AI and Ethics (2023) 3:223–240
https://doi.org/10.1007/s43681-022-00195-z

ORIGINAL RESEARCH

Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers

Pravik Solanki[1] · **John Grundy**[2] · **Waqar Hussain**[2]

Received: 1 January 2022 / Accepted: 27 June 2022 / Published online: 19 July 2022

Abstract
Artificial intelligence (AI) offers much promise for improving healthcare. However, it runs the looming risk of causing individual and societal harms; for instance, exacerbating inequalities amongst minority groups, or enabling compromises in the confidentiality of patients' sensitive data. As such, there is an expanding, unmet need for ensuring AI for healthcare is developed in concordance with human values and ethics. Augmenting "principle-based" guidance that highlight adherence to ethical ideals (without necessarily offering translation into actionable practices), we offer a solution-based framework for operationalising ethics in AI for healthcare. Our framework is built from a scoping review of existing solutions of ethical AI guidelines, frameworks and technical solutions to address human values such as self-direction in healthcare. Our view spans the entire length of the AI lifecycle: data management, model development, deployment and monitoring. Our focus in this paper is to collate actionable solutions (whether technical or non-technical in nature), which can be steps that enable and empower developers in their daily practice to ensuring ethical practices in the broader picture. Our framework is intended to be adopted by AI developers, with recommendations that are accessible and driven by the existing literature. We endorse the recognised need for 'ethical AI checklists' co-designed with health AI practitioners, which could further operationalise the technical solutions we have collated. Since the risks to health and wellbeing are so large, we believe a proactive approach is necessary for ensuring human values and ethics are appropriately respected in AI for healthcare.

Keywords Artificial intelligence · Machine learning · Healthcare · Medicine · Ethics · Human values

1 Introduction

Although the exponential growth of Artificial Intelligence (AI) for healthcare is promising, its benefits are being increasingly overshadowed by its propensity to cause individual or societal harm quickly at a large scale [1, 2]. AI for healthcare encompasses numerous approaches, including machine learning (ML) algorithms on structured text- or image-based data, and natural language processing (NLP) on unstructured data such as clinical notes or medical journals [3]. These approaches are being applied to the prediction (e.g. predicting the presence of type 2 diabetes from clinical risk factors, [4] or the risk of suicide from social media posts [5]), detection (e.g. detecting breast cancer tumours from mammography scans [6]), and management of diseases (e.g. using NLP-driven chatbots to deliver cognitive behavioural therapy [7]).

This progress has endowed AI with significant hype [8], despite many ethical challenges threatening health and human rights remaining unaddressed [9]. A prominent example of this is how melanoma-detecting AI algorithms are presently trained largely on images of white skin, making them inaccurate at detecting melanoma in darker-skinned people (despite the fact that melanoma is more lethal in African populations) [10]. If ethical issues such as these are not promptly addressed, we risk an 'AI winter' taking place, whereby public trust and the potential benefits of AI for healthcare could be swiftly lost [6]. Regulatory policy to address these issues at the legal level [11] and governance frameworks to address these issues at the organisational level [12] are slowly emerging, but there remain few practical recommendations that developers and users of AI for healthcare can utilise throughout the AI lifecycle [13].

✉ Pravik Solanki
Pravik.solanki@gmail.com

[1] Monash University (Faculty of Medicine), Melbourne, VIC, Australia

[2] Monash University (Faculty of Information Technology), Melbourne, VIC, Australia

Existing ethical AI guidelines have two issues: firstly, very few are specific to healthcare [14], despite the fact that AI for healthcare involves unique ethical issues [2]; and secondly, they emphasise adherence to 'ethical principles' [15] without complementary translations into actionable practices [16, 17]. As such, there remains a pressing need to operationalise ethics throughout the development pipeline of AI for healthcare [18].

To address this gap, we first provide an overview of the unique ethical issues that arise when human values are compromised in AI for healthcare. We then propose a framework for operationalising ethics, grounded in existing guidelines that provide actionable solutions. To ensure our framework is accessible and utilisable, we organise it in accordance to the AI lifecycle. In going through each stage of the development pipeline, we outline implementable recommendations that adopt, but also go beyond an abstract consideration of 'ethical principles' [19]. Our framework is of direct relevance to those developing AI for healthcare (including software developers and data scientists, who we collectively refer to as 'developers'), and those utilising AI for healthcare (including clinicians and health informaticians, who we collectively refer to as 'users'). Although we focus on healthcare, our framework may also be useful to those developing and utilising AI in other domains. Important facilitators of our framework include adherence to governance models [20] and the creation of best-practice ethics checklists [2, 21], both of which are emerging as AI for healthcare continues to develop.

2 Methodology

To contextualise and assemble our framework, we performed two literature reviews. Firstly, we performed a scoping review on the range of ethical issues that may arise in AI for healthcare. This involved searching PubMed and Web of Science on literature at the intersection of AI, healthcare, and ethics. Titles and abstracts of the first 200 articles from each repository were assessed, and relevant articles were reviewed in full.

We then used two conceptual frameworks to organise the ethical issues identified. First, we use Schwartz' theory of basic human values, an empirically-validated framework comprised of human values that are assumed to be universal across all cultures [22]. Of the ten total human values in this theory, we reference only the four found to be most cited in software engineering literature [23]. Acknowledging the broad nature of these human values, we further subcategorise them into specific, granular ethical principles, as outlined in a recent scoping review of AI ethics publications [14]. The one-to-many mapping of human values to ethical principles is arbitrary, and is used only as a foundation to present an organised overview of ethical issues identified in the AI literature.

Second, we performed a scoping review of existing frameworks, guidelines and recommendations pertaining to ethical AI for healthcare. We chose Scopus and Google Scholar to identify relevant articles, searching for literature at the intersection of AI, healthcare, and existing guidelines for ethical AI. Noting that the publication of generic ethical AI guidelines has increased exponentially over recent years [14], we focussed on scholarship at the intersection of ethical AI and healthcare wherever possible. We assessed the first 200 articles identified by Scopus and Google Scholar, then adopted a forward and backward snowballing approach to identify papers offering actionable solutions for operationalising ethics throughout the AI lifecycle. Drawing upon existing literature to separate the AI lifecycle into distinct stages [2, 24, 25], we then present the actionable recommendations that can be operationalised at each stage of the pipeline by developers and/or users to ensure ethical AI for healthcare.

3 Key human values and ethical issues in AI for healthcare

A consideration of human values is largely lacking in software engineering [23], which is used for the majority of AI applications [26]. This is despite the fact that purposefully aligning AI with human values can produce many benefits, such as improving cancer care and patient engagement [27]. Conversely, if human values are compromised, a multitude of ethical issues can arise [8], which we outline extensively in the following subsections. Here we provide a taxonomy of human values from the social sciences and review the ethical issues corresponding to these values that arise in AI for healthcare.

Schwartz' Theory of Basic Human Values describes ten human values validated by empirical research conducted across over 70 countries [22]. A literature review found four values to be the most frequently cited in 1,350 recently published software engineering publications: security, benevolence, universalism, and self-direction [23]. Independently, a scoping review conducted a thematic analysis of 84 AI ethics guidelines [14], identifying 11 key ethical principles referenced across guidelines. Table 1 presents these human values [22], together with the corresponding ethical principles associated with them [14]. The mapping between human values and ethical principles is arbitrary, and is used only to outline and conceptualise the many ethical issues arising in AI for healthcare.

Table 1 Human values and corresponding ethical principles

Human value [22]	Ethical principle [14]
Security (safety, harmony, and stability of society, of relationships, and of self)	Non-maleficence (protection from harm, precaution, prevention, non-subversion)
Self-direction (independence in thought and action; creating, exploring, being curious)	Freedom and autonomy (consent, choice, self-determination, liberty, empowerment)
	Dignity
	Privacy (protection of personal or private information)
Benevolence (preserving and enhancing others' welfare, voluntary concern for others' welfare)	Beneficence (benefits, well-being, peace, social good, common good)
	Responsibility (accountability, liability, acting with integrity)
	Trust
	Transparency (explainability, understandability, interpretability, acts of communication and disclosure)
	Solidarity (social security and cohesion)
Universalism (understanding, appreciation, tolerance, and protection for the welfare of all people and for nature)	Justice and fairness (consistency, inclusion, equality, equity, non-discrimination, respect for diversity, plurality, accessibility, redress)
	Sustainability (conserving environment and natural resources)

Descriptions of human values and ethical principles adapted from [22] and [14] respectively

3.1 Security

Security encapsulates feelings of safety and stability of oneself, one's relations, and society at large, comprising the ethical principle of non-maleficence.

3.1.1 Non-maleficence

Non-maleficence involves minimising foreseeable harm in terms of discrimination, violation of privacy, and bodily harm, and is cited considerably more than beneficence in current AI guidelines, suggesting it is a greater priority for AI to avoid harm than to do good [14]. Since AI for healthcare is evolving so rapidly, there is a concern that harms will only be recognised and addressed after they have occurred [28].

Safety is a key priority in AI for healthcare, especially since relatively few initiatives are backed by empirical evidence [2]. Technical failures, such as AI chatbots that cease to function properly [29], or AI initiatives that fail during a network failure [30], may result in unintended harm. In addition, AI may also lack interpersonal or cultural competency, which could hinder therapeutic relationships and cause unintended psychological distress [29].

3.2 Self-direction

Self-direction encapsulates a sense of personal independence, comprising the ethical principles of freedom and autonomy; dignity; and privacy.

3.2.1 Freedom and autonomy

Freedom and autonomy refer to the preservation of self-determination, which includes informed consent and the right to withdraw consent [14]. Informed consent involves the disclosure of relevant information, the individual being competent and fully comprehending this information, and the individual voluntarily accepting participation [31]. The process of informed consent also involves clarifying concerns or misconceptions, including the possibility of third parties accessing confidential data [32].

Obtaining consent may be difficult for black box algorithms that are too opaque to be fully understood by humans [33], and may not be practically feasible for social media or other large datasets encompassing millions of individuals [34]. In terms of AI-based mobile health apps, users may assume as high an ethical commitment to confidentiality as in professional healthcare, even though this is often not the case [35]. Although the terms and conditions of health apps may be presented, these are seldom read or comprehended by users [31].

3.2.2 Dignity

Dignity refers to the preservation of human decency and rights [14]. One consideration is the dignity of developers, who may have no training in healthcare and could find emotional content traumatic [34]. Other considerations relate to how individuals might form therapeutic relationships with AI. Possible ethical issues with this include mentally ill individuals wrongly believing they are interacting with a human or other force; individuals

feeling upset if a humanoid robot invokes a creepy sense of repulsion (termed the "uncanny valley"); and individuals not having an easy way to safely end the therapeutic relationship [30].

3.2.3 Privacy

Privacy is a human right, involving individuals' information being carefully protected and securely managed [9]. Information relating to individuals' health can be highly sensitive [35], and clinicians therefore have an ethical obligation to maintain confidentiality. To what degree this obligation to maintain confidentiality extends to AI in healthcare—which often require access to large amounts of sensitive health data to be effective—remains a matter of debate [36]. The right to privacy is recognised to be in ongoing tension with the open advancement of AI for healthcare [31], with key issues relating to data collection, data management, and working with individuals' social media data.

Data collection can entail numerous ethical issues. Data may be collected by multiple sources (e.g. smartphone geolocation and online forum activity) [31], by AI robots [29], or by passive means (e.g. screen taps or voice inflection) [35]. A key issue is whether individuals feel comfortable about their data being collected in each of these cases, including when they are unaware that it is taking place [31].

Data management involves additional considerations. Privacy can be threatened through poor security practices, such as leaving a laptop with sensitive data unattended in public. Privacy can also be threatened through hacking by non-authorised parties [30]. Regarding social media and mental health, users generally expect their privacy to be respected online [37], with most feeling their data should not be used for mental health research without their explicit consent [34]. Social media data may be problematic on numerous counts: it may be an inaccurate representation of an individuals' mental state (given one's self-portrayal on social media can differ markedly from offline behaviours); it may include other users who have been 'tagged' (who themselves have a right to privacy); and it may be difficult to anonymise [32, 34]. Finally, if a user decides to 'drop out' from social media after their data has been collected, their health data may still be stored [34].

3.3 Benevolence

Benevolence encapsulates a sense of enhancing and maintaining 'good' for oneself and others, comprising the ethical principles of beneficence; responsibility; trust; transparency; and solidarity.

3.3.1 Beneficence

Beneficence involves contributing to individual and societal wellbeing [14]. Although AI has the potential to do much good, key ethical issues relate to the known limitations of AI, and the impact of AI on clinicians' decision-making.

The known limitations of AI should be recognised. Whilst a clinician may be able to monitor societal risks as per their professional code (e.g. the risk of domestic violence or child abuse), AI initiatives built for other purposes could miss these signs entirely [30]. Additionally, those who are high-risk, such as those with severe depression, may need more comprehensive treatment than AI initiatives alone [35].

AI-driven decisions could also have unforeseen impacts on clinicians' decision-making. For instance, if a patients' high-risk genetic mutation is known to radiologists, the number of missed breast lesions on MRI scans is known to decrease considerably [38]. AI-driven predictions could therefore sway clinicians' own assessments of risk. If only patients deemed high-risk by the AI are offered further screening or treatment (without appropriate input from clinicians), this has the potential to become a self-fulfilling prophecy [39].

3.3.2 Responsibility and trust

Responsibility involves attributing accountability and liability, as well as acting with transparency and integrity in a way that builds trust [14]. For developers, clearly stating the limitations of their work is a key consideration [37]; for instance, most suicide risk predictions do not predict when an individual may attempt suicide, and hence whether involuntary restraint is warranted [40].

The lines of responsibility between developers, implementers and AI are not clearly defined, particularly with 'black box' algorithms that cannot be easily understood [34]. For instance, if a suicide occurs and was not detected by the AI algorithm, is it the fault of the algorithm, the developer, the clinician, or the manufacturer? This question remains unresolved [40]. Other areas of ambiguity are how conflicts between an AI-driven decision and a clinician's impression should be resolved, such as if an AI algorithm detects an individual to be at high risk, but the clinician disagrees [40]. Although clinicians have professional standards (e.g. a duty of care) to which they are held accountable, AI have no in-built responsibility to maintain these standards [35], and have no ability to experience the moral consequences of poor decisions (e.g. emotional distress) [30].

Trust is built through a transparent culture amongst developers, implementers, and patients, ensuring that practices fulfil public expectations [14]. Trust may be lost when AI have many false positive or negative results [41], are perceived to be incompetent [30], or make use of public data in

a disrespectful way [37]. A breakdown of trust is not only harmful for AI initiatives, but could also be harmful for clinicians and healthcare as a whole [30].

3.3.3 Transparency

Transparency involves a sense of interpretability or explainability of AI-based decisions [14]. Interpretability can be understood as 'how' a model arrives at a decision, whereas explainability can be understood as 'why' a model arrives at a decision [41]. Key issues relate to AI having low interpretability and/or explainability, and shortcomings of AI not being fully disclosed.

Although some AI algorithms are more interpretable (e.g. regression models) compared to others, empirical findings suggest that uninterpretable algorithms (e.g. deep learning algorithms developed from millions of data points) tend to perform better [31]. The issue with 'black boxes' algorithms is that humans can understand the inputs and outputs, but cannot clearly understand the process connecting the two [34]. In many cases, such algorithms also lack explainability, which poses issues; for instance, an individual may be told they are at high risk of developing an illness for reasons that remain undeterminable from the AI model. Individuals remain divided about whether they would like to be told their AI-derived high risk status of an illness if it is unexplainable, given this can be distressing [31]. Additionally, if implementers cannot understand models in the first place, they may be unable to discover sources of bias or challenge AI-driven decisions [31].

Other issues of transparency relate to disclosure of an AI intervention's shortcomings. This includes the rate of false positive or false negative results (including for specific groups) [2], changes to model performance over time [31], and the presence of bias [41]. The amount of information to be disclosed to implementers or patients about AI algorithms, and how this should best be done, remains contested [14].

3.3.4 Solidarity

Solidarity refers to the special consideration of vulnerable populations and those of low socioeconomic status, including the potential need to redistribute benefits of AI to these groups [14]. For instance, NLP algorithms may be developed only in the English language, such that they cannot be applied to cultural groups using other languages [32].

The implementation of AI for healthcare can cause harm to vulnerable populations. For instance, some individuals could find the notion of AI (and the tracking of their behaviour) distressing, such as those with schizophrenia fearing mass surveillance [34]. Moreover, AI could be used for ulterior motives, such as health insurance companies using AI to identify high-risk individuals whose premiums they wish to raise [42]. Finally, although AI can provide some form of healthcare in resource-poor areas, this could be used as a justification to not further develop physical (as opposed to virtual) mental health services in these areas of need [29].

3.4 Universalism

Universalism encapsulates a sense of appreciation towards people and the planet, comprising the ethical principles of justice and fairness, and sustainability.

3.4.1 Justice and fairness

Justice and fairness refer to representing the full diversity of society (rather than the privileged few) while safeguarding against discrimination towards vulnerable groups, and upholding individuals' right to challenge AI-based decisions [14]. If people of different gender, ethnic and other sociodemographic backgrounds are not represented in the research, design and development of AI, these interventions could implicitly ignore the needs of these groups [2]. For instance, gaps in training data could stem from the lack of non-binary gender identities in electronic health records, or from undocumented migrants with low access to healthcare [40]. Moreover, training data may also reflect systemic biases based on gender, race, and other sociodemographic characteristics—for instance, the disproportionate number of African–Americans who suffer from schizophrenia [33]. If such data is used for predictions, these algorithms could simply exacerbate existing disparities [40, 41]. Economic factors (e.g. different billing rates for specific groups) could also influence what and how information is collected in healthcare [41]. Finally, social media users may be more likely to be young and Caucasian compared to the overall population, meaning the health of other groups could be ignored [34].

Since AI-driven decisions are not absolute truths, another potential issue is whether implementers can challenge questionable outputs. This is particularly important in high-risk situations; for instance, suicide prediction algorithms' false positives could result in an individual being wrongly detained by a health service [43], whilst false negatives could mean an otherwise preventable suicide is missed [40]. Due to these possibilities, delegating decision-making to "machines alone" has been criticised [44].

3.4.2 Sustainability

Sustainability involves considering the environment and minimising the ecological footprint of AI initiatives [14]. This is of importance for all AI initiatives, with relevant factors specific to healthcare not yet identified.

4 A framework for operationalising ethics in AI for healthcare

Although there has been significant work in mapping the ethical principles in AI for healthcare, this understanding alone does not translate readily into actionable practices [16, 19]. To address this gap, here we present our framework for operationalising ethics in AI for healthcare across the development pipeline, as illustrated in Fig. 1. Firstly, ethical AI is a product of numerous layers of influence: the professional practices of developers and users; the governance of these individuals at an organisational level; and the regulation of individuals and organisations at a legal level. Secondly, the AI lifecycle involves three major stages, which must progressively be fulfilled before the subsequent stage is initiated. These stages are as follows [2, 24, 25]:

1. *Data management* involves: (A) data being collected; (B) data being appropriately protected with best-security practices; (C) data being cleaned (including pre-processing and augmentation where appropriate); and (D) data being reported.
2. *Model development* involves: (A) an AI model being trained on a dataset; (B) an AI model being verified in its performance on test dataset/s; and (C) an AI model being reported.
3. *Deployment and monitoring* involves: (A) a model being deployment in a real-life setting with stakeholder engagement and user-centered design; (B) updates and ongoing validation; and (C) supervision and auditing.

We present our operationalised ethical AI framework across these three stages, providing actionable solutions to prevent, mitigate and address ethical issues. Throughout our work, we identify guidelines that provide actionable recommendations on specific ethical needs (e.g. protecting classifiers from adversarial attacks). All of these guidelines are further explained in our framework, and have been summarised in Table 2 for the reader's convenience (see Supplemental Material for unabridged version). We end by highlighting how these solutions cannot simply be actioned in a vacuum, but must be supported by rapidly-evolving governance and regulatory frameworks.

4.1 Data management

4.1.1 Data collection

Data can be collected via numerous methods. People are more likely to report behavioural symptoms to digital agents than humans, but sociocultural sensitivity can only be provided by a human agent; hence, care should be taken to ensure methods of data collection are appropriate [41], and the collected data itself is more diverse and inclusive [46–49]. Moreover, the variables and collected data should be justified by having demonstrated pertinence to healthcare; unnecessary variables should not be collected [18]. This includes avoiding datasets known to be imbalanced or biased [50–53] and data types that will not be used, such as photographs of people when training text-based

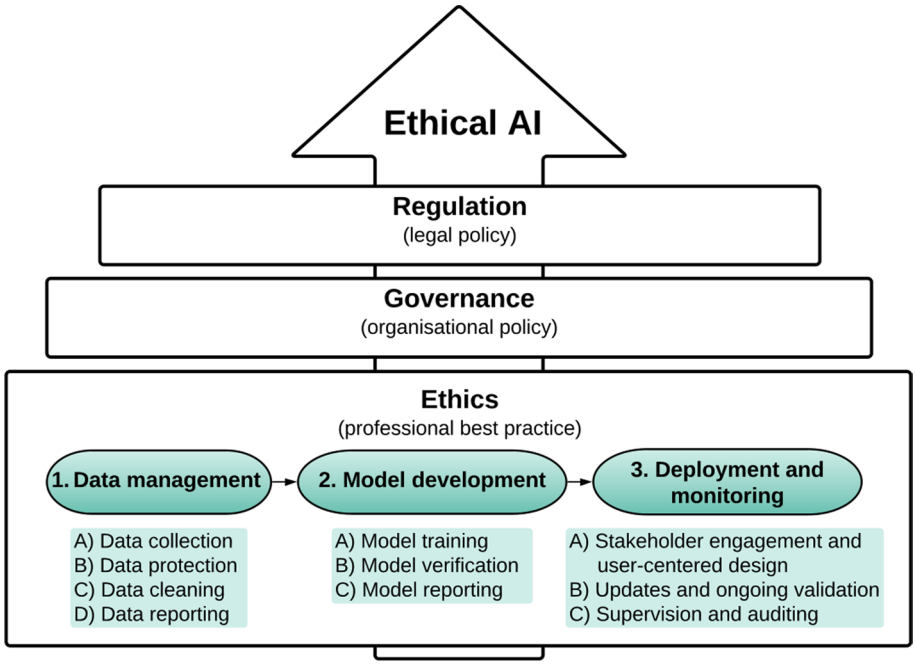

Fig. 1 Ethical AI framework

Table 2 Guidelines addressing specific needs in AI for healthcare and the ethical principles they address

ML/AI main components		Solution / Guideline	Ethical Principles
Data	**Collection & handling**	1. *Avoid using common yet biased 'benchmark' datasets [45]*	
		2. *Use diverse datasets [46-49]*	
		3. *Voluntarily increase diversity of imbalanced datasets [50-53]*	
		4. *(Synthetically) create diverse datasets for training and accessing models [54, 55]*	
		5. *Increase team diversity and AI literacy of AI, ML and data science teams [56-62]*	
		6. *Educate teams about AI benefits and issues e.g. biased data usage, thus build AI-literate workforce [50-53, 62, 63]*	
		7. *Raise team awareness about ethical AI challenges e.g. using common yet biased datasets for AI systems [45, 63, 64]*	
		8. *Improve data label quality to achieve improved unbiasedness and model accuracy [45, 65, 66]*	
		9. *Understand fairness and its contextual challenges to make better accuracy vs fairness trade-offs [55, 64, 67]*	
		10. *Use data pre-processing techniques to address data induced unfairness rather than over constraining the model [68, 69]*	
	Protection (privacy/ security)	1. *De-identify personal information within datasets by Anonymizing (removing private data from a record) [51][70-72]*	
		2. *De-identify personal information Pseudonymizing (replacing sensitive entries with artificially generated ones) [73-75]*	
		3. *Use Differential Privacy to protect unwanted personal information exposure [51][74, 76]*	
		4. *Generate (new privacy preserving) synthetic data [50][77, 78]*	
		5. *Use tools/frameworks that combine various privacy techniques for data privacy protection [79-81]*	
		6. *Ensure data integrity and privacy in health data systems [82]*	
		7. *Use privacy protection tools that also allow transparency of the applied privacy process [78]*	
		8. *Avoid loss of data (critical to the owner of the information) [72, 83]*	
		9. *Apply privacy preserving 'temper-free' frameworks and hybrid models that ensure privacy and safety [84, 85]*	
		10. *Guard against (adversarial) security attacks using state of the art techniques [86, 87]*	
	Data cleaning	1. *Perform data debiasing e.g. improve labelling to reduce subgroup representational imbalance [45, 65]*	
		2. *Tackle systematic inequities for under-represented socioeconomic groups to allow classifier debiasing [45, 88, 89]*	
		3. *Apply 'bias transforming' metrics to achieve substantive/true fairness and tackle inequalities [90]*	
		4. *Use fairness evaluation metrics to identify gaps in aspired vs achieved fairness of AI systems [90, 91]*	
		5. *Apply fairness evaluation metrics to quantify concepts like Heterogeneity, Diversity and Inclusion [90]*	
		6. *Prevent discrimination through data manipulation techniques e.g., Reweighing, re/sampling etc. [92-95]*	
		7. *Apply discrimination prevention techniques on data to address direct and indirect discrimination [91, 96, 97]*	
		8. *Identify and select data cleaning tools, models or framework suitable to your context, domain or needs [98-100]*	
		9. *Use data standardization to improve data understandability and interoperability [62]*	
		10. *Improve data quality by removing noise within data and placing data integrity constraints [65, 99-101]*	
	Data reporting	1. *Report data characteristics using Datasheets to enhance transparency and highlight societal biases [102, 103]*	
		2. *Use data characteristics to encourage transparency and data usage aligned with deployment context [102-104]*	
		3. *Treat datasets as infrastructure and develop documents similar to those used in software engineering [105, 106]*	
		4. *Apply privacy-preserving algorithms when publicising e-health data [74, 107]*	
		5. *Share AI services information using appropriate templates for transparency and public awareness [108, 109]*	
		6. *Use dataset statements to report biases in data and enable transparent model development [110-112]*	

algorithms [32]. Although social media data may be easy to collect, the privacy of users should be acknowledged and respected [152]. There remains no consensus of the threshold of health data that requires individual consent for use, but in some cases, deidentification (including of name, date of birth, address, and/or health card number) may be deemed satisfactory for bypassing this step [153]. Since training AI on imbalanced datasets data can fuel inequalities, a special effort should be made to gain data from understudied, underserved or vulnerable populations [2, 154]. Applying data pre-processing techniques [68, 69] to improve the quality of data (regarding diversity and inclusion) often proves more effective than dealing with unfairness resulting from feeding biased data into models for decision making and predictions [54, 55].

Technical efforts should be complemented by more holistic, humanistic initiatives. This involves raising practitioners' awareness of the contextual nature and demands of particular ethical principles such as fairness, and actively building ethically-aware and diverse teams to address potential systemic biases [45, 56–64].

Table 2 (continued)

Model	**Model training**	1. *Train models using ethical storytelling and reward outcomes that align with 'trained' cultural values [113, 114]*	
		2. *Train models using ethics-aware scenarios concerning fairness, bias/discrimination and social justice [115]*	
		3. *Decompose model bias* into 'bias-variance-noise' to identify and separate sources of discrimination [64, 68]	
		4. *Add ethics-inspired regularization to penalize unfair decisions* made by models under training [116, 117]	
		11. Apply human-centered approaches and keep Explainability to (potentially impacted) humans in mind [118, 119]	
		12. Keep humans in the loop to achieve ethical model training and desired ethical outcomes [120, 121]	
		13. *Adhere to interpretability based on regulatory and efficiency demands and users' needs* [118, 119, 122-125]	
		14. *Apply hybrid approaches to enhance interpretability* [126, 127]	
		15. Apply causal reasoning to identify and understand and discrimination based on protected attributes *[128]*	
		16. *Implement both mathematical and ethically grounded fairness definitions* [67, 129]	
		17. *Make models and classifier safe from adversarial attacks* using game theory and GANs [130-133]	
		18. *Secure models through the use of adversary-aware learning algorithms* [134-136]	
		19. Apply a combination of classifiers or multiple classifiers to guard against attacks on models [137, 138]	
	Model verification	1. *Avoid unethical corrections to models* that sacrifice fairness for accuracy [68, 139]	
		2. *Use post-hoc techniques and auditing to identify under-preforming groups* and address racial bias [140-143]	
	Model reporting	20. *Encourage and ensure transparency by reporting* model performance, intended use cases etc.[123, 144]	
		21. Use model reporting frameworks tools to respond to transparency needs of user and other stakeholders [123, 144]	
Deployment and monitoring	**Stakeholder engagement and UCD**	22. *Ensure appropriate human-AI interaction* [51]	
		23. Follow Human – AI guidelines to guide users understanding of the system and invoke the desired response [145]	
		24. Keep users' needs of freedom of choice, transparency and explainability in mind [146, 147]	
	Supervision and auditing	25. *Ensure ethical governance* through ongoing internal audits to avoid potential negative societal impact [148, 149]	
		26. Perform ethics-based auditing to compare aspired vs practiced ethics and take corrective measures [150, 151]	

Key for Ethical principles: Justice & Fairness, Privacy (including Freedom & Autonomy), Security & Non-maleficence, Inclusion, Human Centricity and Responsibility, Transparency, Trust & Standardization.

4.1.2 Data protection

There is always a relatively high risk of adversarial attacks on health data systems [155], and hence, data should be stored, handled and used with best security practices e.g. the use of de-identification, anonymization [51, 70–72] and other approaches such as Differential Privacy to to procted data from malicious attacks and maintain and data integrity [50, 73–75, 77, 78, 82].

A number of practical tools, techniques and frameworks have been developed to used for procting data used in ML [79–81, 84, 85] some of which allow a degree of transparency to the applied privacy processes [78]. One such framework is detailed in [84] for building data security into health information systems. While there is an inherent trade-off between the integrity of data and maintenance of privacy, a verification system for ensuring that both integrity and privacy are preserved in health data systems is detailed in [156]. As evidenced by the numerous attacks and security breaches commonly reported in the media, many malicious actors continue to develop a variety of adversarial attacks on sensitive data, posing ongoing threats to data security and privacy measures [51, 74, 76]. This requires practitioners to ensure data privacy and security measures are up to date, using state-of-the-art defence techniques such as DP [86, 87] to prevent unnecessary data exposure and preserve data integrity [82].

Datasets should also be scrutinised for the possibility of 'poisoning attacks' that falsely skew the data. Practical methods for assessing this are detailed in [85] for quantitative datasets and in [87] for image-based datasets. If a deep neural network is planned, steps to protect classifiers from poisoning attacks are as detailed in [132].

'Sanitising' datasets (i.e. creating an identical dataset with identifying details purposely altered) has been shown not to guarantee individual privacy [157]. However, a novel 'DataSynthesizer' tool for creating synthetic databases ensures strong privacy features, and is further outlined in [78]. There are innovative hybrid models specifically designed to support healthcare big data protection (e.g. [84]), that support data security with techniques like masking encryption, activity monitoring, granular access control, and end point validation. Other methods involving k-anonymity and l-diversity have been shown to have numerous limitations in protecting the privacy of individuals; hence, a novel t-closeness requirement, further detailed in [72], is recommended instead.

Data protection (and the protection process transparency) also entails clearly defining those allowed to access the data, with access levels layered for different agents (users, administrators, etc.) [93], and a log populated that shows who has accessed the data, when they accessed the data, and what actions they have taken [93]. In addition, data sharing practices should be transparently disclosed [18]; data should never be shared with private companies without explicit consent [153]. If data is to be shared between parties, one method to do so securely is detailed in [158]. An alternative approach is to share synthetic data that mimic original data with differential privacy; an open-source tool that be utilised to achieve this is further detailed in [73].

4.1.3 Data cleaning

When assessing a dataset, vulnerable groups should be explicitly identified and should each have sufficient completeness of data [159]. In some cases, simply avoiding the use of biased or imbalanced datasets (e.g. those used in *Fitzpatrick Skin Type Classification System* [45]) can help address potential social injustice resulting from the AI system developed. Methods to automatically identify the appropriate data cleaning activities for an incomplete or biased dataset are still emerging, but could be achieved with the use of frameworks such as *MLClean* (that removes data poisoning to help training accurate and fair models [100]) and tools such as like *HoloClean, Activeclean* [98, 99] and *Universal Cleanser* [160]. A number of techniques that can be employed to prevent discrimination during data cleaning are outlined in [161]. If a classification algorithm is planned, specific pre-processing methods can be applied to prevent discrimination, including minimally intrusive data modifications that lead to an unbiased dataset [92, 96] with additional techniques involving selective instances relabelling ('massaging'), suppression and reweighing [96, 162]. If data from vulnerable groups remains inadequate despite a concerted effort, numerous techniques may be employed such as synthetically creating diverse datasets [54, 55] using techniques such as convex optimization [95] to address the inadequacy of available data. missing data and training the models. Other techniques, including resampling (oversampling and under sampling [163]) and removing poisoned samples [101], can also be applied to tackle data imbalance [93–95].

4.1.4 Data reporting

All datasets should have accompanying documentation to enhance transparency and help ML engineers identify and understand issues in training data [104, 154, 164]. For ML, documentation should encompass the following areas: motivation, composition, collection process, pre-processing, uses, distribution, maintenance, impact and challenges [102]. For NLP, documentation should encompass the following areas: curation rationale, language variety, speaker demographic, annotator demographic, speech situation, text characteristics, and recording quality [110]. The full documentation processes are outlined in [45, 164] for ML datasets, and in [110, 111] for NLP data. Similarly, to engender trust in AI, systems can be accompanied by AI service *FactSheets* shared by service providers that declare the purpose, performance, safety, security of the AI service, i.e. something that can be examined and audited by AI service consumers and regulators [108]. One such *FactSheet* template is provided by [109], describing the diversity of contexts and circumstances in which AI systems are developed and deployed.

4.2 Model development

4.2.1 Model training

Before a model is trained, a pre-analysis plan should be written, clearly stipulating the goals of the model, technical approaches to be taken (e.g. the type of loss function), and the research question/s and outcome/s of interest [2]. Training and test datasets should be clearly separated, such that entries do not occur in both groups simultaneously [28].

Where feasible, ethical or value-based constraints should be set up as guiding criteria beyond the usual cost and efficiency optimisation focussed on during model development to ensure models are developed in a responsible manner [165]. Models can then be 'rewarded' and reinforced when outputs aligned with predefined ethical or regulatory constraints. A practical example of this process is outlined in [165]. This can also be achieved by 'raising' intelligent systems using storytelling that communicates tacit knowledge and cultural values, e.g. using (deep and inverse) reinforcement learning to 'punish' agents when their actions are misaligned with our values or otherwise undesirable to humans [113–116]. Model training can also be nudged towards ethics using training them on ethical datasets such as ETHICS and presenting the model with contrasting ethical and unethical scenarios to enhance its ability to analyse ethical expectations and behaviours [115]. Where feasible, predictor variables that are actionable should be preferentially selected over those that are not [41]. Chosen outcome variables should be clinically relevant [28], unbiased for marginalised groups [2], and could include 'soft' endpoints such as patient satisfaction [147]. Specific criteria that can be used to ensure the model is unbiased are specified in [123]. For classification algorithms that use stream-based data, a method to ensure algorithmic fairness is outlined in [166]. Although models that are less interpretable tend to have higher accuracy [123], interpretable models are generally preferred, as this aids transparency and trust [9, 167].

Model training could be made more ethical by moving away from predominantly algorithm-centred approaches and towards human-centred approaches, e.g. keeping 'humans in the loop' and also by selecting models with the most probability of satisfying users' explainability goals [119]. The principle of explainability can be made more achievable by preferring inherently interpretable models to train for applications that make high-stakes decisions impacting humans instead of using black box models known to cause problems e.g. in healthcare, criminal justice systems and other domains [122]. Creating awareness about the level of explainability in ML models among engineers can go a long way to increase the chances of building explainable AI systems. ML engineers working on potentially explainable health care AI systems can benefit from a comprehensive classification of ML models, regarding their degree of explainability and other related techniques, as presented in [123]. These principles can be further operationalized to target *explainable medicine* by developing explainable interfaces that allowing experts to understand machine diagnosis/outcome and its influencing factors or causality [124, 125]. For a particular project context or specific parameter, It may be possible to convert black box models to 'glass box' explainable models, as argued in [122]. The notion of fairness can be highly contextual, and therefore challenging to satisfy from both a group and individual perspective. For quantitative simplicity, it is the mathematical definitions of fairness that are predominantly implemented, e.g. similar percentages of false positives and/or false negatives for the different socioeconomic groups under consideration [67]. However, depending on the context, these mathematical implementations of fairness could be complemented by more ethically grounded ones [67, 129]

Models should also be trained to minimise the effect of adversarial attacks or of poisoned data samples. Evasion attacks can be protected against by devising systems composed of multiple classifiers, as outlined in [137], or other adversary-aware learning algorithms (e.g. one-and-a-half-class multiple classifiers) as outlined in [135]. A 'hardness of evasion' score for measuring a model's sturdiness against adversarial attacks is detailed in [138]. Another aspect that protects against adversarial attacks is the 'resilience' of a model (i.e. perturbations in input causing minimal changes to output), which can be increased by smoothing ML decision boundaries [168]. Models may be trained only considering the subset of features that cannot be manipulated by an attacker, an approach further outlined in [136]. For training sturdy models using a Generative Adversarial Network approach, two models may be trained on similar datasets as outlined in [131]; alternatively, a 'MinMax game' technique for adversarial regularization may be taken, as outlined in [130]. After a model has been trained, poisoned samples in training data may be removed via data sanitization processes outlined in [101], or via 'machine unlearning' to remove the effects of poisoned data as outlined in [142]. To plan against an attack, the equilibrium between an attacker and the system can be identified using zero-sum game theory, as detailed in [134].

4.2.2 Model verification

The performance metrics of an AI algorithm should be carefully chosen depending on its purpose, including whether it is more important to have few false positives or few false negatives [28]. Importantly, the performance of an AI algorithm must not only be considered on an overall scale; the performance on marginalised groups should be specifically evaluated [2, 41]. Methods to slice data and identify subsets of the data where the model has poor performance are detailed in [141].

Post-hoc classifier auditing can also be performed periodically to discover under-performing subgroups, thus iteratively improving model performance for subgroups of concern and improving fairness [143]. Practices to address fairness need to be accompanied by the team's realistic acknowledgment of the inherent challenges of fairness, including contextual trade-offs to balance fairness with accuracy. Some biases may even be correctible with post-hoc techniques involving re-labelling, such as for racial bias [140]. Biases may be balanced with variance of performance estimates using out-of-sample bootstrap techniques [139]. Importantly, post-hoc model correction methods (e.g. severity scoring in clinical tasks based on randomizing predictions) must be carefully considered and should be justifiable from an ethical perspective [68]. While retraining models is possible and often necessary, other 'machine unlearning' techniques can also be applied to remove the effects of poisoned data without having to retrain the model [142].

Where possible, the AI algorithm should be tested across multiple healthcare systems, socioeconomic groups, or age ranges [28, 169]; if the proposed scale of application is particularly large, proportionally greater scrutiny is warranted [140, 170].

4.2.3 Model reporting

When reporting the final, verified model, the data source, participants, predictors and outcome variables should be clearly reported, as should the features of the model (e.g. regression coefficients) and the specific environment/s in which it has been verified [28]. Such comprehensive disclosure of model-related information significantly improves the transparency of the model and thus trust in the built technology. Clear documentation with 'model cards' is recommended for summarising a model's attributes and performance; this is further detailed in [144]. These cards clarify

model's intended use cases, and disclose conditions, context and intended application domains as well as the scenarios that are suitable or unsuitable for trained ML models [144]. The stakeholders are likely to appreciate model-related documentation when it is aligned with their explainability needs. Identifying the target audience and the reasons why they need these explanations can boost their trust in AI-based systems, significantly more than handling black box models to them without any explanations of how these models come up with the decisions [123]. Besides enhancing transparency, documentation detailing benchmarked evaluation (e.g. across different cultural, demographic, or phenotypic groups) can encourage reuse for practitioners trying to implement these models in their context [144]. Transparency can also be promoted by ad-hoc and external post-hoc techniques that make models more interpretable, some of which are further explained in [123].

4.3 Deployment and monitoring

4.3.1 Stakeholder engagement and user-centered design

To ensure needs are being met, stakeholders should be consulted extensively throughout the AI lifecycle, and particularly during deployment [171]. Relevant stakeholders include knowledge experts (clinicians, ML researchers, health informaticians, implementations experts), decision-makers (hospital administrators, institutional leadership, regulatory agencies, and government bodies), and users (clinicians, laboratory technicians, patients and their family members) [28, 147]. To assist in thoughtful design, efforts should be made for the team developing and implementing AI to feature sociocultural diversity [159], including an appropriate representation of women [172], ensuring that the needs of these groups are not implicitly ignored by the AI initiative. Crucially, there should be a clearly identified clinical problem that is to be tackled [28], noting how ML outputs are not always actionable and hence may not necessarily help solve a problem [169]. A questionnaire for reviewing whether development aligns with data ethics principles is available at [173].

Moreover, there is currently a paucity of resources for clinicians to use AI [41]. Hence, a purposeful effort should be undertaken to appropriately train users, encompassing numerous dimensions as outlined in [174]: understanding how AI works; building patient trust; appraising evidence; assessing training data; and mitigating bias. The nature of AI in healthcare as one of shared responsibility should also be emphasised [175]. To increase understanding, outputs should be visualised where possible, making them easier for users to interpret [169]. To assist in transparency, users should be provided with sufficient meta-information about the model (including addressing the questions 'who', 'which', 'what', 'why', 'when' and 'where'), an approach further detailed in the explainable AI framework provided by [146].

When integrating AI into clinical practice, what the system can do (and how well it can do so) should be clearly specified to the user [51]. Information and alerts should be provided to the user in a context-specific manner (i.e. taking into account the user's current tasks and environment) and should be easy to dismiss if not relevant [51]. Clinicians should be able to disagree with AI-driven recommendations [41], and should be able to override them if they believe they have sound clinical reasons for doing so [147]. Additionally, the AI system should be built to accommodate the diverse needs of all its different users (e.g. doctors, nurses, and patients), and to integrate as smoothly as possible into the existing healthcare workflow [159, 176].

The final AI model should be accessible, particularly for marginalised populations [18, 147]. If model developers are charging a licensing fee to healthcare services, one method to promote accessibility is to reduce or waive fees for organisations working in disadvantaged settings [147]. AI should not be used for the purposes of allocating treatment, as most patients are opposed and believe those decisions require their collaborative input [153]. Patients should be informed about the use of AI, the risks, and expected shortcomings of its predictions [167]. Patients should be clear when they are communicating with a human and when they are communicating with an AI system [123]. The language of the AI system should not reinforce unfair stereotypes; a validated set of guidelines for human-AI interaction is available in [51]. Moreover, where possible, implementers and patients should be asked for their preferences, upholding their freedom of choice; choice architecture is further detailed in [147]. If possible, the AI system should alert a trained healthcare professional if the patient is exhibiting acute high-risk behaviours [30].

4.3.2 Updates and ongoing validation

The AI system should have a manual or automatic method to be updated over time [28]. Where possible, the AI should learn from users' behaviour and continually update its operation as a result of user interactions [51]. The AI system should also have a mechanism for users to provide feedback [169], and where unforeseen or unjust mistakes have occurred, a mechanism for adequate redress or reparations to take place [123]. To validate the AI system, performance metrics should be systematically and continuously evaluated after the AI has been deployed [159]. For formally evaluating real-life performance, a prospective clinical trial design may be considered [159], which should be conducted across diverse population groups [167].

4.3.3 Supervision and auditing

AI systems in healthcare bring new risks and amplifies existing ones, in large part due to their capacity for operating automatically and at scale [177]. Operationalising ethics in AI for healthcare necessitates an active governance model, one that implements policies, procedures, and standards to align practices with some overarching ethical principles [178]. Adherence to such governing principles ensures desired outcomes are delivered appropriately and with minimum risks [20]. However, on its own, a "principles based" approach is insufficient for ensuring the implementation of ethical principles [13, 19, 179, 180]. As such, in ethical AI literature, an active governance approach is preferred [181]. Many AI governance approaches emphasise the necessity of regulatory influence to ensure compliance [181, 182], the importance of human-centred governance to achieve shared goals [177], and even the possibility of going beyond the Westernised human rights-based approach to ethics [183].

Once a model of governance—such as that presented in [20]—is in place, the governance can be supported through mechanisms such as internal and external ethics-based auditing [184]. Internal audits can be facilitated by performing algorithmic-use risk analyses that are informed by social impact assessment of similar models. Ongoing internal audits help to avoid potential negative societal impact from the developed AI systems [148]. Organisations that design and deploy AI systems can apply ethics-based auditing through structured processes to assess adherence of their AI systems to ethical principles to take necessary corrective actions to address the identified gaps and enhance user satisfaction and trust in AI systems [150, 184].

Of late, due to high-profile failures reported in the media, AI systems have not enjoyed positive public repute. To make amends and restore trust in these systems, a holistic and transparent approach is needed, one that requires necessary buy-in from an institutional perspective. The institutional approach should ensure clarity about the guiding organizational values used to develop the system, and may involve carrying out red teaming exercises to identify AI risks, conducting third party auditing to identify areas of concern, and communicating any known or potential incidents related to the system [151].

5 Discussion

5.1 Ethical principles and actionable solutions

The ethical principles identified by [14] are preferentially addressed at different stages of the AI development pipeline. Since the distribution of ethical issues is not uniform across the AI lifecycle, we suggest that rather than considering a single ethical principle as done in some guidelines [17], developers should focus on the principles most pertinent to the specific stage of the AI lifecycle they are working on. Trade-offs between ethical principles should be carefully reasoned in this context and clearly documented [123], making use of frameworks such as that offered in [185] to evaluate ethical tensions. In addition to the AI lifecycle perspective, institutional and cultural values must also be proactive upheld for implemented changes to be meaningful and effective.

5.2 Bridging the gap between ethical principles and practices

5.2.1 Improving organisational engagement amongst AI developers

At present, a key barrier and opportunity is that AI developers have only limited awareness of existing solutions that support ethical AI development, with organisational barriers persisting [186]. For example, developers may be unaware how to document their datasets [102] or to present their models in an explainable manner [146], despite these guidelines being available.

As suggested by [186], these issues could be overcome by simple organisational mechanisms, rather than technical solutions. Such measures could include conducting ethics workshops; sending out newsletters with information on ethical AI developments; inviting external speakers to raise awareness; increasing peer networks; and following ethical AI news streams. If adequate resources were available, additional team roles such as a 'Responsible AI Champion' could also be created [186, 187]. These organisational norms could promote collective responsibility and help overcome attitudes such as "it's not my job".

Another opportunity is to associate ethical values with an organisation's beliefs. Aligning AI ethics to organisational values could get more engagement from the employees and consumers, especially when those values are translatable into actionable and operational messages such as: 'we will anonymise your personal data and it is never going to be sold to any third party" [187].

5.2.2 Evolution rather than revolution of existing practices

Empirical research in software engineering has noted that human values can be embedded into software development by evolution, rather than revolution, of existing practices [186]. Hence, we believe that actionable solutions for addressing ethical issues are most likely to appeal to practitioners if they can be easily integrated into existing workflows. Examples of this include guidelines of reporting data [102, 110] and reporting models [144] which provide an

unambiguous, user-friendly structure for completing tasks. Such solutions are straightforward to implement in AI development through evolution of workflows, avoiding a need to adopt an idealistic and impractical revolutionary approach.

From a technical standpoint, since AI development is so closely related to software engineering [26], AI developers should draw from existing relevant practices in software engineering [188, 189]. Exploratory processes in ML share many similarities to scientific programming, and would benefit from lessons learned in embracing uncertainty [190]. Moreover, software engineering has had decades of development in data management, which could be readily adapted to AI practices in many cases [190]. To be effective, AI for healthcare must also consider accessibility, user-centeredness and product design, aspects that can be readily informed by software engineering literature [51].

5.2.3 Limitations of AI governance and regulation

Governance and regulatory mechanisms may provide the necessary support to align AI development with ethical principles. However, organisational leaders may view the adoption of ethical procedures as costly, time-consuming, and counterproductive for holding a competitive edge. Another barrier is how governance and regulatory mechanisms often develop at a relatively slow pace, with regulations like the GDPR providing little support to small business to facilitate easy implementation [191]. Additionally, if auditors do not adequately understand ever-developing AI techniques, discriminatory features (such as image-based AI systems discriminating against darker skin shades) could slip by unnoticed.

5.3 Future directions

Although we have collated recommendations for addressing ethical issues at each stage of the AI lifecycle, a next step is to further expand and formalise these guidelines into categories linked to individual ethical principles such as transparency, fairness, and justice. This can enable convenient accessibility for developers to the set of practices required to embed a particular ethical principle into the system. We then plan to identify and link these guidelines to the AI development team roles that would action them, as well as to governance roles who would be responsible for monitoring and overseeing effective implementation of these guidelines. We also plan to experiment by containerizing and color-coding guidelines and practices based on their relevance to individuals, teams, governance mechanisms, and leadership initiatives. This would enable the implementation of guidelines from both a technical and non-technical standpoint. Integration of guidelines into existing AI pipelines can take many shapes, but an approach supported by current literature is developing 'AI ethics checklists' [2, 17, 21, 192].

We are beginning to see some attempts to utilize AI ethics checklists to address some ethical concerns, such as the checklist for fairness in AI by Madaio et al. [17]. Issues with current ethical AI checklists include a lack of detailed, technical, and actionable activities to embed ethical principles into AI development, and restricted scope dealing with only a particular ethical concern (e.g. fairness) [17]. In any case, such checklists are an expressed priority when ML practitioners are asked directly about their needs, and should therefore be co-designed with their close input [2, 17, 21, 192].

A comprehensive ethical AI checklist that could address every aspect of AI development lifecycle, would need to be developed and comprehensively validated through active involvement of other stakeholders [193]. This approach would allow raising of necessary flags, and notifying relevant stakeholders when and where ethical concerns are raised, which could be linked to existing solutions that we have identified in this work. Although ethical come with their own limitations (such as the risk of people delegating their thinking to checklists alone) [17], they provide a promising next step for further operationalising ethics in AI for healthcare, one that should be taken in conjunction with advancements in governance and regulation.

6 Conclusion

The exponential development of AI in healthcare is promising, but a naïve application of AI to healthcare may lead to a wide array of ethical issues, resulting in avoidable risks and harms. As such, there is a growing need for ensuring AI for healthcare is developed and implemented in an ethical manner. In this paper, we recognise and go beyond solutions that offer principle-based guidance (e.g. adherence to 'fairness'), adopting a solution-based framework that AI developers can use to operationalise ethics in AI for healthcare across all stages of the AI lifecycle—data management, model development, and deployment and monitoring. We emphasise solutions that are actionable (whether technical or non-technical), and therefore utilizable by AI developers. Finally, we acknowledge the growing need for 'ethical AI checklists' co-designed with health AI practitioners, which could further operationalize existing solutions into the AI lifecycle.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s43681-022-00195-z.

Author contributions PS: manuscript writeup and editing; JG: manuscript review and editing; WH: methodology, framework

conceptualization, literature review, supervision, funding acquisition, manuscript review and editing.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This work is supported by a grant from HumaniSE Lab, Monash University.

Data availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose. On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, New York (2016)
2. Chen, I.Y., et al.: Ethical machine learning in health care. Annu. Rev. Biomed. Data Sci. **4**, 123–144 (2021)
3. Jiang, F., et al.: Artificial intelligence in healthcare: past, present and future. Stroke Vasc. Neurol. **2**(4), 230–243 (2017)
4. Zhang, L., et al.: Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study. Sci. Rep. **10**(1), 4406 (2020)
5. Coppersmith, G., et al.: Natural language processing of social media as screening for suicide risk. Biomed. Inform. Insights **10**, 1178222618792860 (2018)
6. Morley, J., et al.: The ethics of AI in health care: a mapping review. Soc. Sci. Med. **260**, 113172 (2020)
7. Fitzpatrick, K.K., Darcy, A., Vierhile, M.: Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment. Health **4**(2), e19 (2017)
8. Guan, J.: Artificial intelligence in healthcare and medicine: promises, ethical challenges and governance. Chin. Med. Sci. J. **34**(2), 76–83 (2019)
9. Gerke, S., Minssen, T., Cohen, G.: Ethical and legal challenges of artificial intelligence-driven healthcare. Artif Intell Healthc. **12**, 295–336 (2020)
10. Noor, P.: Can we trust AI not to further embed racial bias and prejudice? BMJ **368**, m363 (2020)
11. Calo, R.: Artificial intelligence policy: a primary and roadmap. Univ. Bologna Law Rev. **3**, 180–218 (2017)
12. Reddy, S., et al.: A governance model for the application of AI in health care. J. Am. Med. Inform. Assoc. **27**(3), 491–497 (2020)
13. Hagendorff, T.: The ethics of ai ethics: an evaluation of guidelines. Mind. Mach. **30**(1), 99–120 (2020)
14. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019)
15. Morley, J., Floridi, L.: An ethically mindful approach to AI for health care. Lancet **395**(10220), 254–255 (2020)
16. Morley, J., et al.: From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci. Eng. Ethics **26**(4), 2141–2168 (2020)
17. Madaio, M.A., et al.: Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–14 (2020)
18. Nebeker, C., Torous, J., Bartlett Ellis, R.J.: Building the case for actionable ethics in digital health research supported by artificial intelligence. BMC Med. **17**(1), 137 (2019)
19. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**(11), 501–507 (2019)
20. Gasser, U., Almeida, V.A.F.: A layered model for AI governance. IEEE Internet Comput. **21**(6), 58–62 (2017)
21. Nebeker, C., Bartlett Ellis, R.J., Torous, J.: Development of a decision-making checklist tool to support technology selection in digital health research. Transl. Behav. Med. **10**(4), 1004–1015 (2020)
22. Schwartz, S.H.: An overview of the Schwartz theory of basic values. Online Read. Psychol. Culture **2**(1), 1–20 (2012)
23. Perera, H., et al.: A study on the prevalence of human values in software engineering publications, 2015–2018. In: Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering, pp. 409–420 (2020)
24. Ashmore, R., Calinescu, R., Paterson, C.: Assuring the machine learning lifecycle: desiderata, methods, and challenges (2019)
25. Amershi, S., et al.: Software engineering for machine learning: a case study. In: International Conference on Software Engineering 2019, IEEE Computer Society: Montreal, Canada
26. Saleh, Z.: Artificial intelligence definition, ethics and standards (2019)
27. Davenport, T., Kalakota, R.: The potential for artificial intelligence in healthcare. Future Healthc. J. **6**(2), 94–98 (2019)
28. Wiens, J., et al.: Do no harm: a roadmap for responsible machine learning for health care. Nat Med **25**(9), 1337–1340 (2019)
29. Fiske, A., Henningsen, P., Buyx, A.: Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J Med Internet Res **21**(5), e13216 (2019)
30. Luxton, D.D., Anderson, S.L., Anderson, M.: Ethical issues and artificial intelligence technologies in behavioral and mental health care. In: Artificial Intelligence in Behavioral and Mental Health Care, pp. 255–276 (2016)
31. Jacobson, N.C., et al.: Ethical dilemmas posed by mobile health and machine learning in psychiatry research. Bull. World Health Organ **98**(4), 270–276 (2020)
32. Fleming, M.N.: Considerations for the ethical implementation of psychological assessment through social media via machine learning. Ethics Behav. **31**(3), 1–12 (2020)
33. Starke, G., et al.: Computing schizophrenia: ethical challenges for machine learning in psychiatry. Psychol Med. **51**, 2515–2521 (2020)
34. Chancellor, S., et al.: A taxonomy of ethical tensions in inferring mental health states from social media. In: Proceedings of the conference on fairness, accountability, and transparency—FAT* '19, pp. 79–88 (2019)

35. Martinez-Martin, N., Kreitmair, K.: Ethical issues for direct-to-consumer digital psychotherapy apps: addressing accountability, data protection, and consent. JMIR Ment. Health **5**(2), e32 (2018)
36. Char, D.S., Abramoff, M.D., Feudtner, C.: Identifying ethical considerations for machine learning healthcare applications. Am. J. Bioeth. **20**(11), 7–17 (2020)
37. Conway, M., O'Connor, D.: Social media, big data, and mental health: current advances and ethical implications. Curr. Opin. Psychol. **9**, 77–82 (2016)
38. Vreemann, S., et al.: The frequency of missed breast cancers in women participating in a high-risk MRI screening program. Breast Cancer Res. Treat. **169**(2), 323–331 (2018)
39. Lysaght, T., et al.: AI-assisted decision-making in healthcare. Asian Bioeth. Rev. **11**(3), 299–314 (2019)
40. Linthicum, K.P., Schafer, K.M., Ribeiro, J.D.: Machine learning in suicide science: applications and ethics. Behav. Sci. Law **37**(3), 214–222 (2019)
41. Walsh, C.G., et al.: Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence. JAMIA Open **3**(1), 9–15 (2020)
42. Dawson, D., et al.: Artificial intelligence: Australia's ethics framework. Data61 CSIRO: Australia (2019)
43. Thieme, A., Belgrave, D., Doherty, G.: Machine learning in mental health. ACM Trans. Comput. Hum. Interact. **27**(5), 1–53 (2020)
44. Carr, S.: "AI gone mental": engagement and ethics in data-driven technology for mental health. J. Ment. Health **29**(2), 125–130 (2020)
45. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. PMLR (2018)
46. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017)
47. Zou, J., Schiebinger, L.: Ensuring that biomedical AI benefits diverse populations. EBioMedicine **67**, 103358 (2021)
48. Codella, N.C., et al.: Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE (2018)
49. Sefala, R., et al.: Constructing a visual dataset to study the effects of spatial apartheid in South Africa. In: Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2) (2021)
50. Chawla, N.V., et al.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
51. Amershi, S., et al.: Guidelines for human–AI interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2019)
52. Amirruddin, A.D., et al.: Synthetic minority over-sampling TEchnique (SMOTE) and logistic model tree (LMT)-adaptive boosting algorithms for classifying imbalanced datasets of nutrient and chlorophyll sufficiency levels of oil palm (*Elaeis guineensis*) using spectroradiometers and unmanned aerial vehicles. Comput. Electron. Agric. **193**, 106646 (2022)
53. Liew, S.-L., et al.: A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. Sci. Data **5**(1), 1–11 (2018)
54. Abebe, R., et al.: Using search queries to understand health information needs in Africa. In: Proceedings of the International AAAI Conference on Web and Social Media (2019)
55. Jo, E.S., Gebru, T.: Lessons from archives: Strategies for collecting sociocultural data in machine learning. In: Proceedings of the 2020 conference on fairness, accountability, and transparency (2020)
56. Rock, D., Grant, H.: Why diverse teams are smarter. Harv. Bus. Rev. **4**(4), 2–5 (2016)
57. Mannix, E., Neale, M.A.: What differences make a difference? The promise and reality of diverse teams in organizations. Psychol. Sci. Public Interest **6**(2), 31–55 (2005)
58. Salazar, M.R., et al.: Facilitating innovation in diverse science teams through integrative capacity. Small Group Res. **43**(5), 527–558 (2012)
59. Ebadi, A., et al.: How can automated machine learning help business data science teams? In: 2019 18th IEEE International Conference on Machine Learning And Applications (ICMLA). IEEE (2019)
60. Colson, E.: Why data science teams need generalists, not specialists. Harv. Bus. Rev (2019)
61. Sanders, N.: A balanced perspective on prediction and inference for data science in industry (2019)
62. He, J., et al.: The practical implementation of artificial intelligence technologies in medicine. Nat. Med. **25**(1), 30–36 (2019)
63. Char, D.S., Shah, N.H., Magnus, D.: Implementing machine learning in health care—addressing ethical challenges. N. Engl. J. Med. **378**(11), 981 (2018)
64. Mehrabi, N., et al.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) **54**(6), 1–35 (2021)
65. Roh, Y., Heo, G., Whang, S.E.: A survey on data collection for machine learning: a big data-ai integration perspective. IEEE Trans. Knowl. Data Eng. **33**(4), 1328–1347 (2019)
66. Kamiran, F., Zliobaite, I.: Explainable and non-explainable discrimination in classification. In: Magnani, L. (ed.) Studies in Applied Philosophy, Epistemology and Rational Ethics, pp. 155–170. Springer, Berlin (2013)
67. Lee, M.S.A., Floridi, L., Singh, J.: Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. AI Ethics **1**(4), 529–544 (2021)
68. Chen, I., Johansson, F.D., Sontag, D.: Why is my classifier discriminatory? Adv. Neural Inf. Process. Syst. **31**, 3543–3554 (2018)
69. Friedler, S.A., et al.: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency (2019)
70. Thanamani, A.S.: Comparison and analysis of anonymization techniques for preserving privacy in big data. Adv. Comput. Sci. Technol **10**(2), 247–253 (2017)
71. Jain, P., Gyanchandani, M., Khare, N.: Big data privacy: a technological perspective and review. J. Big Data **3**(1), 1–25 (2016)
72. Li, N., T. Li, Venkatasubramanian, S., t-Closeness: privacy beyond k-anonymity and I-diversity. In: 2007 IEEE 23rd international conference on data engineering. IEEE: Istanbul, Turkey (2007)
73. Li, H., et al.: DPSynthesizer: differentially private data synthesizer for privacy preserving data sharing. Proc. VLDB Endowm. **7**(13), 1677–1680 (2014)
74. Hassan, M.U., Rehmani, M.H., Chen, J.: Differential privacy techniques for cyber physical systems: a survey. IEEE Commun. Surv. Tutor. **22**(1), 746–789 (2019)
75. Ye, H., et al.: Secure and efficient outsourcing differential privacy data release scheme in cyber–physical system. Future Gener. Comput. Syst. **108**, 1314–1323 (2020)
76. Dong, J., Roth, A., Su, W.J.: Gaussian differential privacy. arXiv: 1905.02383 (2019)
77. Surendra, H., Mohan, H.: A review of synthetic data generation methods for privacy preserving data publishing. Int. J. Sci. Technol. Res. **6**(3), 95–101 (2017)

78. Ping, H., J. Stoyanovich, Howe, B.: DataSynthesizer. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management, pp. 1–5 (2017)
79. Erlingsson, Ú., Pihur, V., Korolova, A.: Rappor: Randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (2014)
80. Dwork, C., Rothblum, G.N.: Concentrated differential privacy. arXiv:1603.01887 (2016)
81. Mironov, I.: Rényi differential privacy. In: 2017 IEEE 30th Computer Security Foundations Symposium (CSF). 2017. IEEE
82. Xu, J., et al.: Privacy-preserving data integrity verification by using lightweight streaming authenticated data structures for healthcare cyber-physical system. Future Gener. Comput. Syst. **108**, 1287–1296 (2020)
83. Rodríguez-Barroso, N., et al.: Federated learning and differential privacy: software tools analysis, the Sherpa.ai FL framework and methodological guidelines for preserving data privacy. Inf. Fus. 64:270–292 (2020)
84. Kaur, P., Sharma, M., Mittal, M.: Big data and machine learning based secure healthcare framework. Procedia Comput. Sci. **132**, 1049–1059 (2018)
85. Baracaldo, N., et al.: Mitigating poisoning attacks on machine learning models. In: Proceedings of the 10th ACM workshop on artificial intelligence and security, pp. 103–110 (2017)
86. Samangouei, P., Kabkab, M., Chellappa, R.: Defense-gan: Protecting classifiers against adversarial attacks using generative models. arXiv:1805.06605 (2018)
87. Jalal, A., Ilyas, A., Daskalakis, C.: The robust manifold defense: adversarial training using generative models (2017)
88. Wiegand, M., Ruppenhofer, J., Kleinbauer, T.: Detection of abusive language: the problem of biased datasets. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (2019)
89. Nam, J., et al.: Learning from failure: de-biasing classifier from biased classifier. Adv. Neural. Inf. Process. Syst. **33**, 20673–20684 (2020)
90. Mitchell, M., et al.: Diversity and inclusion metrics in subset selection. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (2020)
91. Wachter, S., Mittelstadt, B., Russell, C.: Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. W. Va. L. Rev. **123**, 735 (2020)
92. Calmon, F., et al.: Optimized pre-processing for discrimination prevention. In: Advances in Neural Information Processing Systems, p. 30 (2017)
93. Clifton, D.A., et al.: Machine learning and software engineering in health informatics. In: 2012 First International Workshop on Realizing AI Synergies in Software Engineering (RAISE). Zurich, pp. 37–41 (2012)
94. Batuwita, R., Palade, V.: Efficient resampling methods for training support vector machines with imbalanced datasets. In: The 2010 International Joint Conference on Neural Networks (IJCNN). IEEE: Barcelona, Spain (2010)
95. Calmon, F.P., et al.: Optimized data pre-processing for discrimination prevention (2017)
96. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. **33**(1), 1–33 (2012)
97. Hajian, S., Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in data mining. IEEE Trans. Knowl. Data Eng. **25**, 1445–1459 (2013)
98. Rekatsinas, T., et al.: Holoclean: holistic data repairs with probabilistic inference. arXiv:1702.00820 (2017)
99. Krishnan, S., et al.: Activeclean: Interactive data cleaning for statistical modeling. Proc. VLDB Endow. **9**(12), 948–959 (2016)
100. Tae, K.H., et al.: Data cleaning for accurate, fair, and robust models: a big data-AI integration approach. In: Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning (2019)
101. Cretu, G.F., et al.: Casting out demons: Sanitizing training data for anomaly sensors. In: 2008 IEEE Symposium on Security and Privacy (sp 2008), pp. 81–95 (2008)
102. Gebru, T., et al.: Datasheets for datasets (2018)
103. Costa-jussà, M.R., et al.: Mt-adapted datasheets for datasets: template and repository. arXiv:2005.13156 (2020)
104. Boyd, K.L.: Datasheets for datasets help ML engineers notice and understand ethical issues in training data. Proc. ACM Hum. Comput. Interact. **5**(CSCW2), 1–27 (2021)
105. Hutchinson, B., 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)
106. Hanna, A., et al.: Towards accountability for machine learning datasets (2021)
107. Sharma, S., Chen, K., Sheth, A.: Toward practical privacy-preserving analytics for IoT and cloud-based healthcare systems. IEEE Internet Comput. **22**(2), 42–51 (2018)
108. Arnold, M., et al.: FactSheets: increasing trust in AI services through supplier's declarations of conformity. IBM J. Res. Dev. **63**(4–5), 1–6 (2019). (**13**)
109. Richards, J., et al.: A methodology for creating AI FactSheets. arXiv:2006.13796 (2020)
110. Bender, E.M., Friedman, B.: Data statements for natural language processing: toward mitigating system bias and enabling better science. Trans. Assoc. Comput. Linguist. **6**, 587–604 (2018)
111. McMillan-Major, A., Bender, E.M., Friedman, B.: Data Statements: documenting the datasets used for training and testing natural language processing systems. In: Presented at: Scholarly Communication in Linguistics: Resource Workshop and Poster Session (2022)
112. Holland, S., et al.: The dataset nutrition label: a framework to drive higher data quality standards. arXiv:1805.03677 (2018)
113. Riedl, M.O., Harrison, B.: Using stories to teach human values to artificial agents. In: Workshops at the Thirtieth AAAI Conference on Artificial Intelligence (2016)
114. Nahian, M.S.A., et al.: Learning norms from stories: a prior for value aligned agents. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (2020)
115. Hendrycks, D., et al.: Aligning ai with shared human values. arXiv:2008.02275 (2020)
116. Aghaei, S., Azizi, M.J., Vayanos, P.: Learning optimal and fair decision trees for non-discriminative decision-making. In: Proceedings of the AAAI Conference on Artificial Intelligence (2019)
117. Calders, T., Verwer, S.: Three naive Bayes approaches for discrimination-free classification. Data Min. Knowl. Disc. **21**(2), 277–292 (2010)
118. Ehsan, U., et al.: Operationalizing human-centered perspectives in explainable AI. In: Extended abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (2021)
119. Liao, Q.V., Gruen, D., Miller, S.: Questioning the AI: informing design practices for explainable AI user experiences. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (2020)
120. Hatherley, J.J.: Limits of trust in medical AI. J. Med. Ethics **46**(7), 478–481 (2020)

121. Sparrow, R., Hatherley, J.: High hopes for "Deep Medicine"? AI, economics, and the future of care. Hastings Cent. Rep. **50**(1), 14–17 (2020)
122. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019)
123. Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fus. **58**, 82–115 (2020)
124. Holzinger, A., et al.: What do we need to build explainable AI systems for the medical domain? arXiv:1712.09923 (2017)
125. Holzinger, A., et al.: Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **9**(4), e1312 (2019)
126. Caruana, R.: Case-based explanation for artificial neural nets. In: Artificial Neural Networks in Medicine and Biology. Springer, Berlin, pp. 303–308 (2000)
127. Donadello, I., Dragoni, M., Eccher, C.: Persuasive explanation of reasoning inferences on dietary data. In: PROFILES/SEMEX@ ISWC (2019)
128. Kilbertus, N., et al.: Avoiding discrimination through causal reasoning. In: Advances in Neural Information Processing Systems, p. 30 (2017)
129. Tsamados, A., et al.: The ethics of algorithms: key problems and solutions. AI Soc. **37**, 215–230 (2022)
130. Deldjoo, Y., Di Noia, T., Merra, F.A.: A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks. ACM Comput. Surv. **1**(1), 1–37 (2020)
131. Goodfellow, I.J., et al.: Generative adversarial nets (2014)
132. Samangouei, P., Kabkab, M., Chellappa, R.: Defense-GAN: protecting classifiers against adversarial attacks using generative models (2018)
133. Ilyas, A., et al.: The robust manifold defense: Adversarial training using generative models. arXiv:1712.09196 (2017)
134. Russu, P., et al.: Secure Kernel machines against evasion attacks. In: Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security, pp. 59–69 (2016)
135. Biggio, B., et al.: One-and-a-half-class multiple classifier systems for secure learning against evasion attacks at test time. In: Multiple Classifier Systems. Günzburg, Germany (2015)
136. Gardiner, J., Nagaraja, S.: On the security of machine learning in malware C&C detection: a survey. ACM Comput. Surv. 49(3):Article 59 (2016)
137. Brückner, M., Scheffer, T.: Stackelberg games for adversarial prediction problems. In: The 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, California (2011)
138. Biggio, B., Fumera, G., Roli, F.: Multiple classifier systems for adversarial classification tasks. In: Multiple Classifier Systems, 8th International Workshop. Reykjavik, Iceland (2009)
139. Tantithamthavorn, C., et al.: An empirical comparison of model validation techniques for defect prediction models. IEEE Trans. Softw. Eng. **43**(1), 1–18 (2016)
140. Obermeyer, Z., et al.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**, 447–453 (2019)
141. Chung, Y., et al., *Automated Data Slicing for Model Validation: A Big data - AI Integration Approach.* ArXiv, 2019.
142. Cao, Y., Yang, J.: Towards making systems forget with machine unlearning. In: 2015 IEEE Symposium on Security and Privacy, pp. 463–480 (2015)
143. Hébert-Johnson, U., et al.: Multicalibration: calibration for the (computationally-identifiable) masses. In International Conference on Machine Learning. PMLR (2018)
144. Mitchell, M., et al.: Model cards for model reporting (2019)
145. Amershi, S., et al.: Guidelines for human-AI interaction. In: Proceedings of the 2019 chi conference on human factors in computing systems (2019)
146. Jaigirdar, F.T., et al.: What information is required for explainable AI?: A provenance-based research agenda and future challenges. In: 2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC), pp. 177–183 (2020)
147. Cohen, I.G., et al.: The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Aff. (Millwood) **33**(7), 1139–1147 (2014)
148. Raji, I.D., 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)
149. Song, C., Shmatikov, V.: Auditing data provenance in text-generation models. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 196–206 (2019)
150. Mökander, J., Floridi, L.: Ethics-based auditing to develop trustworthy AI. Mind. Mach. **31**(2), 323–327 (2021)
151. Brundage, M., et al.: Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv:2004.07213 (2020)
152. Gupta, A., Katarya, R.: Social media based surveillance systems for healthcare using machine learning: a systematic review. J Biomed Inform **108**, 103500 (2020)
153. McCradden, M.D., et al.: Ethical concerns around use of artificial intelligence in health care research from the perspective of patients with meningioma, caregivers and health care providers: a qualitative study. CMAJ Open **8**(1), E90–E95 (2020)
154. Jo, E.S., Gebru, T.: Lessons from archives: strategies for collecting sociocultural data in machine learning. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 306–316 (2020)
155. Finlayson, S.G., et al.: Adversarial attacks against medical deep learning systems (2019)
156. Xu, J., et al.: Privacy-preserving data integrity verification by using lightweight streaming authenticated data structures for healthcare cyber–physical system. Future Gener. Comput. Syst. **108**, 1287–1296 (2020)
157. Suster, S., Tulkens, S., Daelemans, W.: A short review of ethical challenges in clinical natural language processing (2017)
158. Lüthi, P., Gagnaux, T., Gygli, M.: Distributed Ledger for provenance tracking of artificial intelligence assets (2020)
159. Rajkomar, A., et al.: Ensuring fairness in machine learning to advance health equity. Ann Intern Med **169**(12), 866–872 (2018)
160. Boselli, R., et al.: Planning meets data cleansing. In: 24th International Conference on Automated Planning and Scheduling. New Hampshire, United States (2014)
161. Hajian, S., Domingo-Ferrer, J.: Direct and indirect discrimination prevention methods. In: Discrimination and privacy in the information society. Springer, Berlin, pp. 241–254 (2013)
162. Kamiran, F., Žliobaitė, I., Calders, T.: Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowl. Inf. Syst. **35**(3), 613–644 (2013)
163. Batuwita, R., Palade, V.: Efficient resampling methods for training support vector machines with imbalanced datasets. In: The 2010 International Joint Conference on Neural Networks (IJCNN). IEEE (2010)
164. Gebru, T., et al.: Datasheets for datasets. Commun. ACM **64**(12), 86–92 (2021)
165. Balakrishnan, A., et al.: Incorporating behavioral constraints in online AI systems. In: The Thirty-Third AAAI Conference on Artificial Intelligence. Honolulu, Hawaii, pp. 3–11 (2019)

166. Zhang, W., Ntoutsi, E.: FAHT: an adaptive fairness-aware decision tree classifier. In: Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao (2019)
167. Yu, K.H., Kohane, I.S.: Framing the challenges of artificial intelligence in medicine. BMJ Qual. Saf. **28**(3), 238–241 (2019)
168. McDaniel, P., Papernot, N., Celik, Z.B.: Machine learning in adversarial settings. IEEE Secur. Priv. **14**(3), 68–72 (2016)
169. Cutillo, C.M., et al.: Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. NPJ Digit Med **3**, 47 (2020)
170. Dent, K.: Ethical considerations for AI researchers (2020)
171. Zhu, H., et al.: Value-sensitive algorithm design. In: Proceedings of the ACM on Human-Computer Interaction. 2(CSCW), pp. 1–23 (2018)
172. Leavy, S.: Gender bias in artificial intelligence. In: Proceedings of the 1st International Workshop on Gender Equality in Software Engineering, pp. 14–16 (2018)
173. DataEthics. Data Ethics Impact Assessment. https://dataethics.eu/ (2021)
174. Shaw, J.A., Sethi, N., Block, B.L.: Five things every clinician should know about AI ethics in intensive care. Intensive Care Med. (2020)
175. Neri, E., et al.: Artificial intelligence: who is responsible for the diagnosis? Radiol Med **125**(6), 517–521 (2020)
176. Polyzotis, N., et al.: Data lifecycle challenges in production machine learning: a survey. SIGMOD Record **47**, 17–28 (2018)
177. Yeung, K., Howes, A., Pogrebna, G.: AI governance by human rights-centred design, deliberation and oversight: an end to ethics washing. In: The Oxford Handbook of AI Ethics. Oxford University Press (2019)
178. Floridi, L., et al.: AI4People-an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach (Dordr) **28**(4), 689–707 (2018)
179. Bietti, E.: From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In: Conference on fairness, accountability, and transparency. Association for Computing Machinery: Barcelona, Spain, pp. 210–219 (2020)
180. Hagendorff, T.: AI virtues: the missing link in putting AI ethics into practice, pp. 1–22 (2021)
181. Rességuier, A., Rodrigues, R.: AI ethics should not remain toothless! A call to bring back the teeth of ethics. Big Data Soc. **7**(2), 1–5 (2020)
182. Schönberger, D.: Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. Int. J. Law Inf. Technol. **27**(2), 171–203 (2019)
183. Smuha, N.A.: Beyond a human rights-based approach to AI governance: promise, pitfalls, Plea. Philos. Technol. (2020)
184. Mökander, J., Floridi, L.: Ethics-based auditing to develop trustworthy AI. Minds Mach. **31**, 323–327 (2021)
185. Xafis, V., et al.: An ethics framework for big data in health and research. Asian Bioeth. Rev. **11**(3), 227–254 (2019)
186. Hussain, W., et al.: Human values in software engineering: contrasting case studies of practice. IEEE Trans. Softw. Eng. **48**(5), 1818–1833 (2020)
187. Ammanath, B., Blackman, R.: Everyone in your organization needs to understand AI ethics. In: Business Ethics. Harvard Business Review: Harvard Business Review (2021)
188. Washizaki, H., et al.: Studying software engineering patterns for designing machine learning systems. In: 2019 10th International Workshop on Empirical Software Engineering in Practice (IWESEP), pp. 49–495 (2019)
189. Serban, A., et al.: Adoption and effects of software engineering best practices in machine learning. In: Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), pp. 1–12 (2020)
190. Wan, Z., et al.: How does machine learning change software development practices? IEEE Trans. Softw. Eng. 1–14 (2019)
191. Politou, E., Alepis, E., Patsakis, C.: Forgetting personal data and revoking consent under the GDPR: challenges and proposed solutions. J. Cybersecur. **4**(1), 1–26 (2018)
192. Holstein, K., et al.: Improving fairness in machine learning systems. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–16 (2019)
193. Lee, M.K., et al.: WeBuildAI. Proc. ACM Hum. Comput. Interact. 3(CSCW):1–35 (2019)