health care on SMPs. Through an in-depth exploration of computational methods, approaches, and evaluation metrics used in various initiatives, we sought to elucidate the current state of the art and identify existing gaps. Furthermore, we assessed the strength of the evidence supporting each identified solution and discussed the implications of our findings for future research and practice. In doing so, we made a unique contribution to the field by highlighting areas that require further exploration and innovation.

Methods: Our research methodology involved a comprehensive literature search across PubMed, Web of Science, and Google Scholar. We used strategic searches through specific filters to identify relevant research papers published since 2012 focusing on the intersection and union of different literature sets. The inclusion criteria were centered on studies that primarily addressed FATE in health care discussions on SMPs; those presenting empirical results; and those covering definitions, computational methods, approaches, and evaluation metrics.

Results: Our findings present a nuanced breakdown of the FATE principles, aligning them where applicable with the American Medical Informatics Association ethical guidelines. By dividing these principles into dedicated sections, we detailed specific computational methods and conceptual approaches tailored to enforcing FATE in AI-driven health care on SMPs. This segmentation facilitated a deeper understanding of the intricate relationship among the FATE principles and highlighted the practical challenges encountered in their application. It underscored the pioneering contributions of our study to the discourse on ethical AI in health care on SMPs, emphasizing the complex interplay and the limitations faced in implementing these principles effectively.

Conclusions: Despite the existence of diverse approaches and metrics to address FATE issues in AI for health care on SMPs, challenges persist. The application of these approaches often intersects with additional ethical considerations, occasionally leading to conflicts. Our review highlights the lack of a unified, comprehensive solution for fully and effectively integrating FATE principles in this domain. This gap necessitates careful consideration of the ethical trade-offs involved in deploying existing methods and underscores the need for ongoing research.

(JMIR Med Inform 2024;12:e50048) doi: 10.2196/50048

KEYWORDS

fairness, accountability, transparency, and ethics; artificial intelligence; social media; health care

Introduction

Background

Machine learning (ML) algorithms have become pervasive in today's world, influencing a wide range of fields, from governance and financial decision-making to medical diagnosis and security assessment. These technologies depend on artificial intelligence (AI) and ML to provide results, offering clear advantages in terms of speed and cost-effectiveness for businesses over time [1]. However, as AI research progresses rapidly, the importance of ensuring that its development and deployment adhere to ethical principles has become paramount.

User data on social media platforms (SMPs) can reveal patterns, trends, and behaviors. Platforms such as Twitter (X Corp) are predominantly used by younger individuals and those residing in urban areas [2]. These platforms often impose age restrictions, leading to a potential bias in algorithms trained on their data toward younger, urban demographics. Social media presents a rich source of data invaluable for health research [3], yet using these data without proper consent poses ethical concerns. Furthermore, social media content is influenced by various social factors and should not always be interpreted at face value. For example, certain topics may engage users from specific regions or demographic groups more than others [4], rendering the data less universally applicable. An additional challenge is the trustworthiness of these data. The issue of bias is further exacerbated when AI or ML software is proprietary with a closed source code, making it challenging to analyze and understand the reasons behind biased decisions [3].

The spread of both misinformation and disinformation is a significant concern on social media [5,6], a problem that became particularly acute during the COVID-19 pandemic. False claims about vaccine safety contributed to public mistrust and hesitancy, undermining efforts to control the virus. In tackling this issue, AI tools have been deployed to sift through information and spotlight reliable content for users [7]. These AI systems are trained using health data from trustworthy sources, ensuring the dissemination of scientifically sound information. On the bright side, social media provides a venue for disseminating new health information, offering valuable insights for the health sector [8]. However, the inherent challenges of social media, such as verifying information authenticity and the risk of spreading misinformation, require careful management to guarantee that the health information shared is accurate and reliable.

Fairness, accountability, transparency, and ethics (FATE) research focuses on evaluating the fairness and transparency of AI and ML models, developing accountability metrics, and designing ethical frameworks [9]. Incorporating a human in the loop is one approach to upholding ethical principles in algorithmic processes. For example, in the case of the Correctional Offender Management Profiling for Alternative Sanctions system used within the US judicial system to predict the likelihood of a prisoner reoffending after release, it is recommended that a judge first review the AI's decision to ensure its accuracy. In summary, recognizing the inherent biases in AI and ML, the implementation of systematic models is crucial for maintaining accountability. Efforts in computer science are directed toward enhancing the transparency of AI and ML, which helps uncover the decision-making processes, identify biases, and hold systems accountable for failures [10,11].

Motivation

The American Medical Informatics Association (AMIA) has delineated a comprehensive set of ethical principles for the governance of AI [12] building on the foundations laid out in the Belmont Report [13]: autonomy, beneficence, nonmaleficence, and justice. These principles are critical for the responsible application of AI in monitoring health care–related data on SMPs [7]. The AMIA expanded these principles to include 6 technical aspects—explainability, interpretability, fairness, dependability, auditability, and knowledge management—as well as 3 organizational principles: benevolence, transparency, and accountability. Furthermore, it incorporated special considerations for vulnerable populations, AI research, and user education [12]. Our review emphasized the concept of FATE, which is prevalent in the AI and ML community [14], and discussed its alignment with the principles outlined by the AMIA.

The discourse on AI ethics is notably influenced by geographic and socioeconomic contexts [15]. There has been extensive debate regarding the best practices for evaluating work produced by explanatory AI and conducting gap analyses on model interpretability in AI [16,17]. Recent advancements in ML interpretability have also been subject to review [18]. Table 1 provides a summary of existing studies that discuss FATE in various contexts. These studies reveal a substantial research gap in understanding how the principles of FATE are integrated within the realm of AI in health care on SMPs. Notably, none of the studies have thoroughly investigated the computational methods commonly used to assess the components of FATE and their intricate interrelationships in this domain.

To bridge the identified research gap, this study focused on three pivotal research questions (RQs):

1. What existing solutions address FATE in the context of health care on SMPs? (RQ 1)
2. How do these solutions identified in response to RQ 1 compare with each other in terms of computational methods, approaches, and evaluation metrics? (RQ 2)
3. What is the strength of the evidence supporting these various solutions? (RQ 3)

Our aim was to enrich the domain of FATE by exploring the array of techniques, methods, and solutions that facilitate social

media interventions in health care settings while pinpointing gaps in the current body of literature. This study encompassed the definitions, computational methods, approaches, and evaluation metrics pertinent to FATE in AI along with an examination of FATE in data sets. The novelty of our research lies in delivering a comprehensive analysis of metrics, computational solutions, and the application of FATE principles specifically within the realm of SMPs. This includes a focus on uncovering further research directions and challenges at the confluence of health care, computer science, and social science.

Table 1. An overview of existing studies focusing on fairness, accountability, transparency, and ethics.

Study	Fairness			Accountability			Transparency			Ethics		
	A[a]	B[b]	C[c]	A	B	C	A	B	C	A	B	C
Mehrabi et al [1], 2021	✓	✓	✓									
Golder et al [19], 2017											✓	✓
Bear Don't Walk et al [20], 2022	✓	✓	✓									
Attard-Frost et al [21], 2022	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
Wieringa [9], 2020				✓	✓	✓						
Adadi and Berrada [22], 2018								✓	✓			
Diogo et al [18], 2019	✓					✓	✓	✓	✓			✓
Chakraborty et al [17], 2017	✓			✓			✓		✓			
Hagerty and Rubinov [15], 2019										✓		✓
Vian and Kohler [23], 2016				✓			✓					

[a]Definitions.

[b]Computational methods and approaches.

[c]Evaluation metrics.

Methods

Research Methodology

Our research methodology was grounded in the approach presented by Kofod-Petersen [24] and adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [25]. We used 2 search databases, PubMed and Web of Science, to ensure the reproducibility of the search results in the identification of records. PubMed was chosen for its comprehensive coverage of biomedical literature, providing direct access to the most recent research in health care and its intersections with AI, rendering it indispensable for studies focused on the FATE principles in the domain. Web of Science was selected for its interdisciplinary scope, diversity of publication sources, and rigorous citation analysis, offering a broad and authoritative overview of global research trends and impacts across computer science, social sciences, and health care. In addition, we used Google Scholar, which is recognized as the most comprehensive repository of scholarly articles [26], known for its inclusivity and extensive coverage across multiple disciplines. However, due to the lack of reproducibility of the search results on Google Scholar, we classified it as *other source* for record identification, as shown in Figure 1. Our search across these databases was conducted without any language restrictions, ensuring a comprehensive and inclusive review of the relevant literature.

We conducted a strategic search using Table 2 as a filter to identify research papers pertinent to our review. The table was designed to allow for customization of groups for retrieving varied sets of literature, aiming to find the intersection among these sets. For group 1, we selected "fairness," "accountability," "transparency," and "ethics." These keywords, being integral components of the FATE framework, were an obvious choice for our search queries. In group 2, we identified "natural language processing" and "artificial intelligence" as our keywords. The selection of "natural language processing" was justified by the predominance of textual data on SMPs, necessitating algorithms adept at processing natural language. The inclusion of "artificial intelligence" reflected its broad applicability beyond traditional ML applications. Given that AI encompasses a wide range of advanced technologies, including sophisticated natural language processing (NLP) techniques, its inclusion ensured the comprehensive coverage of relevant studies. Finally, the terms "social media" and "healthcare" were directly pertinent to our review, making their inclusion essential. Consequently, our aim was to encompass a wide spectrum of studies relevant to the topic of our review.

On the basis of Table 1, our initial strategy involved using the intersection of groups as follows: ([group 1, search term 1 ∩ group 2, search term 1] AND [group 1, search term 1 ∩ group 2, search term 2]) ∩ ([group 1, search term 1 ∩ group 3, search term 1] AND [group 1, search term 1 ∩ group 3, search term 2]), which, for simplicity, we condensed to (group 1, search term 1 ∩ group 2, search term 1 ∩ group 2, search term 2 ∩ group 3, search term 1 ∩ group 3, search term 2), as outlined in the search query presented in Textbox 1.

For our queries, we implemented year-based filtering in PubMed and conducted a parallel topic search in Web of Science, limiting the results to articles published since 2012. However, this

approach yielded only 2 publications from each database, a tally considered inadequate for our purposes. Consequently, we opted to broaden our search by applying the union of 2 intersections. The initial formula ([group 1, search term 1 ∩ group 2, search term 1] AND [group 1, search term 1 ∩ group 2, search term 2]) ∪ ([group 1, search term 1 ∩ group 3, search term 1] AND [group 1, search term 1 ∩ group 3, search term 2]) was streamlined to group 1, search term 1 ∩ ([group 2, search term 1 ∩ group 2, search term 2] ∪ [group 3, search term 1 ∩ group 3, search term 2]), as detailed in the search query in Textbox 2, while maintaining the same year range.

Our search queries resulted in 442 records from PubMed and 327 records from Web of Science, as shown in Figure 1. Subsequently, we eliminated duplicates across the 3 sources, consolidating the findings into 672 records for initial screening. During the screening phase, we applied specific inclusion criteria based on an analysis of titles and abstracts to refine the selection: (1) the study primarily addressed FATE principles in the context of health care on SMPs (inclusion criterion 1); (2) the study reported empirical findings (inclusion criterion 2); (3) the study elaborated on definitions, computational methods, approaches, and evaluation metrics (inclusion criterion 3).

This process narrowed down the field to 172 records eligible for full-text assessment. At this stage, we applied our quality criteria to further assess eligibility: (1) we confirmed through full-text screening that the study adhered to inclusion criteria 1, 2, and 3 (quality criterion 1); (2) the study articulated a clear research objective (quality criterion 2).

Ultimately, this led to the selection of 135 articles for inclusion in our review. The complete list of these articles is available in Multimedia Appendix 1 [1-3,5-11,15-23,26-141].

Figure 1. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram for record selection.

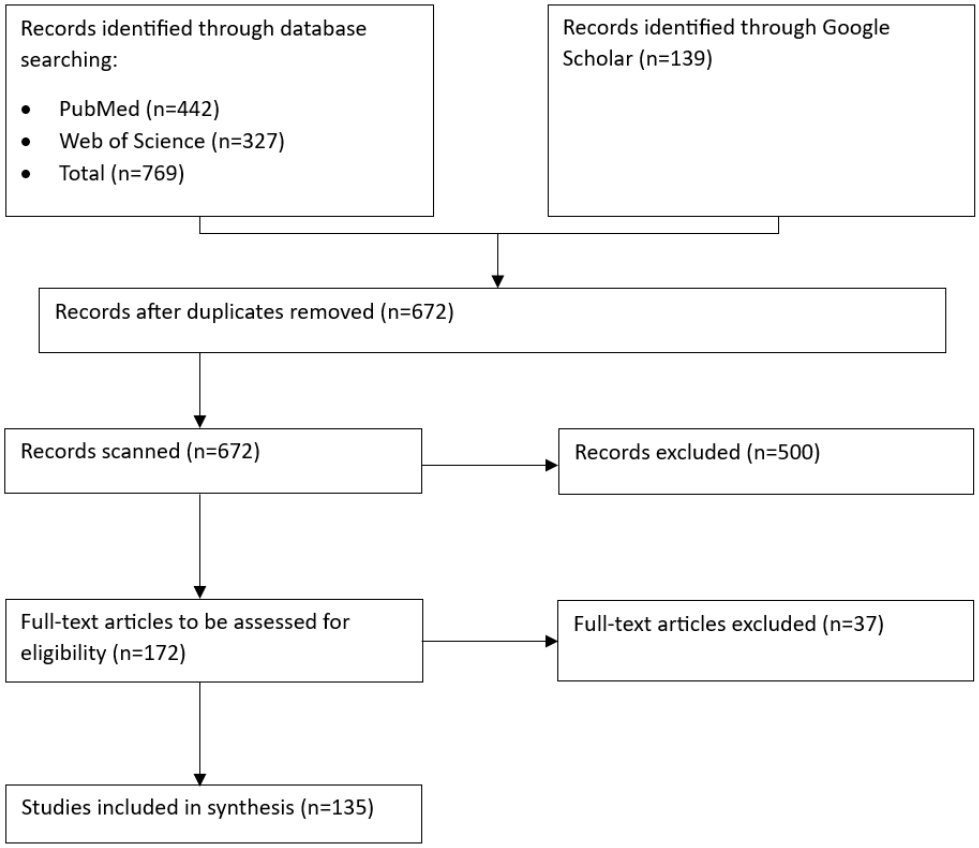

Table 2. Search strategy for finding research articles.

	G1[a]	G2[b]	G3[c]
T1[d]	Quality[e]	Natural language processing	Social media
T2[f]	N/A[g]	Artificial intelligence	Health care

[a]G1: group 1.

[b]G2: group 2.

[c]G3: group 3.

[d]T1: search term 1.

[e]{Fairness, Accountability, Transparency, Ethics}

[f]T2: search term 2.

[g]N/A: not applicable.

Textbox 1. The initial query to the databases.

- ("Fairness" OR "Accountability" OR "Transparency" OR "Ethics") AND ("NLP" or "Natural Language Processing") AND ("AI" OR "Artificial Intelligence") AND ("Healthcare" AND "Social Media")

Textbox 2. Modified query to the databases.

- ("Fairness" OR "Accountability" OR "Transparency" OR "Ethics") AND ((("NLP" or "Natural Language Processing") AND ("AI" OR "Artificial Intelligence")) OR ("Healthcare" AND "Social Media"))

Data Items and Data-Charting

In our review, we incorporated the following data items: (1) approaches and definitions related to each component of FATE; (2) mathematical formulations and algorithms designed to address FATE; (3) methodologies for the integration of FATE principles into AI and ML systems, particularly within health care settings on SMPs; (4) characteristics of the AI or ML systems under study, encompassing their type, application areas within health care, and the specific roles that SMPs play in these systems; (5) outcomes from the formal evaluation or assessment of FATE aspects within the studies, such as their impact on decision-making processes; (6) challenges and barriers reported in the implementation of FATE principles in AI or ML systems; (7) use of frameworks or tools developed to support or evaluate FATE in AI and ML systems; and (8) engagement of stakeholders throughout the AI and ML system's life cycle, including their perspectives on FATE.

The data-charting process involved 3 researchers, each independently extracting pertinent data from the selected sources with a particular focus on the aforementioned data items. For methodical organization and analysis, the extracted information was documented in Microsoft Excel spreadsheets (Microsoft Corp). These spreadsheets were organized alphabetically by the last name of the first author of each article and included references to the corresponding data items as presented in the studies. To consolidate the compiled data, one researcher was tasked with merging the information from these spreadsheets. This step aimed to synthesize the data and ensure a coherent presentation of our findings. The merging process entailed a thorough review and amalgamation of the data charted by each researcher, emphasizing the consolidation of similar approaches and methodologies as identified in the studies.

Results

Definitions, Computational Methods, and Approaches to Fairness

Overview

The understanding of fairness among the public is diverse [26]. The AMIA classifies fairness as a technical principle, emphasizing its importance in creating AI systems that are free from bias and discrimination [12]. This study reviewed various approaches to achieving fairness, with a particular focus on perspectives that facilitate the quantification of fairness in the context of AI for health care on SMPs. The mathematical formulations used to measure fairness are presented in Multimedia Appendix 2 [27-32,142,143]. The following subsections offer a comprehensive examination of approaches to ensure fairness.

Calibrated Fairness

Calibrated fairness seeks to balance providing equal opportunities for all individuals with accommodating their distinct differences and needs [33]. For instance, in the context of social media, a calibrated fair algorithm aims to ensure equal access to opportunities, such as visibility for all users, while also considering specific factors, such as language or location, to offer a personalized experience. In health care, such an algorithm would ensure that all patients have access to the same standard of care yet take into account variables such as age and health status to tailor the best possible treatment plan. The objective is to find a balance between treating everyone equally and acknowledging individual differences to achieve the most equitable outcomes. Fairness metrics, including the false positive rate difference [29] and the equal opportunity difference [34], are used to evaluate the degree of calibrated fairness. Common

computational methods used to achieve calibrated fairness include the following: (1) preprocessing—modifying the original data set to diminish or eliminate the impact of sensitive attributes (eg, gender and ethnic background) on the outcome of an ML model [35]; (2) in-processing—integrating fairness constraints into the model's training process to ensure calibration with respect to sensitive attributes [35]; (3) postprocessing—adjusting the model's output after training to calibrate it in relation to sensitive attributes [35]; (3) adversarial training—training the model on adversarial examples, which are designed to test the model's fairness in predictions [36].

Each of the approaches to achieving calibrated fairness in AI systems has a specific application context that is influenced by various factors. Preprocessing aims to directly mitigate biases in the data before the model's training phase but may present challenges in preserving the integrity of the original data, potentially resulting in the loss of important information. In contrast, in-processing involves the integration of fairness constraints during the model's learning process, which, while aiming to ensure fairness, might compromise model performance due to the added constraints. Postprocessing, which adjusts the model's outputs after training, may appear as a straightforward solution but often falls short in addressing the root causes of bias, thus providing a superficial fix. Adversarial training stands out as a promising approach by challenging the model's fairness through specially designed examples; however, its effective implementation can be complex and resource intensive. Each method has inherent trade-offs between fairness, accuracy, and complexity. The choice among them depends on the specific circumstances of the application, including the nature of the data, the criticality of the decision-making context, and the specific fairness objectives.

Statistical Fairness

Statistical fairness considers various factors, including demographic information, that may be pertinent to the concept of fairness within a specific context. Among the widely recognized statistical definitions of fairness are demographic parity, equal opportunity, and equal treatment [37]. The measure of "demographic parity" is used to reduce data bias by incorporating penalty functions into matrix-factorization objectives [38], whereas the "equal opportunity" metric is crucial for ensuring that decisions are devoid of bias [39]. In the realm of social media, individual notions of fairness might encompass issues such as unbiased content moderation, equitable representation of diverse perspectives and voices, and transparency in the algorithms used for content curation and ranking. Common approaches for measuring statistical fairness include the following: (1) equalized odds—this approach evaluates fairness by examining the differences in true positive and false positive rates across various groups [40]; (2) theorem of equal treatment—this approach assesses fairness by comparing how similar individuals from different groups are treated [41].

Moreover, several toolkits have been developed for measuring statistical fairness in ML and AI models. For instance, Aequitas, as introduced by Saleiro et al [42], generates reports aiding in equitable decision-making by policy makers and ML researchers. The AI Fairness 360 toolkit [43] provides metrics and algorithms designed to reduce statistical biases that lead to the unfair treatment of various groups by ML models [44]. Another toolkit, Fairlearn [45], offers algorithms aimed at addressing disparities in the treatment of different demographic groups by an ML model.

Intersectional Fairness

This approach integrates multiple intersecting identity facets, such as race, gender, and socioeconomic status, into decision-making processes concerning individuals [46]. Its objective is to guarantee equitable treatment for all stakeholders, recognizing that the confluence of these identities may exacerbate marginalization and discrimination. Within the realm of social media, an algorithm designed with intersectional fairness in mind ensures that content is neither recommended nor censored in a manner that is prejudiced against a user's race, gender, or socioeconomic status. Similarly, in health care, an algorithm that incorporates intersectional fairness aims to prevent the disproportionate allocation of medical treatments and resources. Intersectional fairness can be operationalized using the worst-case disparity method, which involves evaluating each subgroup individually and comparing the best and worst outcomes to ascertain the precision of the fairness score. Subsequently, the ratio of the maximum to minimum scores is calculated, with a ratio nearing 1 indicating a more equitable outcome [46]. Other prevalent methods and strategies for achieving intersectional fairness include the following: (1) constraint-based methods—these are designed to honor specific fairness constraints, such as providing equal treatment to different groups identified by multiple attributes, through mathematical optimization [47]; (2) causal inference methods—these aim to ensure that the algorithm's outputs are unbiased by examining the causal relationships between inputs and outputs [48]; (3) decision trees and rule-based systems—these are used to guarantee that the algorithm's decisions are informed by relevant factors and free from bias [49].

Constraint-based methods are adept at enforcing predefined fairness goals; however, the complexity of defining and optimizing these goals poses a significant challenge. In contrast to constraint-based methods, causal inference methods do not necessitate predefined fairness constraints but require a thorough comprehension of the data at hand. Erroneous assumptions regarding causality can result in flawed assessments of fairness. Decision trees and rule-based systems, owing to their interpretability, facilitate the understanding of algorithmic decisions. However, their simplicity may be a limitation as they may not adequately address the complexities inherent in various data sets. To mitigate some of the discussed shortcomings, supervised ranking, unsupervised regression, and reinforcement in fairness evaluation can be approached through pairwise evaluation [50]. This technique involves assessing an AI model's performance by comparing its outputs against a preselected set of input data pairs.

Definitions, Computational Methods, and Approaches to Accountability

Overview

The AMIA considers accountability a fundamental organizational principle, stressing that organizations should bear the responsibility for continuously monitoring AI systems. This includes identifying, reporting, and managing potential risks. Furthermore, organizations are expected to implement strategies for risk mitigation and establish a system for the submission and resolution of complaints related to AI operations [12]. In the following subsections, we explore prevalent views on accountability within the ML and AI community. In addition, we provide summaries of the measurements for different accountability components as identified in the reviewed literature, which can be found in Multimedia Appendix 3 [51-54,144].

Legal Accountability

Legal accountability encompasses the obligations of entities involved in designing, developing, deploying, and using AI systems for health care purposes on social media [55]. This responsibility includes ensuring that AI systems are developed and used in compliance with relevant laws and regulations in addition to addressing any adverse effects or impacts that might arise from their use. Legal accountability also covers issues such as data protection and privacy along with the duty to prevent the use of AI systems for discriminatory or unethical purposes. Commonly used conceptual methods for achieving legal accountability include the following: (1) transparency—this method involves making AI systems transparent, ensuring that their decision-making processes are explainable and comprehensible [56] (there are existing frameworks designed to enhance transparency in the accountability of textual models [57]); (2) documentation—this involves maintaining detailed records of the systems' design, development, and testing processes, as well as documenting the data used for training them [58] (an initiative toward accountability is the implementation of model cards, which are intended to outline an ML model's limitations and disclose any biases that it may be susceptible to [59]); (3) adjudication—this refers to the creation of procedures for addressing disputes and grievances associated with the use of ML and AI systems [60].

Overall, the pursuit of legal accountability should be carefully balanced with the autonomy of stakeholders and must not hinder innovation.

Ethical Accountability

Ethical accountability ensures that AI systems make decisions that are transparent, justifiable, and aligned with societal values [61]. This encompasses addressing data privacy, securing informed consent, and preventing the perpetuation of existing biases and discrimination. Ethical concerns specific to the use of AI in health care include safeguarding patient privacy, handling sensitive health data responsibly, and avoiding the reinforcement of existing health disparities [62]. Common strategies for achieving ethical accountability include the following: (1) ethical impact assessment—this approach entails assessing the ethical risks and benefits of the system and weighing the trade-offs between them [63]; (2) value alignment—this strategy involves embedding ethical principles and values into the design and development of the system, ensuring that its operations are in harmony with these values [64]; (3) transparency and explanation—this is accomplished by offering clear, understandable explanations of the system's functionality and making its data and algorithms openly available [65]; (4) stakeholder engagement—this involves the active participation of a diverse group of stakeholders, including users, developers, and experts, in all phases of the AI or ML system's life cycle [66].

When crafting ethical AI for disseminating health care–related information on social media, the application of these methodologies varies according to specific tasks. Ethical impact assessments, for instance, are valuable for evaluating the potential advantages, such as enhanced patient engagement via personalized dissemination of health care information, against risks, including privacy breaches and the spread of misinformation. The value alignment method plays a crucial role in pinpointing essential ethical values such as patient privacy, information accuracy, nondiscrimination, and accessibility. This method also supports the performance of regular audits to verify that AI systems continuously reflect these ethical standards. Finally, approaches to stakeholder engagement establish a platform for transparent and continuous communication between stakeholders and developers, thereby promoting a cooperative atmosphere in development.

Technical Accountability

Technical accountability ensures that developers and designers of AI and ML systems are held responsible for maintaining standards of security, privacy, and functionality [67]. This responsibility encompasses the implementation of adequate mechanisms to monitor and manage AI algorithms and address arising technical issues. Within the realms of social media and health care, technical accountability further entails the use of AI technologies to foster ethical decision-making, safeguard user privacy, and ensure that decisions are made fairly and transparently [68]. Common strategies for achieving technical accountability include the following: (1) logging—the practice of recording all inputs, outputs, and decisions to trace the system's performance and pinpoint potential problems [69]; (2) auditing—conducting evaluations to check the system's performance, detect biases, and ensure compliance with ethical and legal standards [70].

Both logging and auditing play critical roles in the development of ethical AI for health care information on social media, each with its unique benefits and challenges. Logging, which captures the inputs, outputs, and decisions of an AI system, is vital for tracking system performance. Nonetheless, the retention of detailed logs, especially those involving sensitive health care information, may introduce privacy concerns and necessitate careful consideration of data protection strategies. Auditing, essential for upholding ethical and legal norms, demands expertise and considerable time to effectively scrutinize complex AI systems. In addition, frameworks designed to enhance AI system accountability are in use. An example is Pandora [71],

representing a significant move toward achieving a holistic approach to accountable AI systems.

Societal Accountability

Societal accountability entails the obligation of stakeholders to ensure that their AI systems align with societal values and interests [72]. This encompasses addressing privacy, transparency, and fairness issues, along with considering the wider social, cultural, and economic impacts that AI systems may have. Achieving societal accountability may require stakeholders to participate in public consultations, develop ethical and transparent regulations and standards for AI use, and enhance public understanding of AI system functionalities and applications. Essentially, it advocates for the development and use of AI systems under the principles of responsible innovation, with society's interests considered at every life cycle stage.

Methods for ensuring societal accountability include the following: (1) regulation and standardization—creating regulations and standards for AI system design and use can help hold these systems accountable to society, safeguarding the rights and interests of all stakeholders [73]; (2) public-private partnerships—fostering collaboration among government agencies, private-sector companies, and other entities to promote the societal accountability of AI and ML systems [74].

To ensure accountability, integrating transparency and fairness into algorithms, designing systems with privacy considerations, and conducting regular audits and evaluations to review AI system performance is critical. Researchers have suggested approaches for holding companies accountable for their AI-related actions [9]. They emphasize the importance of pinpointing specific decision makers within a company responsible for any errors, a crucial step for ensuring equitable accountability. The entity or individuals determining accountability should possess comprehensive knowledge of legal, political, administrative, professional, and social viewpoints regarding the error to guarantee fair and unbiased judgments. Moreover, the consequences imposed on decision makers should be appropriately matched to their areas of responsibility, considering each individual's level of responsibility within the company's hierarchy when deciding on these consequences.

Definitions, Computational Methods, and Approaches to Transparency

Overview

According to the AMIA, transparency is an organizational principle that asserts that an AI system must operate impartially, not favoring its host organization. This principle ensures fairness, treating all stakeholders equally without privileging any party. Moreover, transparency requires stakeholders to be clearly informed that they are interacting with an AI system and not a human [12]. Adadi and Berrada [22] presented a nuanced view on transparency, defining it as the degree to which the workings of an AI system are comprehensible to humans. This definition encompasses providing explanations for the system's decision-making processes, clarifying the data used for system training, and certifying the system's neutrality and nondiscriminatory nature. The balancing act between transparency and privacy presents challenges. For instance, in the analysis of mental health data on SMPs, the difficulty does not lie in pinpointing user-specific attributes (as data are often aggregated) but in the application of these data [75]. Here, transparency intersects with the ethical principle of autonomy, which demands that systems protect individual independence, treat users respectfully, and secure informed consent [12]. Guaranteeing autonomy is particularly crucial in the deployment of AI-powered depression detection systems on social networks [76]. The following subsections will delve into the nuances of transparency in AI, emphasizing the importance of openness in data and algorithmic procedures. This focus is particularly critical in the context of data derived from SMPs. We also introduce some metrics for assessing transparency in Multimedia Appendix 4 [77-81].

Algorithmic Transparency

Algorithmic transparency is the clarity with which one can comprehend the manner in which an AI algorithm or model produces its outputs or decisions [82]. Within the context of AI for health care on SMPs, transparency entails the ability to lucidly grasp the processes and methodologies used in the creation, dissemination, and evaluation of social media interventions for health care objectives [83]. This encompasses an understanding of the data sources that inform these interventions, the algorithms or models that analyze the data and generate the interventions, and the criteria for assessing intervention effectiveness. Algorithmic transparency is crucial for identifying and addressing potential biases or errors in interventions and fostering trust among stakeholders, including patients, health care providers, and regulatory bodies. Several computational techniques can enhance algorithmic transparency: (1) feature importance analysis—this technique identifies the most impactful features or variables in the model's output, shedding light on the decision-making process [84]; (2) model interpretability—this involves designing models whose outputs are easily understood and interpreted by humans [85] (for instance, decision trees and logistic regression models are more interpretable compared to more complex models [86]; detailed discussions of model interpretability will follow in a dedicated subsection); (3) explanation generation—this technique produces explanations for a model's outputs, offering insights into its decision-making process through visualizations or natural language descriptions [87].

Feature importance analysis enhances the comprehension of a model's decision-making process, yet it may not fully elucidate the complex interactions among features or their combined effect on the model's decisions, especially in the case of sophisticated deep neural networks. Models that are inherently interpretable, such as decision trees and logistic regression, promote user trust and facilitate the validation of model behaviors. However, these models might not offer the same level of power and precision as more complex models such as deep neural networks, which restricts their effectiveness in analyzing health care–related social media interactions. On the other hand, explanation generation seeks to clarify the model's reasoning for stakeholders. Nonetheless, guaranteeing that these

explanations are both accurate and reflective of the model's inner workings poses a considerable challenge.

Data Transparency

Data transparency pertains to the comprehensibility of how data are collected, stored, and used in the development of an AI system [88]. Within the realm of AI for health care on SMPs, data transparency delineates the degree to which health care organizations and providers maintain openness and clarity regarding the collection, storage, and use of patient data [89]. This aspect is critical to the design and implementation of social media campaigns, encompassing the provision of explicit information to patients about the nature of the data being collected, their intended uses, the entities granted access, and the measures in place for their protection. By adopting a transparent approach to data collection and use, health care organizations can foster trust among patients and encourage more robust engagement in social media–driven health interventions. Such transparency can significantly enhance patient health outcomes as individuals are more inclined to engage in interventions in which they feel informed, comfortable, and confident. Examples of computational methods to enhance data transparency include the following: (1) data visualization—this method entails the creation of graphical representations of data to simplify user understanding and interpretation [90]; (2) data profiling—this process analyzes data to ascertain their structure, quality, and content, aiding in the identification of issues such as missing values and inconsistencies [91]; (3) data lineage analysis and provenance tracking—this approach tracks the movement of data through various systems and processes to verify their accuracy and reliability [81,92].

A critical consideration in implementing any of the data transparency methods is ensuring that the autonomy and privacy of all stakeholders are upheld.

Process Transparency

Process transparency denotes the capability to comprehend the procedures involved in the development and deployment of an AI system, including the testing and validation methodologies used [93]. Within the sphere of social media and health care, this notion extends to the clarity of decision-making processes that govern the prioritization, display, and dissemination of health-related information on SMPs. This encompasses transparency regarding the algorithms and computational methods used to curate and showcase health-related content as well as the policies and guidelines governing the moderation of user-generated content pertaining to health. Enhancing process transparency allows users to place greater trust in the information and interventions presented to them and affords researchers increased confidence in the data they examine. Several computational techniques can facilitate enhanced process transparency in AI systems: (1) auditability and monitoring—this involves integrating auditing and monitoring functions within the AI system, including tracking the system's performance, detecting biases or other ethical concerns, and pinpointing instances of underperformance [94]; (2) open-source development—this entails the open and transparent creation of AI systems, where the code, data, and models are made accessible to the public. Such transparency fosters enhanced scrutiny and accountability of the system by external parties, including regulators and the general public [95].

Adopting these methods while recognizing their limitations and taking into account additional ethical considerations can foster greater transparency in AI applications for health care interventions on SMPs.

Explainability and Interpretability

According to the AMIA, the concepts of explainability and interpretability in AI are closely intertwined in the context of transparency. Explainability necessitates that AI developers articulate the functions of AI systems using language appropriate to the context, ensuring that users have a clear understanding of the system's intended use, scope, and limitations. Conversely, interpretability concentrates on the system's capability to elucidate its decision-making processes [12]. It is common for researchers to use the terms explainability and interpretability interchangeably [18,96].

In the realm of social media interventions for health care, explainability and interpretability pertain to comprehending how an AI system processes social media data, identifies pertinent information, and bases its recommendations or decisions on those data [97]. Research conducted by Amann et al [98] delves into the explainability aspects of AI in health care from 4 perspectives: technological, medical, legal, and that of the patient. The authors highlighted the critical role of explainability in the medical domain, arguing that its absence could compromise fundamental ethical values in medicine and public health. The pursuit of explainability and interpretability in AI systems remains a vibrant area of research. For AI systems that apply social media interventions in health care, various methods, including feature selection techniques and visualizations, can facilitate a deeper understanding among health care professionals of the AI system's underlying mechanisms and the factors influencing its decision-making process. As Barredo Arrieta et al [99] noted, techniques for interpretability in AI involve the design of models with clear and comprehensible features, which can aid in identifying the factors that impact the AI's decisions, thus simplifying the understanding and explanation of the outcomes. The existing computational approaches to achieving explainability and interpretability include the following: (1) partial dependence plots (PDPs) [98,100]—PDPs elucidate the relationship between specific input variables and the predicted outcome, offering insights into the rationale behind an AI model's decisions; (2) local interpretable model-agnostic explanations (LIME)—LIME elucidates the outputs of ML models by creating a simpler, interpretable model that approximates the behavior of the original model [101]; (3) Shapley additive explanations (SHAP)—unlike LIME, SHAP explains the outputs of ML models by calculating the contribution of each input feature to the final output [102]; (4) counterfactual explanations—this approach identifies the minimal changes required in the input features to alter the model's output, providing insights into alternative decision pathways [103]; (5) using mathematical structures for analyzing ML model parameters—techniques such as concept activation vectors, t-distributed stochastic

neighbor embedding, and singular vector canonical correlation analysis are used for this purpose [104]; (6) attention visualization [105]—techniques for visualizing attention in transformer-based language models used across various NLP tasks on SMPs help reveal the models' inner workings and potential biases; (7) explanation generation—this involves creating natural language or visual explanations for an AI system's decisions (using techniques such as saliency maps, LIME [101], and SHAP [102] in conjunction with NLP methods enhances the generation of comprehensible explanations); (8) applying inherently interpretable models—models such as fuzzy decision trees, which graphically depict the decision-making process akin to standard decision trees, clarify how decisions are made and identify the most influential factors [106]; (9) model distillation—this technique trains a simpler model to approximate the decision boundaries of a more complex model, thereby facilitating the creation of an interpretable model while maintaining the original's performance [107].

While all the aforementioned methods significantly contribute to the explainability and interpretability of AI and ML systems in this domain, it is crucial to recognize their inherent limitations in practical applications. Specifically, PDPs may face challenges with complex unstructured data such as natural language. SHAP can become computationally intensive when dealing with a large number of input features, which is typical in complex models. LIME might yield inconsistent outcomes, and the interpretations from attention visualization techniques necessitate detailed analysis by experts. Explanation generation, which is often dependent on the aforementioned methods, can inherit their flaws, potentially resulting in misleading explanations. Finally, models that are inherently interpretable or refined through distillation techniques might oversimplify, failing to fully encapsulate the complexities of health care interventions on SMPs.

Definitions, Computational Methods, and Approaches to Ethics

Overview

Ethics encompasses a wide range of considerations, many of which align with the AI principles recognized by the AMIA. In the realm of AI, ethics generally pertains to the study and practice of crafting and applying AI technologies in ways that are fair, transparent, and advantageous to all stakeholders [108]. The objective of ethical AI is to ensure that AI systems and their decisions are in harmony with human values, uphold fundamental human rights, and do not cause harm or discrimination to individuals or groups. This encompasses issues related to privacy, data protection, bias, accountability, and explainability [109].

Within the sphere of social media, the digital surveillance of public health data from SMPs should adhere to several key principles: (1) beneficence, ensuring that surveillance contributes to better public health outcomes; (2) nonmaleficence, ensuring that the use of data does not undermine public trust; (3) autonomy, either through the informed consent of users or by anonymizing personal details; (4) equity, ensuring equal access for individuals to public health interventions; and (5) efficiency, advocating for legal frameworks that guarantee continuous access to web platforms and the algorithms that guide decision-making [110]. AI-mediated health care interventions must consider affordability and equity across the wider population. In addition, health-related data gathered from social platforms need to be scrutinized for various biases such as population and behavioral biases using appropriate metrics [111]. The following subsections offer insights into different ethical viewpoints and the methods used to evaluate how well AI systems align with these ethical standards. We also present summaries of quantifications of key ethical elements in Multimedia Appendix 5 [112-115].

Philosophical Ethics

Our review concentrated primarily on the practical application of ethical principles in AI rather than exploring the purely philosophical dimensions of ethics. Consequently, this subsection focuses on a set of general ethical principles directly pertinent to AI. Kazim and Koshiyama [116] examined various philosophical aspects of ethics and supported a human-centric approach to AI. This perspective underscores the significance of designing and using AI systems in ways that uphold human autonomy, dignity, and privacy [116]. Within the realm of health care interventions on social media, the philosophical ethics of AI can be specifically perceived as the application of ethical principles and values to the development and use of AI-powered tools and technologies [117]. This entails scrutinizing the potential benefits and risks associated with using AI to gather, analyze, and interpret health-related data from SMPs. It also involves ensuring that the deployment of such technologies adheres to the ethical principles recognized by the AMIA, including autonomy, beneficence, and nonmaleficence [12]. The ultimate goal is to foster the development and use of AI technologies that enhance health outcomes while minimizing the potential risks and harms that could emerge from their application. Examples of computational methods and models for addressing philosophical ethics include the following: (1) Methods and models focused on the simulation and modeling of ethical dilemmas, such as those using model-based control and Pavlovian mechanisms, are instrumental. These approaches offer valuable insights into the likely outcomes of diverse ethical decisions [118]. (2) Game theory experiments serve as a pivotal means to model and analyze decision-making processes in social contexts, encompassing ethical dilemmas. Notable examples of these experiments include the ultimatum game, the trust game, and the prisoner's dilemma [119]. (3) The field of data analytics provides methods and models that leverage statistical methods and ML algorithms to scrutinize data. This analysis aims to unearth patterns or insights pertinent to ethical questions or dilemmas [120].

Overall, while methods and models for simulating and modeling ethical dilemmas are capable of effectively representing various scenarios and predicting outcomes, there is a risk that they might oversimplify the complexities inherent in real-world ethics and fail to fully encapsulate the nuances of human ethical reasoning. Although game theory experiments provide insightful perspectives on human behavior in ethical dilemmas, they possess an abstract nature that may limit their practical applicability in realistic situations. Moreover, the efficacy of

data analytics methods is heavily dependent on the quality and quantity of the available data. Thus, the application of these methodologies in AI for health care–related interventions on social media should be approached with caution. It is essential to ensure that such applications are in alignment with broader ethical principles.

Professional Ethics

In the context of health care interventions via social media, professional ethics refers to a set of guidelines and principles that guide the behavior of health care professionals engaging with social media as part of their practice [121]. These guidelines may cover aspects such as patient privacy; confidentiality; informed consent; and the appropriate use of SMPs for disseminating health information, which includes avoiding conflicts of interest or biased behavior [122]. Algorithms that are designed to detect and flag fraudulent behavior among stakeholders can play a crucial role in identifying potential breaches of professional ethics [123]. Various modeling approaches, such as the living laboratory model, can support the development of health care professional ethics on SMPs [124]. Some researchers call for the development and implementation of local policies at health care organizations to govern the social media activities of health care professionals, highlighting the significant risks associated with the dissemination of information in health care–related social media endeavors [125].

While enforcing professional ethics is vital, it poses challenges, particularly when the methods used may infringe on the autonomy of stakeholders. The strategies mentioned, although essential for upholding ethics, could inadvertently overstep boundaries, thus eliciting concerns regarding the autonomy and privacy of the individuals involved.

Legal Ethics

Legal ethics refers to the ethical considerations related to complying with the laws, regulations, and policies surrounding health care data privacy and security. This encompasses safeguarding the confidentiality of patient data, adhering to informed consent and data-sharing agreements, and complying with relevant legal and ethical standards [126,127]. Furthermore, it necessitates ensuring that AI models used in social media interventions for health care are developed and used in conformity with applicable regulations and standards. The existing regulatory and ethical oversight frameworks include the following: (1) the Health Insurance Portability and Accountability Act (HIPAA)—this framework is dedicated to implementing privacy regulations for health care data [145]; (2) the General Data Protection Regulation (GDPR)—it mandates compliance with data protection laws and adherence to other relevant legal and regulatory frameworks governing the use of AI in health care and social media interventions [128]; (3) ethical review boards—advocating for Ethics by Design, this approach involves integrating the services of an ethical review board into the development process of any product within an organization [129].

Both HIPAA and the GDPR are pivotal in the realm of data protection; however, they face intrinsic limitations, with HIPAA being constrained by jurisdictional reach and the GDPR being constrained by the specific subjects it safeguards. The Ethics by Design concept encourages the responsible and ethical development of AI. Nonetheless, this approach could potentially decelerate the innovation process due to the additional layer of review and oversight required during the deployment phase.

Other Ethical Considerations

Guttman [130] highlighted a range of ethical concerns tied to health promotion and communication interventions, including issues related to autonomy, equity, the digital divide, consent, and the risk of unintended adverse effects such as stigmatization of certain groups through the use of derogatory terms to describe their medical conditions. The author stressed the importance of identifying and addressing these issues in the context of health care–related communication interventions [130]. This involves safeguarding the privacy and confidentiality of patient data, respecting patient autonomy and consent, and ensuring that the use of SMPs does not harm the patient [131]. Gagnon and Sabus [132] recognized the concerns that health care professionals may have regarding the use of SMPs due to potential factual inaccuracies. Nevertheless, they argued that using social media in health care does not inherently breach ethical principles as long as evidence-based practices are followed, digital professionalism is upheld through controlled information sharing, and the potential benefits of disseminated information outweigh the risks [132].

Bhatia-Lin et al [133] suggested a rubric approach for the ethical use of SMPs in research that is applicable to health care–associated research involving social media surveillance. Wright [63] introduced a framework for assessing the ethical implications of a wide range of technologies whose comprehensiveness renders it a suitable baseline for evaluating the ethical implications of using AI in social media and health care contexts. Various tools, methods, and approaches can aid in ensuring the ethical use of AI within the health care domain on SMPs: (1) data visualization tools—these tools are designed to present complex ethical data in a clear and accessible manner, thus aiding health care professionals and other stakeholders in understanding and making informed decisions [134]; (2) sentiment analysis of social media posts related to health care interventions—this technique identifies ethical issues and concerns, such as biases or stigmatization of certain patient groups, by analyzing the sentiment of social media content [135]; (3) crowdsourcing platforms for ethical feedback—these platforms are developed to gather insights from a wide range of individuals on the ethical implications of AI systems and their recommendations, ensuring the inclusion of diverse perspectives and values (this approach highlights potential ethical concerns that development teams may otherwise overlook [136]); (4) fairness-aware ML algorithms—these algorithms are designed to address and mitigate unfairness in both the training data and the algorithmic decision-making process with the goal of promoting equity [137]; (5) privacy-preserving data analysis—this method emphasizes the protection of sensitive data from unauthorized access while enabling meaningful analysis, thus balancing privacy with utility [138,139]; (6) human-in-the-loop approaches by incorporating human oversight and decision-making into AI systems, these approaches aim to

ensure that technology aligns with social values and ethical principles, thereby promoting responsible use [140]; (7) value-sensitive design—this approach focuses on identifying and integrating social values and ethical principles into the design and development of AI systems, thereby promoting their alignment with societal ethics [141].

In summary, each method has distinct applications and limitations. For instance, sentiment analysis of health care–related social media posts is effective in identifying ethical issues such as biases or stigmatization, yet it is susceptible to misinterpretation due to the inherent ambiguity of natural language. On the other hand, human-in-the-loop approaches may introduce subjectivity and diminish the efficiency of automated systems. Consequently, stakeholders involved in applying AI in social media within the health care domain should be cognizant of these methods' inherent limitations before implementation.

Discussion

Principal Findings and Future Research Directions

Overview

Health care providers leverage social media to advertise their services, engage with individuals, and cultivate community bonds [146]. SMPs enable medical professionals to interact with patients and gather feedback, thereby enhancing patient care. Moreover, social media acts as a medium for health promotion via peer support and disease awareness initiatives and enabling web-based consultations between physicians and patients [147]. To combat misinformation, implementing rigorous fact-checking measures is imperative for the dissemination of accurate health information. It is also vital to oversee the use of these platforms by health professionals to ensure the protection of patient confidentiality.

The key findings of this study are outlined in the following sections.

RQ 1: What Existing Solutions Address FATE in the Context of Health Care on SMPs?

There are 4 identified solutions to FATE in health care discussions on SMPs. First, fairness in this domain is tackled through calibrated, statistical, and intersectional approaches. Calibrated fairness seeks to balance equal opportunities with individual differences, such as language or location. Statistical fairness uses demographic data to prevent biases. Intersectional fairness examines various aspects of an individual's identity. Second, accountability in health care on SMPs is ensured by adhering to legal standards, incorporating ethical principles into system design, and maintaining technical functionality and privacy, as well as through societal regulation and standardization. These measures include protecting data privacy, preventing discriminatory or unethical use of AI systems, conducting ethical impact assessments, enhancing transparency, involving stakeholders, carrying out audits and evaluations, and holding decision makers responsible. Third, transparency in AI within health care on social media emphasizes the importance of understanding AI systems, including their algorithms, data sources, and decision-making processes. Transparency is vital for comprehending how interventions are crafted, disseminated, and assessed, playing a significant role in identifying and rectifying biases or errors, fostering trust among stakeholders, and improving participation in social media–based health interventions. Fourth, ethics in health care on SMPs focuses on the development of AI technologies that are fair, transparent, and beneficial. This encompasses considerations of privacy, data protection, bias, accountability, and explainability. Upholding professional and social ethics, such as ensuring patient privacy and autonomy, is crucial. The primary aim is to guarantee the ethical use of AI in health care on SMPs while reducing potential risks and adverse effects.

RQ 2: How Do the Different Solutions Identified in Response to RQ 1 Compare to Each Other in Terms of Computational Methods, Approaches, and Evaluation Metrics?

The various solutions identified in response to RQ 1 can be compared based on computational methods, approaches, and evaluation metrics. These solutions encompass strategies for achieving calibrated, statistical, and intersectional fairness through a variety of computational methods, including data preprocessing, postprocessing, adversarial training, and decision tree use. Key evaluation metrics for assessing these solutions are equal opportunity and equalized odds. Accountability can be examined from multiple perspectives: legal accountability, achieved through regulatory measures and public-private partnerships; technical accountability, emphasizing logging and auditing; and ethical accountability, focusing on the identification of ethical risks through methods such as ethical impact assessments, value alignment, and stakeholder engagement. Transparency is attainable through several strategies: algorithmic transparency, data transparency, process transparency, and the interpretability and explainability of models. Enhancements in algorithmic transparency can be achieved through feature importance analysis, interpretability techniques for models, and the generation of explanations. Data transparency improvements are facilitated by data visualization, profiling, lineage analysis, and provenance tracking. Process transparency can be bolstered by auditability, monitoring, and adoption of open-source development practices. Although interpretability and explainability remain burgeoning research areas, there is a diverse range of methods for attaining these goals, each suitable for specific contexts. The promotion of ethics in health care on SMPs involves the use of simulation, modeling, data analytics, sentiment analysis, crowdsourcing, and automated systems considering both professional and social ethics.

RQ 3: What Is the Strength of the Evidence Supporting the Different Solutions?

The strength of the evidence supporting the solutions is variable and influenced by research quality, methodology, and the statistical significance of the findings. Concepts such as calibrated, statistical, and intersectional fairness are grounded in substantial research. Computational methods, including data preprocessing, adversarial training, and the use of decision trees, are widely adopted, although the extent of evidence supporting

their efficacy varies. Evaluation metrics such as equal opportunity and equalized odds rely on well-established statistical measures, but their applicability and effectiveness can differ across studies. Within the ethics domain of health care on SMPs, the principles of privacy protection and bias mitigation are robustly supported by research; however, the evidence for the effectiveness of specific solutions may vary. Techniques such as simulation, modeling, data analytics, and crowdsourcing are commonly used, with their success dependent on the specific application context. Due to the rapidly evolving nature of this field, consulting current and reputable sources is essential for accessing the latest research findings.

The findings from this study contribute to the evolving landscape of AI applications within health care on SMPs by enhancing the understanding of the ethical considerations essential for deploying AI in health care. They delineate practical pathways for leveraging social media to improve patient care and engagement. This study offers insights into achieving fairness in this domain through calibrated, statistical, and intersectional approaches, presenting methodologies that balance personalized care with broader demographic considerations and effectively address biases. It identifies accountability measures such as transparency, documentation, adjudication, stakeholder engagement, logging, and auditing as essential for the design and regulation of AI, ensuring its responsible use in health care contexts. Achieving public transparency presents technical and practical challenges; however, entities involved in AI applications within health care should provide comprehensive reports on decision-making factors, data origins and use, and solid scientific evidence supporting their decisions to stakeholders upon request. Finally, ethical considerations, encompassing philosophical, professional, and legal dimensions, should drive the implementation of the 3 core components of FATE: fairness, accountability, and transparency.

Our study identified several research gaps in AI systems within health care on SMPs. First, primary challenge in the integration of AI and health care on SMPs is the collection and use of data that accurately represent diverse populations without inherent biases. Trustworthy data sets are crucial for training large language models for clinical applications, yet these data sets often lack diversity in key demographics such as age, ethnicity, or medical history. This shortfall can result in AI predictions that disproportionately benefit certain groups. Moreover, the process of obtaining informed consent on SMPs is complicated by the limited understanding users have of how their data might be used for health care research. A common scenario involves the use of patient-generated data from web-based health forums or social media support groups where consent is ambiguously defined, thereby raising ethical and privacy concerns. Second, the operationalization of the broad set of ethical principles defined by the AMIA into a cohesive FATE framework presents significant challenges. The pursuit of a unified approach that addresses the components of FATE simultaneously is hampered by potential conflicts among these principles. For example, increasing transparency by making AI decision-making processes more comprehensible can inadvertently risk patient privacy and system security by exposing sensitive data or proprietary algorithms. Third, the application of FATE principles in real-world health care interventions on SMPs is critically underdocumented. There is a notable absence of comprehensive case studies that detail the implementation, challenges, and outcomes of ethical frameworks in practice. Such documentation is essential for grasping how theoretical ethical considerations are translated into practical impacts and for pinpointing areas that need adjustment when applying these principles. The effectiveness and ethical considerations of AI-driven public health campaigns on platforms such as Twitter and Facebook, for instance, are largely unexplored in a manner that would provide actionable insights into their real-world impact and ethical ramifications. Fourth, the current landscape of evaluating FATE in AI systems, particularly at the intersection of health care and social media, is characterized by a lack of methods that can be universally applied across different models and data types. The specific challenges of the health care domain on SMPs, which include the necessity to analyze diverse data formats in real time, call for the development of model-agnostic tools for ethical assessment. Most existing methods are designed for particular models or data types and do not comprehensively address the wide range of health care applications on social media. Furthermore, there is an absence of a clear strategy for assessing the impact of various AI-assisted interactions between health care and social media domains.

Given the identified gaps, our study proposes 5 research directions. First, research should focus on the development of comprehensive models that integrate the FATE framework with the broader ethical principles outlined by the AMIA. This involves pioneering methodologies that ensure a balanced consideration of all ethical dimensions, aiming to uphold each without compromising the significance or effectiveness of the others. For medical professionals and researchers, this direction represents a shift toward creating AI systems in health care that are both technologically advanced and ethically robust, ensuring equitable and responsible AI use in patient care and data management. Second, investigations are needed into merging computational methods with ethical evaluations to devise sophisticated mathematical formulations capable of quantitatively assessing ethical components in AI applications within health care on SMPs. By developing robust metrics and evaluation frameworks, researchers can bridge the theoretical ethical considerations with practical computational methods. This effort aims to facilitate the integration of ethical principles into the design and evaluation of AI technologies, ensuring that they meet the highest standards of medical ethics and patient care. Third, exploration is required into ethical trade-offs by focusing on understanding and mitigating inherent conflicts between different ethical components within the FATE framework. By systematically examining these trade-offs, research could aim to find innovative solutions that minimize conflicts, such as between transparency and privacy or between fairness and accountability. For the medical and research community, acknowledging and navigating these trade-offs is crucial for the development and implementation of AI systems that are both ethically responsible and effective in achieving health care goals. Fourth, investigation is necessary into the application of FATE principles in real health care interventions on SMPs. This direction seeks to understand the ethical impact of these technologies on users and society. Focusing on the

ethical implications of AI-driven health care solutions, from patient engagement strategies to public health campaigns on social media, this research direction aims to ensure that they positively contribute to user well-being and societal health standards. Fifth, a strategic approach should be identified to evaluate the impact of AI-assisted interactions within health care and social media from a FATE perspective. This includes analyzing these interactions to develop universal, model-agnostic metrics that assess the ethical dimensions of AI applications across various platforms. Once established, such metrics could be integrated into social networks, guiding the regulation of AI use in health care on SMPs. For medical professionals and researchers, these metrics would provide a framework for consistently evaluating and ensuring the ethical integrity of AI technologies, promoting safer and more beneficial health care interactions on social media.

Limitations

The primary limitation of our study stems from the scarcity of comprehensive research that thoroughly explores all dimensions of FATE in the context of AI applications in health care on SMPs. This scarcity reflects not only existing research gaps but also the early stage of scholarly inquiry in this interdisciplinary area. Consequently, our review may not fully encapsulate the complex and multidimensional nature of how FATE intersect and manifest in the deployment of AI within health care settings on social media. This limitation is significant because it suggests that our understanding of FATE issues in this context may rely on an incomplete picture, thus impacting the generalizability of our findings across all potential AI applications in health care on social media.

In addition, identifying the precise population of studies relevant to FATE in AI and health care on SMPs is made more challenging by the heterogeneity and dynamism of SMPs as well as the diversity of AI applications within health care. SMPs are rapidly evolving, introducing new functionalities and altering user interactions, which in turn influences how AI technologies can be applied and examined within these contexts. The challenge of compiling a representative collection of studies that fully encompasses this range contributes to potential gaps in our review, limiting the degree to which our findings can be seen as representative of the field as a whole.

Moreover, the fast-paced advancement of technology, along with the continual evolution of both SMPs and AI, imposes a temporal limitation on our study. Research that was up-to-date at the time of our review may soon become outdated as new technologies emerge and existing ones advance. This swift pace of change implies that the ethical challenges identified today may evolve, new challenges may surface, and previously proposed solutions may become obsolete or less applicable. Therefore, the applicability of our findings is inherently limited by this temporal aspect, underscoring the necessity for ongoing research to continuously refresh our understanding of FATE within AI in health care on SMPs.

Conclusions

Our review sheds light on the current state of FATE in health care AI as applied to SMPs. It offers a critical analysis of the methodologies, computational techniques, and evaluative strategies used in various studies. By examining the successes and identifying the shortcomings of current practices, this review stimulates further innovation in the field. It challenges existing paradigms on how AI technologies can be both technologically advanced and ethically robust, ensuring fairness, accountability, and transparency in their application.

The practical implications of this work are substantial. First, it guides future research by identifying recent trends and research gaps, suggesting that researchers focus on creating more robust, fair, and ethical AI systems. This includes using diverse data sets that more accurately represent the global population and using evaluation metrics that comprehensively assess the systems' impacts on all stakeholders. Second, this review underscores the importance of integrating FATE principles throughout the AI system development life cycle, from conceptualization to deployment. For practitioners in health care and technology, this signifies a move toward more inclusive, transparent, and ethically guided development processes. Such a transition not only addresses biases and accountability issues but also boosts patient trust and engagement with AI-driven health care solutions on social media.

Third, the insights from this review are invaluable for policy makers and regulatory bodies, aiding in the creation of nuanced regulations and guidelines that ensure that AI technologies positively contribute to health care outcomes without compromising ethical standards or patient rights. Furthermore, by simplifying complex concepts, this review acts as an educational tool for a broad audience, including health care providers, AI developers, patients, and the general public. Raising awareness about the importance of FATE in health care AI fosters more informed participation in discussions and decision-making regarding AI use in health care, particularly on SMPs.

Ultimately, this study aids in the pursuit of ethical development and deployment of AI systems in health care. By providing an in-depth analysis of the current achievements and future directions for FATE in health care AI on social media, it advocates for the adoption of best practices that balance ethical considerations with technological innovations. The implications of this study extend beyond academia, affecting how AI technologies are conceptualized, developed, and implemented in health care on social media, thereby shaping a future where AI-driven health care solutions are not only effective and innovative but also ethically responsible, equitable, and transparent. This ensures that these technologies serve the best interests of society.

Data Availability

All data generated or analyzed during this study are included in this published paper and its supplementary information files.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Studies selected for the review.
[DOCX File , 29 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Fairness evaluation metrics with mathematical formulation.
[DOCX File , 27 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Accountability evaluation metrics with mathematical formulation.
[DOCX File , 20 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Transparency evaluation metrics with mathematical formulation.
[DOCX File , 20 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Ethics evaluation metrics with mathematical formulation.
[DOCX File , 18 KB-Multimedia Appendix 5]

References

1. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv. Jul 13, 2021;54(6):1-35. [doi: 10.1145/3457607]
2. Mashhadi A, Winder SG, Lia EH, Wood SA. No walk in the park: the viability and fairness of social media analysis for parks and recreational policy making. Proc Int AAAI Conf Web Soc Media. May 22, 2021;15(1):409-420. [doi: 10.1609/icwsm.v15i1.18071]
3. Leonelli S, Lovell R, Wheeler BW, Fleming L, Williams H. From FAIR data to fair data use: methodological data fairness in health-related social media research. Big Data Soc. May 03, 2021;8(1). [doi: 10.1177/20539517211010310]
4. Singhal A, Baxi MK, Mago V. Synergy between public and private health care organizations during COVID-19 on Twitter: sentiment and engagement analysis using forecasting models. JMIR Med Inform. Aug 18, 2022;10(8):e37829. [FREE Full text] [doi: 10.2196/37829] [Medline: 35849795]
5. Kington RS, Arnesen S, Chou WY, Curry SJ, Lazer D, Villarruel AM. Identifying credible sources of health information in social media: principles and attributes. NAM Perspect. 2021;2021:10.31478/202107a. [FREE Full text] [doi: 10.31478/202107a] [Medline: 34611600]
6. Pershad Y, Hangge PT, Albadawi H, Oklu R. Social medicine: Twitter in healthcare. J Clin Med. May 28, 2018;7(6):121. [FREE Full text] [doi: 10.3390/jcm7060121] [Medline: 29843360]
7. Flores L, Young SD. Ethical considerations in the application of artificial intelligence to monitor social media for COVID-19 data. Minds Mach (Dordr). 2022;32(4):759-768. [FREE Full text] [doi: 10.1007/s11023-022-09610-0] [Medline: 36042870]
8. Pirraglia PA, Kravitz RL. Social media: new opportunities, new ethical concerns. J Gen Intern Med. Feb 8, 2013;28(2):165-166. [FREE Full text] [doi: 10.1007/s11606-012-2288-x] [Medline: 23225258]
9. Wieringa M. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020. Presented at: FAT* '20; January 27-30, 2020; Barcelona, Spain. URL: https://dl.acm.org/doi/abs/10.1145/3351095.3372833 [doi: 10.1145/3351095.3372833]
10. Hutchinson B, Smart A, Hanna A, Denton E, Greer C, Kjartansson O, et al. Towards accountability for machine learning datasets: practices from software engineering and infrastructure. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021. Presented at: FAccT '21; March 3-10, 2021; Virtual event, Canada. [doi: 10.1145/3442188.3445918]
11. Johnson SL. AI, machine learning, and ethics in health care. J Leg Med. 2019;39(4):427-441. [doi: 10.1080/01947648.2019.1690604] [Medline: 31940250]

12. Solomonides AE, Koski E, Atabaki SM, Weinberg S, McGreevey JD, Kannry JL, et al. Defining AMIA's artificial intelligence principles. J Am Med Inform Assoc. Mar 15, 2022;29(4):585-591. [FREE Full text] [doi: 10.1093/jamia/ocac006] [Medline: 35190824]
13. The Belmont report. U.S. Department of Health and Human Services. URL: https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html [accessed 2023-12-05]
14. Shin D. The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. Int J Hum Comput Stud. Feb 2021;146:102551. [doi: 10.1016/j.ijhcs.2020.102551]
15. Hagerty A, Rubinov I. Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. arXiv. Preprint posted online July 18, 2019. [FREE Full text] [doi: 10.48550/arXiv.1907.07892]
16. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. In: Proceedings of the IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). 2018. Presented at: DSAA 2018; October 1-3, 2018; Turin, Italy. [doi: 10.1109/dsaa.2018.00018]
17. Chakraborty S, Tomsett R, Raghavendra R, Harborne D, Alzantot M, Cerutti F, et al. Interpretability of deep learning models: a survey of results. In: Proceedings of the 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation. 2017. Presented at: SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017; August 4-8, 2017; San Francisco, CA. [doi: 10.1109/uic-atc.2017.8397411]
18. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: a survey on methods and metrics. Electronics. Jul 26, 2019;8(8):832. [doi: 10.3390/electronics8080832]
19. Golder S, Ahmed S, Norman G, Booth A. Attitudes toward the ethics of research using social media: a systematic review. J Med Internet Res. Jun 06, 2017;19(6):e195. [FREE Full text] [doi: 10.2196/jmir.7082] [Medline: 28588006]
20. Bear Don't Walk OJ4, Reyes Nieva H, Lee SS, Elhadad N. A scoping review of ethics considerations in clinical natural language processing. JAMIA Open. Jul 2022;5(2):ooac039. [FREE Full text] [doi: 10.1093/jamiaopen/ooac039] [Medline: 35663112]
21. Attard-Frost B, De los Ríos A, Walters DR. The ethics of AI business practices: a review of 47 AI ethics guidelines. AI Ethics. Apr 13, 2022;3(2):389-406. [doi: 10.1007/s43681-022-00156-6]
22. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138-52160. [doi: 10.1109/access.2018.2870052]
23. Vian T, Kohler JC. Medicines Transparency Alliance (MeTA): pathways to transparency, accountability, and access: cross-case analysis and review of phase II. World Health Organization. May 25, 2016. URL: https://tinyurl.com/3vhjyysd [accessed 2023-12-05]
24. Kofod-Petersen A. How to do a structured literature review in computer science. Norwegian University of Science and Technology. 2018. URL: https://research.idi.ntnu.no/aimasters/files/SLR_HowTo2018.pdf [accessed 2024-03-13]
25. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. Oct 02, 2018;169(7):467-473. [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
26. Saha D, Schumann C, Mcelfresh DC, Dickerson JP, Mazurek ML, Tschantz MC. Measuring non-expert comprehension of machine learning fairness metrics. In: Proceedings of the 37th International Conference on Machine Learning. 2020. Presented at: PMLR 2020; July 13-18, 2020; Virtual event. URL: https://proceedings.mlr.press/v119/saha20c.html [doi: 10.1145/3375627.3375819]
27. Mehrabi N, Gupta U, Morstatter F, Steeg GV, Galstyan A. Attributing fair decisions with attention interventions. In: Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022). 2022. Presented at: TrustNLP 2022; July 14, 2022; Seattle, WA. [doi: 10.18653/v1/2022.trustnlp-1.2]
28. Hertweck C, Heitz C, Loi M. On the moral justification of statistical parity. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021. Presented at: FAccT '21; March 3-10, 2021; Virtual event. [doi: 10.1145/3442188.3445936]
29. Yao H, Chen Y, Ye Q, Jin X, Ren X. Refining language models with compositional explanations. arXiv. Preprint posted online March 18, 2021. [FREE Full text]
30. Markoulidakis I, Kopsiaftis G, Rallis I, Georgoulas I. Multi-Class Confusion Matrix Reduction method and its application on Net Promoter Score classification problem. In: Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference. 2021. Presented at: PETRA '21; June 29-July 2, 2021; Corfu, Greece. [doi: 10.1145/3453892.3461323]
31. Vergeer P, van Schaik Y, Sjerps M. Measuring calibration of likelihood-ratio systems: a comparison of four metrics, including a new metric devPAV. Forensic Sci Int. Apr 2021;321:110722. [doi: 10.1016/j.forsciint.2021.110722] [Medline: 33684845]
32. Lagioia F, Rovatti R, Sartor G. Algorithmic fairness through group parities? The case of COMPAS-SAPMOC. AI Soc. Apr 28, 2022;38(2):459-478. [doi: 10.1007/s00146-022-01441-y]

33. Saxena NA, Huang K, DeFilippis E, Radanovic G, Parkes DC, Liu Y. How do fairness definitions fare?: examining public attitudes towards algorithmic definitions of fairness. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 2019. Presented at: AIES '19; January 27-28, 2019; Honolulu, HI. [doi: 10.1145/3306618.3314248]
34. Park Y, Hu J, Singh M, Sylla I, Dankwa-Mullan I, Koski E, et al. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. JAMA Netw Open. Apr 01, 2021;4(4):e213909. [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.3909] [Medline: 33856478]
35. Xu J, Xiao Y, Wang WH, Ning Y, Shenkman EA, Bian J, et al. Algorithmic fairness in computational medicine. EBioMedicine. Oct 2022;84:104250. [FREE Full text] [doi: 10.1016/j.ebiom.2022.104250] [Medline: 36084616]
36. Tao G, Sun W, Han T, Fang C, Zhang X. RULER: discriminative and iterative adversarial training for deep neural network fairness. In: Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2022. Presented at: ESEC/FSE '22; November 14-18, 2022; Singapore. [doi: 10.1145/3540250.3549169]
37. Chouldechova A, Roth A. A snapshot of the frontiers of fairness in machine learning. Commun ACM. Apr 20, 2020;63(5):82-89. [doi: 10.1145/3376898]
38. Yao S, Huang B. Beyond parity: fairness objectives for collaborative filtering. arXiv. Preprint posted online March 24, 2017. [FREE Full text]
39. Zhang Y, Zhou L. Fairness assessment for artificial intelligence in financial industry. arXiv. Preprint posted online December 16, 2019. [FREE Full text] [doi: 10.5260/chara.21.2.8]
40. Ghassami A, Khodadadian S, Kiyavash N. Fairness in supervised learning: an information theoretic approach. arXiv. Preprint posted online January 13, 2018. [FREE Full text] [doi: 10.1109/isit.2018.8437807]
41. Malawski M. A note on equal treatment and symmetry of values. In: Nguyen NT, Kowalczyk R, Mercik J, Motylska-Kuźma A, editors. Transactions on Computational Collective Intelligence XXXV. Berlin, Heidelberg. Springer; 2020.
42. Saleiro P, Kuester B, Hinkson L, London J, Stevens A, Anisfeld A, et al. Aequitas: a bias and fairness audit toolkit. arXiv. Preprint posted online November 14, 2018. [FREE Full text] [doi: 10.48550/arXiv.1811.05577]
43. Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev. Jul 1, 2019;63(4/5):4:1-:15. [doi: 10.1147/jrd.2019.2942287]
44. Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim HC, et al. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. Biol Psychiatry Cogn Neurosci Neuroimaging. Sep 2021;6(9):856-864. [FREE Full text] [doi: 10.1016/j.bpsc.2021.02.001] [Medline: 33571718]
45. Bird S, Dudík M, Edgar R, Horn B, Lutz R, Milan V, et al. Fairlearn: a toolkit for assessing and improving fairness in AI. Microsoft. Sep 22, 2020. URL: https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf [accessed 2023-12-05]
46. Ghosh A, Genuit L, Reagan M. Characterizing intersectional group fairness with worst-case comparisons. arXiv. Preprint posted online January 05, 2021. [FREE Full text]
47. Zafar MB, Valera I, Gomez-Rodriguez M, Gummadi KP. Fairness constraints: a flexible approach for fair classification. J Mach Learn Res. 2019;20(75):1-42.
48. Chakraborti T, Patra A, Noble JA. Contrastive fairness in machine learning. IEEE Lett Comput Soc. Jul 7, 2020;3(2):38-41. [doi: 10.1109/locs.2020.3007845]
49. Rosenfeld A, Richardson A. Explainability in human–agent systems. Auton Agent Multi-Agent Syst. May 13, 2019;33:673-705. [doi: 10.1007/s10458-019-09408-y]
50. Narasimhan H, Cotter A, Gupta M, Wang S. Pairwise fairness for ranking and regression. Proc AAAI Conf Artif Intell. Apr 03, 2020;34(04):5248-5255. [doi: 10.1609/aaai.v34i04.5970]
51. Kaur D, Uslu S, Durresi A, Badve S, Dundar M. Trustworthy explainability acceptance: a new metric to measure the trustworthiness of interpretable ai medical diagnostic systems. In: Proceedings of the 15th International Conference on Complex, Intelligent and Software Intensive Systems. 2021. Presented at: CISIS-2021; July 1-3, 2021; Asan, Korea. [doi: 10.1007/978-3-030-79725-6_4]
52. Bucher M, Herbin S, Jurie F. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In: Proceedings of the Computer Vision – ECCV 2016. 2016. Presented at: ECCV 2016; October 11-14, 2016; Amsterdam, The Netherlands. [doi: 10.1007/978-3-319-46454-1_44]
53. Kynkäänniemi T, Karras T, Laine S, Lehtinen J, Aila T. Improved precision and recall metric for assessing generative models. arXiv. Preprint posted online April 15, 2019. [FREE Full text]
54. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. arXiv. Preprint posted online August 13, 2020. [FREE Full text] [doi: 10.48550/arXiv.2008.05756]
55. Zaki MM, Jena AB, Chandra A. Supporting value-based health care - aligning financial and legal accountability. N Engl J Med. Sep 09, 2021;385(11):965-967. [doi: 10.1056/NEJMp2105625] [Medline: 34478249]
56. Blacklaws C. Algorithms: transparency and accountability. Philos Trans A Math Phys Eng Sci. Sep 13, 2018;376(2128):20170351. [doi: 10.1098/rsta.2017.0351] [Medline: 30082299]
57. Kim B, Park J, Suh J. Transparency and accountability in AI decision support: explaining and visualizing convolutional neural networks for text information. Decis Support Syst. Jul 2020;134:113302. [doi: 10.1016/j.dss.2020.113302]

58. Dubberley S, Murray D, Koenig A. Digital Witness: Using Open Source Information for Human Rights Investigation, Documentation, and Accountability. Oxford, UK. Oxford University Press; 2020.
59. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019. Presented at: FAT* '19; January 29-31, 2019; Atlanta, GA. [doi: 10.1145/3287560.3287596]
60. King J. The instrumental value of legal accountability. In: Accountability in the Contemporary Constitution. Oxford, UK. Oxford University Press; 2013.
61. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell. Nov 04, 2019;1:501-507. [doi: 10.1038/s42256-019-0114-4]
62. Kass NE, Faden RR. Ethics and learning health care: the essential roles of engagement, transparency, and accountability. Learn Health Syst. Sep 18, 2018;2(4):e10066. [FREE Full text] [doi: 10.1002/lrh2.10066] [Medline: 31245590]
63. Wright D. A framework for the ethical impact assessment of information technology. Ethics Inf Technol. Jul 8, 2010;13:199-226. [doi: 10.1007/s10676-010-9242-6]
64. Arnold T, Kasenberg D, Scheutz M. Value alignment or misalignment – what will keep systems accountable? Association for the Advancement of Artificial Intelligence. 2017. URL: https://hrilab.tufts.edu/publications/arnoldetal17aiethics.pdf [accessed 2023-12-05]
65. Iyer R, Li Y, Li H, Lewis M, Sundar R, Sycara K. Transparency and explanation in deep reinforcement learning neural networks. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018. Presented at: AIES '18; February 2-3, 2018; New Orleans, LA. [doi: 10.1145/3278721.3278776]
66. Fukuda‐Parr S, Gibbons E. Emerging consensus on 'ethical AI': human rights critique of stakeholder guidelines. Glob Policy. Jun 19, 2021;12(S6):32-44. [doi: 10.1111/1758-5899.12965]
67. Wachter S, Mittelstadt B, Floridi L. Transparent, explainable, and accountable AI for robotics. Sci Robot. May 31, 2017;2(6):eaan6080. [doi: 10.1126/scirobotics.aan6080] [Medline: 33157874]
68. Ozga J. The politics of accountability. J Educ Change. 2020;21:19-35. [doi: 10.1007/s10833-019-09354-2]
69. Ko RK, Kirchberg M, Lee BS. From system-centric to data-centric logging - accountability, trust and security in cloud computing. In: Proceedings of the Defense Science Research Conference and Expo. 2011. Presented at: DSR 2011; August 3-5, 2011; Singapore. [doi: 10.1109/dsr.2011.6026885]
70. Raji I, Smart A, White R, Mitchell M, Gebru T, Hutchinson B, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020. Presented at: FAT* '20; January 27-30, 2020; Barcelona, Spain. [doi: 10.1145/3351095.3372873]
71. Nushi B, Kamar E, Horvitz E. Towards accountable AI: hybrid human-machine analyses for characterizing system failure. Proc AAAI Conf Hum Comput Crowdsourc. Jun 15, 2018;6(1):126-135. [doi: 10.1609/hcomp.v6i1.13337]
72. Vesnic-Alujevic L, Nascimento S, Pólvora A. Societal and ethical impacts of artificial intelligence: critical notes on European policy frameworks. Telecommun Policy. Jul 2020;44(6):101961. [doi: 10.1016/j.telpol.2020.101961]
73. Kerikmäe T, Pärn-Lee E. Legal dilemmas of Estonian artificial intelligence strategy: in between of e-society and global race. AI Soc. Jul 01, 2020;36:561-572. [doi: 10.1007/s00146-020-01009-8]
74. Reich MR. The core roles of transparency and accountability in the governance of global health public-private partnerships. Health Syst Reform. 2018;4(3):239-248. [doi: 10.1080/23288604.2018.1465880] [Medline: 30207904]
75. Conway M, O'Connor D. Social media, big data, and mental health: current advances and ethical implications. Curr Opin Psychol. Jun 2016;9:77-82. [FREE Full text] [doi: 10.1016/j.copsyc.2016.01.004] [Medline: 27042689]
76. Laacke S, Mueller R, Schomerus G, Salloch S. Artificial intelligence, social media and depression. A new concept of health-related digital autonomy. Am J Bioeth. Jul 2021;21(7):4-20. [doi: 10.1080/15265161.2020.1863515] [Medline: 33393864]
77. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform. Oct 2013;46(5):830-836. [FREE Full text] [doi: 10.1016/j.jbi.2013.06.010] [Medline: 23820016]
78. Crawley AW, Divi N, Smolinski MS. Using timeliness metrics to track progress and identify gaps in disease surveillance. Health Secur. 2021;19(3):309-317. [doi: 10.1089/hs.2020.0139] [Medline: 33891487]
79. Zhai C, Cohen WW, Lafferty J. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. ACM SIGIR Forum. Jun 23, 2015;49(1):2-9. [doi: 10.1145/2795403.2795405]
80. Weiss DJ, Nelson A, Gibson HS, Temperley W, Peedell S, Lieber A, et al. A global map of travel time to cities to assess inequalities in accessibility in 2015. Nature. Jan 18, 2018;553(7688):333-336. [FREE Full text] [doi: 10.1038/nature25181] [Medline: 29320477]
81. Burgess K, Hart D, Elsayed A, Cerny T, Bures M, Tisnovsky P. Visualizing architectural evolution via provenance tracking: a systematic review. In: Proceedings of the Conference on Research in Adaptive and Convergent Systems. 2022. Presented at: RACS '22; October 3-6, 2022; Virtual event. [doi: 10.1145/3538641.3561493]
82. Diakopoulos N, Koliska M. Algorithmic transparency in the news media. Digit J. Jul 27, 2016;5(7):809-828. [doi: 10.1080/21670811.2016.1208053]

83. Stellefson M, Paige SR, Chaney BH, Chaney JD. Evolving role of social media in health promotion: updated responsibilities for health education specialists. Int J Environ Res Public Health. Feb 12, 2020;17(4):1153. [FREE Full text] [doi: 10.3390/ijerph17041153] [Medline: 32059561]
84. Valko M, Hauskrecht M. Feature importance analysis for patient management decisions. Stud Health Technol Inform. 2010;160(Pt 2):861-865. [FREE Full text] [Medline: 20841808]
85. Lipton ZC. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue. Jun 01, 2018;16(3):31-57. [doi: 10.1145/3236386.3241340]
86. Slack D, Friedler SA, Scheidegger C, Dutta Roy C. Assessing the local interpretability of machine learning models. arXiv. Preprint posted online February 9, 2019. [FREE Full text]
87. Stepin I, Alonso JM, Catala A, Pereira-Farina M. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access. Jan 13, 2021;9:11974-12001. [doi: 10.1109/access.2021.3051315]
88. Bertino E, Merrill S, Nesen A, Utz C. Redefining data transparency: a multidimensional approach. Computer. Jan 2019;52(1):16-26. [doi: 10.1109/MC.2018.2890190]
89. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med. Jan 2019;25(1):30-36. [FREE Full text] [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
90. Li Q. Overview of data visualization. In: Embodying Data. Singapore. Springer; Jun 20, 2020.
91. Azeroual O, Saake G, Schallehn E. Analyzing data quality issues in research information systems via data profiling. Int J Inf Manage. Aug 2018;41:50-56. [doi: 10.1016/j.ijinfomgt.2018.02.007]
92. Tang M, Shao S, Yang W, Liang Y, Yu Y, Saha B, et al. SAC: a system for big data lineage tracking. In: Proceedings of the IEEE 35th International Conference on Data Engineering (ICDE). 2019. Presented at: ICDE 2019; April 8-11, 2019; Macao, China. [doi: 10.1109/icde.2019.00215]
93. Leslie D. Understanding artificial intelligence ethics and safety. arXiv. Preprint posted online June 11, 2019. [FREE Full text]
94. Shneiderman B. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Trans Interact Intell Syst. Oct 16, 2020;10(4):1-31. [doi: 10.1145/3419764]
95. Brundage M, Avin S, Wang J, Belfield H, Krueger D, Hadfield G, et al. Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv. Preprint posted online April 15, 2020. [FREE Full text] [doi: 10.48550/ARXIV.2004.07213]
96. Janssen M, Hartog M, Matheus R, Yi Ding A, Kuk G. Will algorithms blind people? The effect of explainable AI and decision-makers' experience on AI-supported decision-making in government. Soc Sci Comput Rev. Dec 28, 2020;40(2):478-493. [doi: 10.1177/0894439320980118]
97. Paredes JN, Teze JC, Martinez MV, Simari GI. The HEIC application framework for implementing XAI-based socio-technical systems. Online Soc Netw Media. Nov 2022;32:100239. [doi: 10.1016/j.osnem.2022.100239]
98. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. Nov 30, 2020;20(1):310. [FREE Full text] [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
99. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. Jun 2020;58:82-115. [doi: 10.1016/j.inffus.2019.12.012]
100. Xiong Z, Cui Y, Liu Z, Zhao Y, Hu M, Hu J. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. Comput Materials Sci. Jan 2020;171:109203. [doi: 10.1016/j.commatsci.2019.109203]
101. Zafar MR, Khan N. Deterministic local interpretable model-agnostic explanations for stable explainability. Mach Learn Knowl Extr. Jun 30, 2021;3(3):525-541. [doi: 10.3390/make3030027]
102. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv. Preprint posted online May 2, 2017. [FREE Full text]
103. Sokol K, Flach P. Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety. In: Proceedings of the AAAI Workshop on Artificial Intelligence Safety, SafeAI 2019. 2019. Presented at: SafeAI 2019; January 27, 2019; Honolulu, HI.
104. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst. Nov 2021;32(11):4793-4813. [doi: 10.1109/TNNLS.2020.3027314] [Medline: 33079674]
105. Vig J. A multiscale visualization of attention in the transformer model. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2019. Presented at: ACL 2019; July 28-August 2, 2019; Florence, Italy. [doi: 10.18653/v1/p19-3007]
106. Fan CY, Chang PC, Lin JJ, Hsieh JC. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Appl Soft Comput. Jan 2011;11(1):632-644. [doi: 10.1016/j.asoc.2009.12.023]
107. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci U S A. Oct 29, 2019;116(44):22071-22080. [FREE Full text] [doi: 10.1073/pnas.1900654116] [Medline: 31619572]

108. Leikas J, Koivisto R, Gotcheva N. Ethical framework for designing autonomous intelligent systems. J Open Innov Technol Mark Complex. Mar 2019;5(1):18. [doi: 10.3390/joitmc5010018]
109. Latonero M. Governing artificial intelligence: upholding human rights and dignity. Data & Society. 2018. URL: https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf [accessed 2023-12-05]
110. Aiello AE, Renson A, Zivich PN. Social media- and internet-based disease surveillance for public health. Annu Rev Public Health. Apr 02, 2020;41:101-118. [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094402] [Medline: 31905322]
111. Olteanu A, Castillo C, Diaz F, Kıcıman E. Social data: biases, methodological pitfalls, and ethical boundaries. Front Big Data. 2019;2:13. [FREE Full text] [doi: 10.3389/fdata.2019.00013] [Medline: 33693336]
112. Dixon L, Li J, Sorensen J, Thain N, Vasserman L. Measuring and mitigating unintended bias in text classification. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018. Presented at: AIES '18; February 2-3, 2018; New Orleans, LA. [doi: 10.1145/3278721.3278729]
113. Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015. Presented at: KDD '15; August 10-13, 2015; Sydney, Australia. [doi: 10.1145/2783258.2783311]
114. Mendes R, Cunha M, Vilela JP, Beresford AR. Enhancing user privacy in mobile devices through prediction of privacy preferences. In: Proceedings of the 27th European Symposium on Research in Computer Security. 2022. Presented at: ESORICS 2022; September 26-30, 2022; Copenhagen, Denmark. [doi: 10.1007/978-3-031-17140-6_8]
115. Datta A, Sen S, Zick Y. Algorithmic transparency via quantitative input influence: theory and experiments with learning systems. In: Proceedings of the IEEE Symposium on Security and Privacy (SP). 2016. Presented at: SP 2016; May 22-26, 2016; San Jose, CA. [doi: 10.1109/sp.2016.42]
116. Kazim E, Koshiyama AS. A high-level overview of AI ethics. Patterns (N Y). Sep 10, 2021;2(9):100314. [FREE Full text] [doi: 10.1016/j.patter.2021.100314] [Medline: 34553166]
117. Nebeker C, Parrish EM, Graham S. The AI-powered digital health sector: ethical and regulatory considerations when developing digital mental health tools for the older adult demographic. In: Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues. Cham, Switzerland. Springer; 2022;159-176.
118. Crockett MJ. Models of morality. Trends Cogn Sci. Aug 2013;17(8):363-366. [FREE Full text] [doi: 10.1016/j.tics.2013.06.005] [Medline: 23845564]
119. Colman AM. Game Theory and its Applications: In the Social and Biological Sciences. London, UK. Psychology Press; 1995.
120. Someh IA, Davern M, Breidbach C, Shanks G. Ethical issues in big data analytics: a stakeholder perspective. Commun Assoc Inf Syst. May 2019;44(34):718-747. [FREE Full text] [doi: 10.17705/1CAIS.04434]
121. Ventola CL. Social media and health care professionals: benefits, risks, and best practices. P T. Jul 2014;39(7):491-520. [FREE Full text] [Medline: 25083128]
122. Ponce SB, M Barry M, S Dizon D, S Katz M, Murphy M, Teplinsky E, et al. Netiquette for social media engagement for oncology professionals. Future Oncol. Mar 2022;18(9):1133-1141. [FREE Full text] [doi: 10.2217/fon-2021-1366] [Medline: 35109663]
123. Drabiak K, Wolfson J. What should health care organizations do to reduce billing fraud and abuse? AMA J Ethics. Mar 01, 2020;22(3):E221-E231. [FREE Full text] [doi: 10.1001/amajethics.2020.221] [Medline: 32220269]
124. Neville P, Waylen A. Social media and dentistry: some reflections on e-professionalism. Br Dent J. Apr 24, 2015;218(8):475-478. [doi: 10.1038/sj.bdj.2015.294] [Medline: 25908363]
125. Ennis-O-Connor M, Mannion R. Social media networks and leadership ethics in healthcare. Healthc Manage Forum. May 2020;33(3):145-148. [doi: 10.1177/0840470419893773] [Medline: 31884833]
126. Garg T, Shrigiriwar A. Managing expectations: how to navigate legal and ethical boundaries in the era of social media. Clin Imaging. Apr 2021;72:175-177. [doi: 10.1016/j.clinimag.2020.11.005] [Medline: 33296827]
127. Kalkman S, Mostert M, Gerlinger C, van Delden JJ, van Thiel GJ. Responsible data sharing in international health research: a systematic review of principles and norms. BMC Med Ethics. Mar 28, 2019;20(1):21. [FREE Full text] [doi: 10.1186/s12910-019-0359-9] [Medline: 30922290]
128. Sharma S. Data Privacy and GDPR Handbook. Hoboken, NJ. John Wiley & Sons; 2019.
129. Leidner JL, Plachouras V. Ethical by design: ethics best practices for natural language processing. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. 2017. Presented at: EthNLP@EACL; April 4, 2017; Valencia, Spain. [doi: 10.18653/v1/w17-1604]
130. Guttman N. Ethical issues in health promotion and communication interventions. Oxford Research Encyclopedias Communication. Feb 27, 2017. URL: https://www.academia.edu/100398082/Ethical_Issues_in_Health_Promotion_and_Communication_Interventions [accessed 2023-12-05]
131. Denecke K, Bamidis P, Bond C, Gabarron E, Househ M, Lau AY, et al. Ethical issues of social media usage in healthcare. Yearb Med Inform. Aug 13, 2015;10(1):137-147. [FREE Full text] [doi: 10.15265/IY-2015-001] [Medline: 26293861]
132. Gagnon K, Sabus C. Professionalism in a digital age: opportunities and considerations for using social media in health care. Phys Ther. Mar 2015;95(3):406-414. [doi: 10.2522/ptj.20130227] [Medline: 24903111]

133. Bhatia-Lin A, Boon-Dooley A, Roberts MK, Pronai C, Fisher D, Parker L, et al. Ethical and regulatory considerations for using social media platforms to locate and track research participants. Am J Bioeth. Jun 2019;19(6):47-61. [FREE Full text] [doi: 10.1080/15265161.2019.1602176] [Medline: 31135323]
134. Davis K, Patterson D. Ethics of Big Data. Sebastopol, CA. O'Reilly Media; Sep 2012.
135. Livingston JD, Milne T, Fang ML, Amari E. The effectiveness of interventions for reducing stigma related to substance use disorders: a systematic review. Addiction. Jan 2012;107(1):39-50. [FREE Full text] [doi: 10.1111/j.1360-0443.2011.03601.x] [Medline: 21815959]
136. Jakesch M, Buçinca Z, Amershi S, Olteanu A. How different groups prioritize ethical values for responsible AI. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022. Presented at: FAccT '22; June 21-24, 2022; Seoul, Republic of Korea. [doi: 10.1145/3531146.3533097]
137. Pastaltzidis I, Dimitriou N, Quezada-Tavarez K, Aidinlis S, Marquenie T, Gurzawska A, et al. Data augmentation for fairness-aware machine learning: preventing algorithmic bias in law enforcement systems. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022. Presented at: FAccT '22; June 21-24, 2022; Seoul, Republic of Korea. [doi: 10.1145/3531146.3534644]
138. Keshk M, Moustafa N, Sitnikova E, Turnbull B. Privacy-preserving big data analytics for cyber-physical systems. Wireless Netw. Dec 20, 2018;28(3):1241-1249. [doi: 10.1007/s11276-018-01912-5]
139. Kayaalp M. Patient privacy in the era of big data. Balkan Med J. Jan 20, 2018;35(1):8-17. [FREE Full text] [doi: 10.4274/balkanmedj.2017.0966] [Medline: 28903886]
140. Enarsson T, Enqvist L, Naarttijärvi M. Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts. Inf Commun Technol Law. Jul 27, 2021;31(1):123-153. [doi: 10.1080/13600834.2021.1958860]
141. Umbrello S, van de Poel I. Mapping value sensitive design onto AI for social good principles. AI Ethics. Feb 01, 2021;1(3):283-296. [FREE Full text] [doi: 10.1007/s43681-021-00038-3] [Medline: 34790942]
142. Hossin M, Sulaiman MN, Mustapha A, Mustapha N, Rahmat RW. A hybrid evaluation metric for optimizing classifier. In: Proceedings of the 3rd Conference on Data Mining and Optimization (DMO). 2011. Presented at: DMO 2011; June 28-29, 2011; Putrajaya, Malaysia. [doi: 10.1109/dmo.2011.5976522]
143. Nguyen AT, Raff E, Nicholas C, Holt J. Leveraging uncertainty for improved static malware detection under extreme false positive constraints. arXiv. Preprint posted online August 9, 2021. [FREE Full text] [doi: 10.48550/arXiv.2108.04081]
144. Jadon S. A survey of loss functions for semantic segmentation. In: Proceedings of the IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology. 2020. Presented at: CIBCB 2020; October 27-29, 2020; Via del Mar, Chile. [doi: 10.1109/cibcb48159.2020.9277638]
145. Hansen E. HIPAA (Health Insurance Portability and Accountability Act) rules: federal and state enforcement. Med Interface. Aug 1997;10(8):96-8, 101. [Medline: 10169779]
146. Grajales FJ3, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G. Social media: a review and tutorial of applications in medicine and health care. J Med Internet Res. Feb 11, 2014;16(2):e13. [FREE Full text] [doi: 10.2196/jmir.2912] [Medline: 24518354]
147. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res. May 12, 2021;23(5):e17917. [FREE Full text] [doi: 10.2196/17917] [Medline: 33978589]

Abbreviations

AI: artificial intelligence
AMIA: American Medical Informatics Association
FATE: fairness, accountability, transparency, and ethics
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
LIME: local interpretable model-agnostic explanations
ML: machine learning
NLP: natural language processing
PDP: partial dependence plot
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RQ: research question
SHAP: Shapley additive explanations
SMP: social media platform

Edited by A Castonguay; submitted 18.06.23; peer-reviewed by G Randhawa, D Valdes, M Arab-Zozani; comments to author 28.10.23; revised version received 21.12.23; accepted 15.02.24; published 03.04.24

Please cite as:
Singhal A, Neveditsin N, Tanveer H, Mago V
Toward Fairness, Accountability, Transparency, and Ethics in AI for Social Media and Health Care: Scoping Review
JMIR Med Inform 2024;12:e50048
URL: https://medinform.jmir.org/2024/1/e50048
doi: 10.2196/50048
PMID: 38568737