Biases in AI: acknowledging and addressing the inevitable ethical issues

Bjørn Hofmann[1,2]*

[1]Centre of Medical Ethics, The University of Oslo, Oslo, Norway, [2]Institute of the Health Sciences, The Norwegian University of Science and Technology (NTNU), Gjøvik, Norway

Biases in artificial intelligence (AI) systems pose a range of ethical issues. The myriads of biases in AI systems are briefly reviewed and divided in three main categories: input bias, system bias, and application bias. These biases pose a series of basic ethical challenges: injustice, bad output/outcome, loss of autonomy, transformation of basic concepts and values, and erosion of accountability. A review of the many ways to identify, measure, and mitigate these biases reveals commendable efforts to avoid or reduce bias; however, it also highlights the persistence of unresolved biases. Residual and undetected biases present epistemic challenges with substantial ethical implications. The article further investigates whether the general principles, checklists, guidelines, frameworks, or regulations of AI ethics could address the identified ethical issues with bias. Unfortunately, the depth and diversity of these challenges often exceed the capabilities of existing approaches. Consequently, the article suggests that we must acknowledge and accept some residual ethical issues related to biases in AI systems. By utilizing insights from ethics and moral psychology, we can better navigate this landscape. To maximize the benefits and minimize the harms of biases in AI, it is imperative to identify and mitigate existing biases and remain transparent about the consequences of those we cannot eliminate. This necessitates close collaboration between scientists and ethicists.

GRAPHICAL ABSTRACT
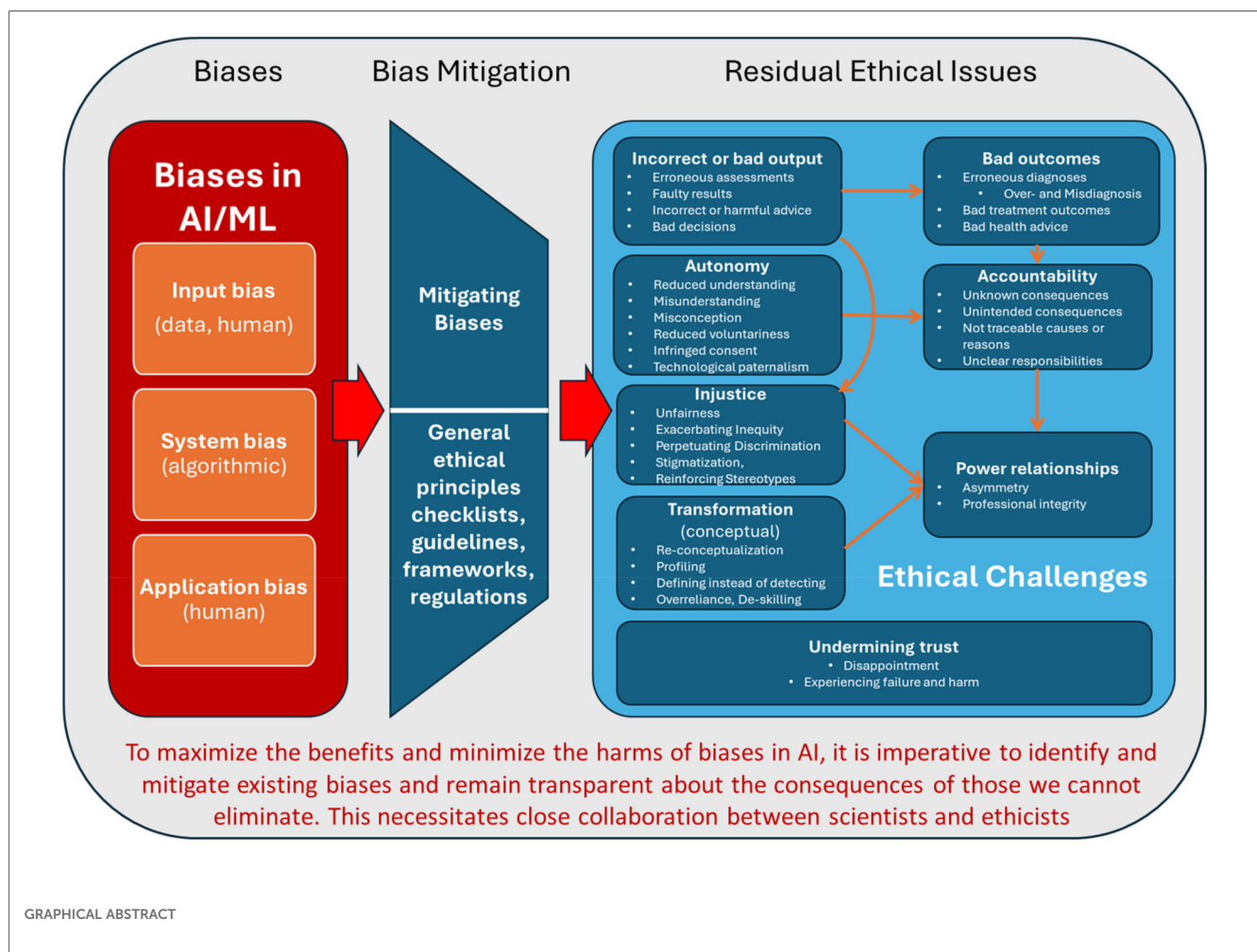
Introduction

The literature on how to identify and assess biases in artificial intelligence (AI) is burgeoning (1–3). So is the literature on how to mitigate such biases (1, 2, 4–10). However, despite great efforts, the problem prevails. So far, biases cannot be eliminated from AI systems. Some biases we therefore have to live with—including their ethical issues.

Correspondingly, there has been a proliferating literature on the ethics of AI (11–20). A wide range of ethical principles, checklists, guidelines, and frameworks have emerged for addressing basic ethical challenges in AI (12, 14–18, 20–35). However, they are rarely tailored to address the ethical aspects of biases.

Hence, there is a need to scrutinize the ethical aspects of biases in AI in more detail. While some studies have addressed specific ethical issues of bias, such as fairness (36), more comprehensive and elaborate analyses are needed.

Accordingly, this article addresses four key questions:

1. What are the biases identified in AI systems? (short overview)
2. What are the basic ethical issues with biases in AI systems?
3. How can biases in AI systems be identified, measured, and mitigated (in order to avoid or reduce their ethical implications)?

4. What can we do to acknowledge and address these (residual) ethical issues with biases in AI?

Very many biases have been identified in AI systems. However, despite great efforts, not all of them seem amenable to mitigation —some we do not know how to mitigate, and others we might not even recognize. Hence, there appear to be unknown residual biases posing epistemic challenges with ethical implications. This article identifies five inevitable ethical challenges with bias in AI (forming the acronym IBATA): Injustice, Bad output/outcome, Autonomy, Transformation, and Accountability.

That is, bias in AI poses special epistemic challenges which are difficult to eliminate and which has important ethical implications (4, 37–40). Unfortunately, general principles, checklists, and frameworks of AI ethics do not seem to be able to address these ethical issues. Therefore, we must identify and mitigate as many biases as possible and strive to reveal the consequences of those that cannot be avoided. Moreover, we must acknowledge and actively address the inevitable ethical challenges with bias to ascertain that the benefits outweigh the harms. Overall, we must strive to use the powerful tool of AI to obtain our goals instead of letting it dictate our values.

For practical reasons, the scope of this study, and its examples, will be limited to healthcare. While the findings may be relevant for AI bias in general, this warrants a separate study.

Artificial intelligence (AI) is used as a generic term, including machine learning and deep learning.

Methods

To address the four questions above narrative reviews are conducted to provide overviews of (1) the biases in AI, (2) the ethics principles, guidelines, and frameworks for artificial intelligence (AI), and (3) of the ways to identify, measure, and mitigate biases in AI. Narrative reviews were conducted according to (41, 42). As other such reviews, this narrative review is "non-quantitative, thematic, educational and …. opinionated" (43).

Initial searches for the topics were done in Google Scholar. Supplemental searches were done in PubMed. Logical search terms were "bias* in AI" and "ethic* in AI". Combinations with "review" and "systematic review" were applied to limit the number of hits. After title and abstract screening, 98 references were included. Snowballing included additional 53 references, and a reviewer suggested additional 19 references (for which I am most thankful).

Data extraction and synthesis: content was extracted from the identified references and synthesized according to the research questions using thematic content analysis. Standard (normative) ethical analysis is applied to identify profound (residual) ethical issues.

Biases in AI systems (RQ1)

Bias is defined as "pervasive simplifications or distortions in judgment and reasoning that systematically affect human decision making" (44). There is a proliferating literature on biases in AI, and the biases are generally divided in three main types (1, 3, 7, 45–47): *input bias*, *system bias*, and *application bias*.

Input biases are biases in the input data for algorithm training. Data can be incomplete, erroneous, or contain biases of a wide range of kinds, e.g., race, sex, age, and socioeconomic status. These biases have many causes, and although they are data-related biases, they originate in human (cognitive and affective) biases, social biases, or organizational biases. Input bias can be revealed by analyzing the data sets. See below. Supplementary Table S1 provides an overview of some major input biases.

System bias is bias in the design and development of algorithms. These biases may originate in selection and sampling (data cleaning, imputation, curation, and treatment of outliers) or in processing and validation of algorithms (48, 49). System bias can be identified and measured by process variables. See below. Supplementary Table S2 provides an overview of some major system biases.

Application bias (also called deployment bias or human bias) stems from the use of the AI systems in practice and is prone to a wide range of human biases (5, 45). Additionally there is bias drift over time (model drift/decay, concept drift) (8, 50). Application bias can be identified and measured by comparative outcome analyses.

A recent systematic review showed that the majority of the studies in the healthcare suffered from input bias and system bias (51). Supplementary Table S3 provides an overview of some major application biases. Figure 1 affords an overview of these three types of biases in AI.

Hence, there is an overwhelming number of biases that can appear in AI systems. Let us now turn to the next question: Which ethical issues do they pose.

Ethical issues with biases in AI (RQ2)

Clearly the vast variety of biases will pose specific ethical issues in particular contexts. However, certain general characteristics of the biases over a variety of contexts may expose some generic ethical issues that are of relevance to a wide range of AI applications. Moreover, while biases may have positive effects, this study will concentrate on its potential negative aspects. The reason for this is that it is crucial that we are aware of and address these in the development, implementation, and use of AI systems.

The most obvious negative implication of bias in AI systems is increased (risk of) *harm* and reduced *safety*, as well as *adverse effects* resulting from erroneous decisions, diagnosis, treatment, or prognosis. The problem is that it violates the ethical principle of non-maleficence (relating to the ancient principle of *primum non nocere*).

Correspondingly, bias may result in poor or erroneous output from the AI system, resulting in bad outcome reducing the *effectiveness* of healthcare services (52). This may origin in a range of the biases listed in S1–3 as well as in model drift, and context ignorance (45). One example is how AI-based tools for assessing skin cancer result in poorer outcomes for populations with diverse skin tones (53). Hence, bias may hamper utility, such as health improvement, and infringe the principle of beneficence.

Yet another obvious ethical challenge following from AI bias is discrimination, unfairness, and stigma. Biases in terms of race, sex, gender, age, socioeconomic status, and ableism are well-documented and undermine the principles of justice and fairness (40, 54–56). As stated in the NIST report "[t]hese biases can negatively impact individuals and society by amplifying and reinforcing discrimination at a speed and scale far beyond the traditional discriminatory practices that can result from implicit human or institutional biases such as racism, sexism, ageism or ableism" (45).

Biases are often latent, that is, they will not be apparent before after long time use (57). This poses a basic *epistemic problem*: the uncertainty of biases adds to the problem of understanding the output from AI systems (explainability, the black-box problem). This challenges the principles of autonomy (and the rule of informed consent) as people are not appropriately informed. It also undermines transparency and accountability.

Several biases can also influence *human agency* as they may reduce human oversight and control, e.g., due to overreliance on AI systems. For example, overreliance on advice has been demonstrated amongst radiologists assessing chest x-rays and making diagnostic decisions (58). Corresponding, the idea that

FIGURE 1
Overview of the three main types of biases in AI.
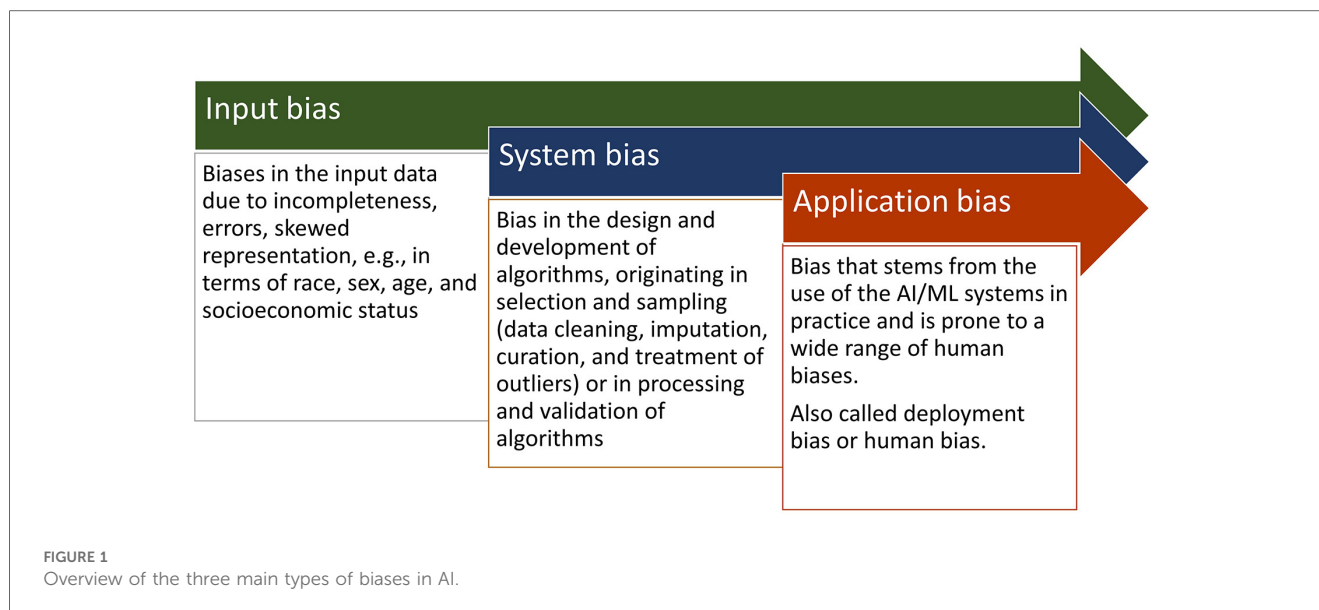
all problems can be solved by technology (techno-solutionism) (59, 60), the belief that technology is always the solution (technochauvinism) (61, 62), or the conception of a technological imperative (63–66), "technological paternalism," (67, 68) or AI-paternalism (69) may reduce human agency as well as challenging the principle of respect for autonomy.

Conceptual challenges raise from biases transforming basic conceptions (70). As pointed out by Floridi: "The digital is deeply transforming reality" (71). Biases may coupling, decoupling, or recoupling features of the world and thereby incite reconceptualization and re-ontologizing of the entities in the world (72). In healthcare this may occur when AI-systems constructed to detect specific conditions (or diagnoses) come to define the same conditions or when AI measures replace human experiences, such as pain or suffering (73). For example, biomarker-based algorithms may change the way we conceptualize, experience, and handle cognitive impairment and Alzheimer's disease. Relatedly, concept drift/creep, model drift, model decay may transform basic conceptions (50). The transformation or re-conceptualization may change social norms and values as well as challenging autonomy and accountability.

Correspondingly, bias may have a *hermeneutic effect*. The output from AI systems may incite new interpretations of agency, personhood, and self-understanding. For example, AI measures may come to (re)define health and disease (wellbeing and suffering) and influence people's interpretation of signs and symptoms, but also of their (self)understanding. This may again challenge their autonomy, integrity, and accountability. It may also instigate hermeneutic epistemic injustice (74).

Moreover, biases may result in a lack of *traceability*, resulting in dissolved or unclear responsibilities (6, 75). Due to lack of transparency in general, and with respect to bias in particular, it can be difficult to hold anybody responsible for errors or harms of bias in AI systems. Again, bias may undermine accountability, and establishing accountability for biased AI outcomes can be difficult (38, 76).

Biases, such as automation complacency (45) or automation bias (5), result in overreliance on AI systems, reduced critical reflection, and deskilling (77, 78). This may change power-relationships and professional integrity, influencing professional ethics. Accordingly, biases in AI systems may reduce trust in such systems and their providers.

Thus, biases in AI systems have a range of ethical implications, raising a series of basic ethical issues, and may undermine several fundamental ethical principles: "the purposes for which AI systems are developed and applied are not in accordance with societal values or fundamental rights such as beneficence, non-maleficence, justice, and explicability" (18). Table 1 provides an overview of ethical implications of AI bias, as well as explanations, examples, and ethical principles or issues arising from these implications.

Hence, a plethora of ethical implications and issues have been identified resulting from biases in AI. Let us now turn to the next question: How can biases be identified, measured, and mitigated? If biases can mitigated, it would resolve or reduce the ethical challenges.

Identifying and mitigating biases in AI (RQ3)

A wide range of approaches have been developed to identify, measure, and mitigate biases (1–3, 7, 79–81). General checklists, such as STARD-AI, TRIPOD-AI, PROBAST, MI-CLAIM, MINIMAR, TEHAI, DECIDE-AI etc aim at avoiding biases.

Correspondingly, there are many methods for detecting and measuring biases in AI systems, such as equalized odds, statistical parity, Context Association Test (CAT), Word Embedding Association Test (WEAT), counterfactual fairness, predictive parity, Categorial Bias Score (CBS), Embedding Coherence Test (ECT) and others. For example, large chest x-ray data sets can be used to demonstrate underdiagnosis bias of artificial intelligence algorithms in under-served patient populations (82). Table 2

TABLE 1 Overview of ethical implications of AI bias, explanations, examples, and ethical principles or issues following from these implications.

Implications	Explanation	Example	Ethical principle, issue
Safety	Increased risk, adverse effect/harm	Erroneous decisions, diagnosis, treatment, prognosis	Non-maleficence, negative utility
Effectiveness	Poor output, bad outcome	Does not improve health	Beneficence
Epistemic: Uncertainty	Lack of understanding	Black-box problem Incomprehensibility	Autonomy Transparency, Accountability
Agency	Lack of control, Paternalism	Tech-reliance, Technochauvinism, Quick Fix	Autonomy
Transformative Conceptual challenges	Reconceptualization Defining > detecting	AI measures frame, form, or replace conceptions and experiences, e.g., of Alzheimer's disease	Accountability Social norms and values Integrity
Hermeneutic changes	New interpretations of agency, personhood, (self)understanding	AI measures come to define basic human experiences, such as health and disease	Autonomy, Integrity, Accountability, Epistemic injustice
Justice	Discrimination, unfairness, stigma	Sex, gender, age, race, rurality, education, economy	Justice Fairness
Traceability	Dissolved or unclear responsibility	Nobody can be held responsible for decisions	Accountability Liability
Under-, Overreliance	Automation bias	Inappropriate reliance on AI	Trust Power(lessness)
De-skilling	Reduced critical reflection	Less experience in examination	Professional (competency and) integrity

gives a brief overview of general checklists for avoiding biases, measures of bias in AI, and bias measurement data sets.

Additionally, there are measures for mitigating biases, such as bias mitigation guidelines (2), checklist (7), bias-handling algorithms (83), debiasing systems, as well as data sets for measuring bias and the effects of bias-mitigating measures (2, 84), as shown in Table 3. Assessing these approaches is beyond the scope of this study, but specific methods can be found in the literature (84). Correspondingly, there are methods to measure and increase fairness by reducing bias (80, 85).

A recent systematic review of electronic health record-based models revealed that 80% of the identified bias mitigation studies reported improved performance after bias mitigation while 13.3% observed unchanged bias after mitigation, and 6.7% found performance variability based on the applied evaluation metrics (1). More specifically, a reduction of racial bias of 84% has been reported by changing the index variable in a commercial prediction algorithm to identify and help patients with complex health needs (56). Yet another example is how group-based training of algorithms for cardiac segmentation in MRI images substantially reduced bias to a standard deviation of 0.89, while making the algorithm impractical (79, 86). Other studies have shown limitations of explainability tools for bias identification (87).

Typical for many mitigation measures is that they are specific and fragmented. They address explicit issues, such as fairness (36, 40), or are directed towards specific biases or processes of AI development (1). However, they may miss out on a range of specific and overarching biases. Moreover, methods for bias measurement and mitigation may themselves be biased. For example what measure you use to estimate fairness (e.g., Equalized odds, Equal opportunity, Precision-recall parity, Predictive equality, Predictive parity, Equal conditional use accuracy, or Equal selectivity), how you choose to estimate these (in terms of true positive rate, area under receiver operating characteristic curve, false positive rate etc), and whether you correct or normalize the calculations, will influence the assessment of bias and fairness (36).

Thus, while novel or evolving approaches, such as algorithmic auditing (88) may further reduce bias in AI, so far we have to address the residual biases. See Figure 2.

Moreover, as we do not know what we do not know about biases in AI systems, there are unknown and unavoidable biases. Other biases may be known, but their effect is unknown. They pose Knightian uncertainty (89). Additionally, biases may stem from indeterminacy, as many key concepts such as pain, suffering, and dysfunction can be defined in many ways. Each definition may have its pros and cons—biasing the outcome of AI systems. Thus, despite great efforts to identify and reduce biases in AI systems, they still pose fundamental epistemic challenges with basic ethical implications.

Acknowledging and addressing ethical issues with biases in AI (RQ4)

How then, can we address the ethical implications of biases in AI? Can they be tackled by applying (some of) the very many ethical principles, approaches, guidelines, checklists, and frameworks that have been developed for ethics in AI? Or do we need other approaches?

Using general ethical principles to address bias problems

Due to the general ethical concerns with AI a wide range of ethical principles, approaches, guidelines, checklists, and frameworks have been developed (11–18, 34, 90). WHO's ethical principles (20), position papers on AI ethics for trustworthy AI (91), as well as regulations, such as the US Algorithmic Accountability Act (92) and the EU Artificial Intelligence Act (93) are but some examples of such efforts.

Several (systematic) reviews provide good overviews of ethical issues and principles (13, 16, 26, 94–97), as illustrated in Figure 3 (with data from 14).

TABLE 2 General checklists for avoiding biases, measures of bias in AI, and bias measurement data sets.

General checklists for avoiding biases	Measures of bias in AI	Bias measurement data sets
Standards for Reporting of Diagnostic Accuracy Study checklist for AI, STARD-AI	Equalized odds	StereoSet, stereotypical biases in gender, profession, race, and religion
Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis for AI, TRIPOD-AI	Statistical parity	WinoBias, identify gender bias in coreference resolution systems
Prediction model risk of bias assessment tool (PROBAST)	The Context Association Test (CAT)	BBQ, bias benchmark for question answering
Minimum information about clinical artificial intelligence modeling, CLAIM/MI-CLAIM	Word Embedding Association Test (WEAT)	BOLD: Dataset and metrics for measuring biases in open-ended language generation.
MINimum Information for Medical AI Reporting, MINIMAR	Counterfactual fairness	
Translational Evaluation of Healthcare AI (TEHAI)	Predictive parity	
Stage-specific reporting guideline for the early and live clinical evaluation of decision-support systems based on artificial intelligence, DECIDE-AI	The Categorial Bias (CB) score	
SPIRIT-AI, a set of recommendations for clinical trial protocols evaluating interventions with an AI component (Liu, 2020)	The Embedding Coherence Test (ECT)	
Recommendations for clinical trial reports evaluating interventions with an AI component, CONSORT-AI	Association without ground truth	
Checklist for AI in Medical Imaging (CLAIM)	Natural language inference	
Consolidated Standards of Reporting Trials-Artificial Intelligence (CONSORT-AI)		

Based on (2, 46, 180).

TABLE 3 Guidelines and checklist for mitigating bias, debiasing systems, as well as data sets for measuring bias and the effects of bias-mitigating measures.

| Bias mitigation measures or strategies	Bias mitigation impact evaluation data sets
Bias mitigation checklist (7)	Corpus of Linguistic Acceptability (CoLA)
Guidelines: • Creating a well-defined goal • Reviewing the training and input data • Using explainable and interpretable models • Feature selection and pre-processing • Regularized model training • Model validation • Algorithmic auditing, • Hyperparameter tuning • Fairness-aware algorithms • Monitoring and feedback	Stanford Sentiment Treebank (SST-2)
Debiasing developed AI systems • Synthetic data augmentation • Re-sampling to balance class distributions • Reweighing to ensure fair representation • Disparate impact remover • Biased embedding correction • Debiasing	Toolkits for detecting and mitigating bias • AI Fairness 360 (AIF360, IBM) • Fairlearn (Microsoft) • What-If Tool (Google) • Fiddler AI

Based on (1, 2, 6, 7, 9, 10, 81).

As can be seen from Table 1, many of the ethical issues and principles at stake for AI in general are relevant with biases in AI as well. For example, transparency and accountability are challenges with bias as well. However, this does not warrant that the general principles or frameworks can address the ethical issues with bias. As pointed out, some biases are latent and unknown while others cannot be eliminated.

Concurrent with the compilation of ethical principles for AI there is an increased awareness of a range of challenges with applying them in practice (14, 98, 99). Vagueness, practical applicability, strong counterforces, and lack of ethical competency are but some of these challenges (14, 100). Additionally, Hagendorff points to other poignant problems in his evaluation of ethics frameworks for ethics in AI. "Currently, AI ethics is failing in many cases. Ethics lacks a reinforcement mechanism. Deviations from the various codes of ethics have no consequences. … Furthermore, empirical experiments show that reading ethics guidelines has no significant influence on the decision-making of software developers. … Distributed

responsibility in conjunction with a lack of knowledge about long-term or broader societal technological consequences causes software developers to lack a feeling of accountability or a view of the moral significance of their work. Especially economic incentives are easily overriding commitment to ethical principles and values" (18).

Moreover, Brent Mittelstadt has pointed out that the generally (bioethical) principle-based approach of AI ethics is inadequate as AI development is substantially different from medical ethics as it lacks "(1) common aims and fiduciary duties, (2) professional history and norms, (3) proven methods to translate principles into practice, and (4) robust legal and professional accountability mechanisms" (101). He continues to point out that the real work of AI ethics is "to translate and implement our lofty principles, and in doing so to begin to understand the real ethical challenges of AI" (101).

While it is beyond the scope of this article to investigate all the ethical principles, checklists, guidelines, frameworks or regulations with respect to the very many biases from AI systems, the mentioned shortcomings indicate that such measures cannot solve all the ethical issues following from (unknown or residual) bias. On a positive note, some frameworks are developed to address specific ethical issues (fairness) of bias in AI (36) and for addressing epistemic-ethical issues in the design of AI systems (55), and can be helpful.

Table 4 provides an overview of how various approaches address the five key ethical challenges with biases in AI (forming the acronym IBATA): Injustice, Bad output/outcome, loss of

FIGURE 2
Residual bias after mitigating various types of bias in AI.
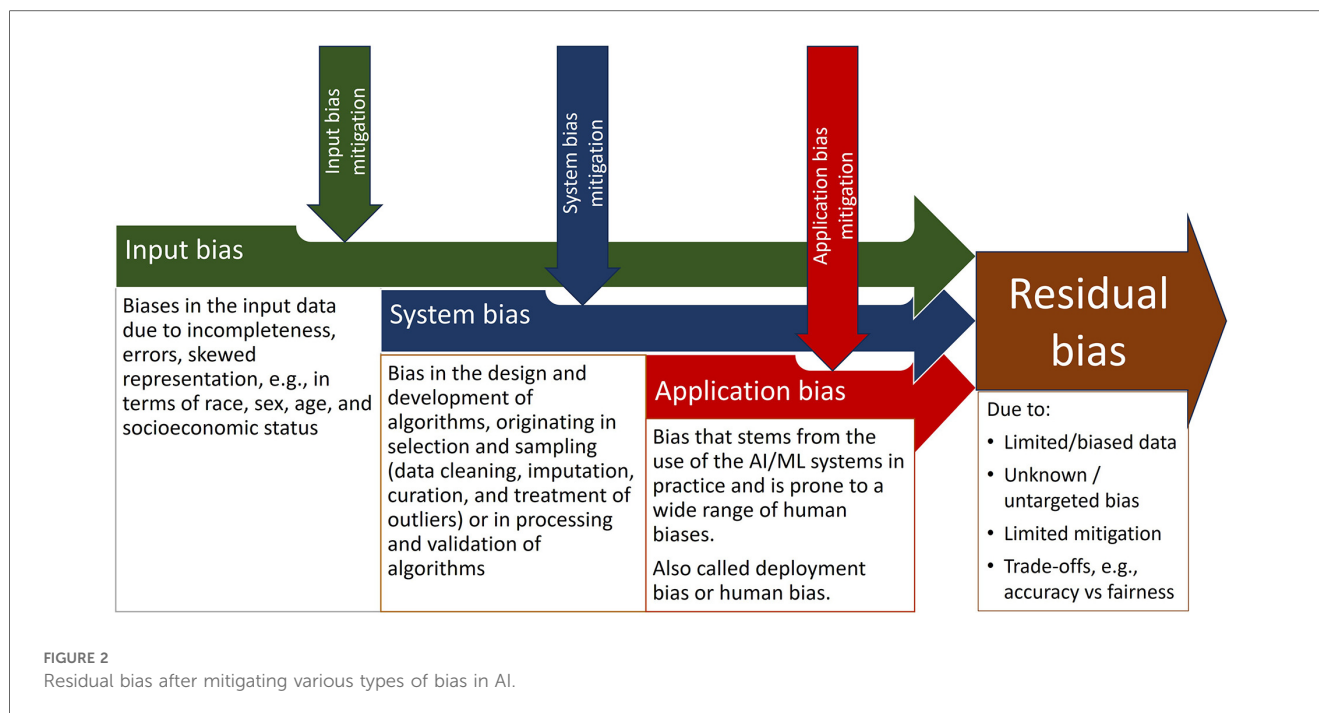
Autonomy, Transformation of basic concepts and values, and loss of Accountability. While several of the frameworks address two or more issues, only one addresses all.

It is also important to notice that very many articles mention ethical principles or frameworks for addressing such issues in AI in general, without demonstrating their application or fruitfulness in the case of bias (83, 119, 120). Others point to ethical challenges with bias (especially fairness) without demonstrating how they can be solved or addressed (121, 122).

Living with residual bias, epistemic challenges, and prevailing ethical issues

As revealed, biases are abundant in AI systems and raise a range of ethical issues. While some of the biases may be mitigated, residual biases appear to prevail. As such, epistemic challenges will occur. The information, suggestions, and advice from AI systems will sometimes be incorrect or imprecise. The knowledge and the derived evidence will occasionally be uncertain and leave us ignorant about crucial factors. Correspondingly, the measures and concepts applied in AI systems may be vague, ambiguous, and change/drift over time. Hence, the output from AI systems (e.g., diagnoses, treatment suggestions, prognoses, decisions etc) may be wrong. Accordingly, the outcomes from such systems may have uncertain efficacy, effectiveness, safety (123), and efficiency (i.e., cost-effectiveness).

This raises a range of ethical issues as elaborated in Table 1. Moreover, they may have regulatory or legal issues (e.g., litigation) and societal challenges (norm creep). Even when applying the rich armamentarium of ethical principles, checklists, guidelines, frameworks, and regulations, some basic issues will prevail: Injustice, Bad output/outcome, loss of Autonomy, Transformation of basic concepts and values, and loss of

Accountability (forming the acronym IBATA). Figure 4 sums up the three main types of biases, mitigating approaches, and basic ethical challenges from residual biases.

How then, can we handle the ethical issues following from biases that cannot be mitigated (because they are unknown or because our mitigation measures are insufficient) or addressed by general approaches in AI ethics? Such ethical issues pose genuine moral dilemmas (124), moral distress (125), moral residue (125, 126), moral doubt (127), and even moral injury (128–130). These challenges are not unique to AI and bias in AI, and a range of approaches have been suggested to address moral residue and moral doubt, such as reflective debriefing, professional counseling, and ethics training (131, 132).

Moreover, as the ethical issues from biases in AI stem from epistemic problems (uncertainty and ignorance), measures to handle uncertainty may be relevant. For example, one can apply a range of strategies to develop uncertainty tolerance (133–142), for uncertainty management (143–150), uncertainty handling (151–153), as well as for increasing comfort with uncertainty (154). In particular, strategies to tolerate and manage uncertainty may help with the cognitive, emotional, behavioural, and moral burden of uncertainty of bias in AI systems (both in terms of whether there is bias and its extension). Importantly, it is crucial to avoid bias numbness, i.e., that "bias is inevitable, so we need not to care".

Correspondingly, one can elaborate on basic concepts, such as outcome measures, in order to reduce bias due to indeterminacy and concept creep. For example, ascertaining that outcome measures directly can be related to human pain, dysfunction or suffering (155) (or wellbeing) can avoid biases due to unclear, vague, or biased concepts.

Corresponding to the de-biasing strategies in AI R&D there are many de-biasing strategies for human biases that may be helpful (156–161). Additionally, we need to pay special attention to
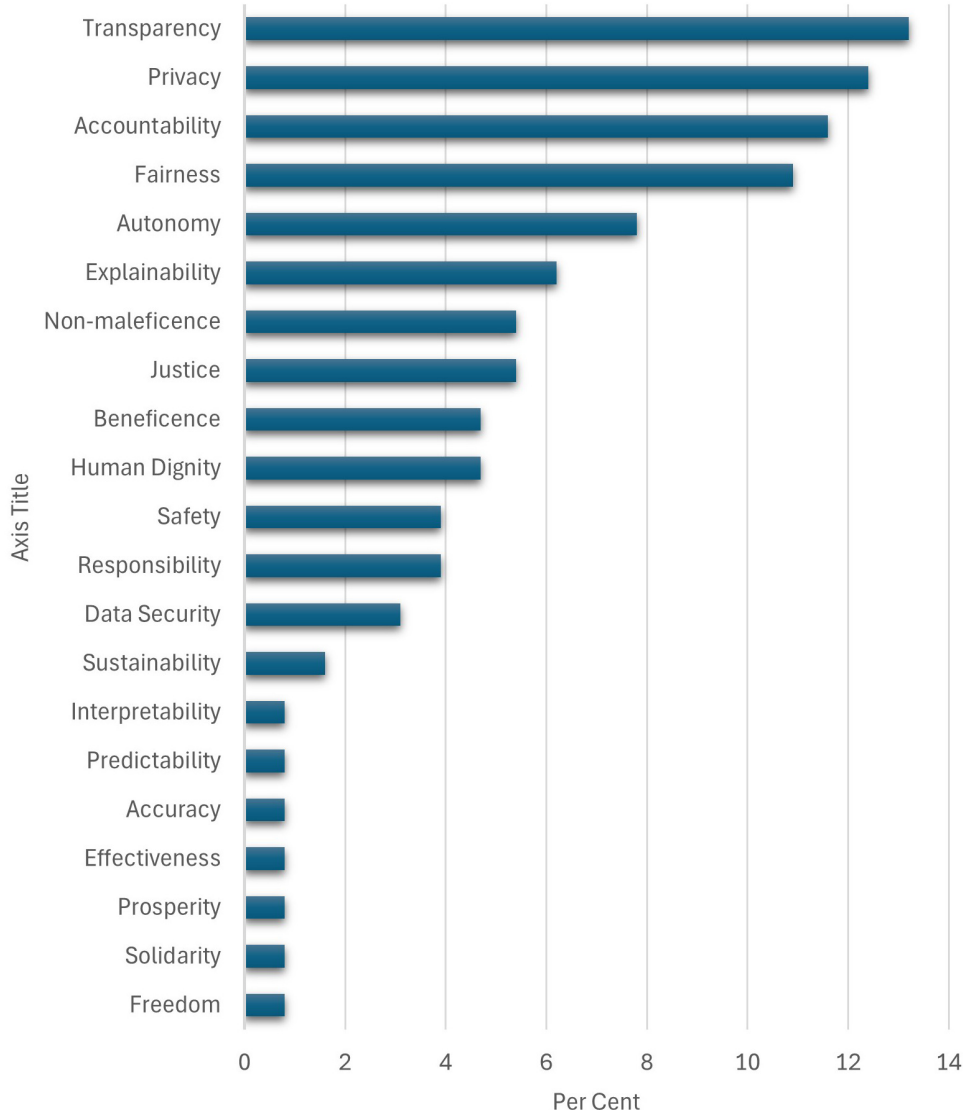
FIGURE 3
Relative frequency in percent of the AI ethics principles identified in a recent systematic review (14).

biases generated by AI, such as overreliance (162), deskilling (77, 78), and acceptance of algorithmic discrimination (162) as they can proliferate or enhance existing bias. Moreover, addressing differences in blaming humans and machines (163, 164) is crucial to address the challenges with accountability.

To maximize autonomy in (biased) AI-based systems it is crucial to be transparent about uncertainty and ignorance about bias and the implications thereof. This is crucial for disclosure in informed consent. Moreover, it is important to be aware of potential paternalism due to bias, e.g., in decision support systems. While paternalism in general is motivated by beneficence, good outcomes may be absent for individuals and groups in biased systems.

To reduce unwarranted transformation, it is crucial to be creative with in envisioning transformative effects of AI systems and their biases. How will the algorithm change our conceptions of the phenomena they handle and the social norms and values

that regulate our behavior. For example, biased AI systems for detecting Alzheimer's disease may change our conceptions of cognitive impairment and our social norms and values (and fears) (165). More generally, we should look for potential conceptual changes (related to health and disease, personal identity, and social status) as well as looping-effects, i.e., human adaptation to classifications and altered classifications (166).

Thus, despite ineliminable (residual) bias in AI systems and unavoidable basic ethical issues, there are measures to face with the ethical aspects of bias in AI. The point in this review has been to identify the ethical issues with bias in AI systems (in healthcare) and not to provide a full-fledged framework to address them. This will be the next step. Nonetheless, the review has provided us with some fruitful initial practical guidance for addressing the basic ethical issues of bias in AI systems, summarized in Table 5.

TABLE 4 Overview of whether the established ethical frameworks or principles for AI mention the ethical issues raised by bias. Dark green means that the issue is more or less addressed. White that it is not addressed. Light green that it is mentioned or implicitly addressed.

| Challenge framework	Injustice	Bad output/ outcome	Autonomy	Transformation	Accountability
AI4People (102)					
Global landscape (103)					
Machine Ethics (104)					
Dynamics of AI Principles (105)					
IEEE Ethically Aligned Design (EAD) (106)					
The European Commission's High-Level Expert (107)					
Group on Artificial Intelligence Report on the Future of Artificial Intelligence (108)					
OECD Recommendation of the Council on Artificial Intelligence (109)					
The Asilomar AI Principles (110)					
AI Now 2019 Report (111)					
Principles for Accountable Algorithms and a Social Impact Statement for Algorithms					
Montréal Declaration for Responsible Development of Artificial Intelligence (112)					
OpenAI Charter (113)					
ITI AI Policy Principles (114)					
Microsoft AI principles (2025) (114)					
Google and DeepMind Ethics Principles (115)					
Google Perspectives on issues in AI governance (116)					
Everyday Ethics for Artificial Intelligence (117)					
Partnership on AI (118)					
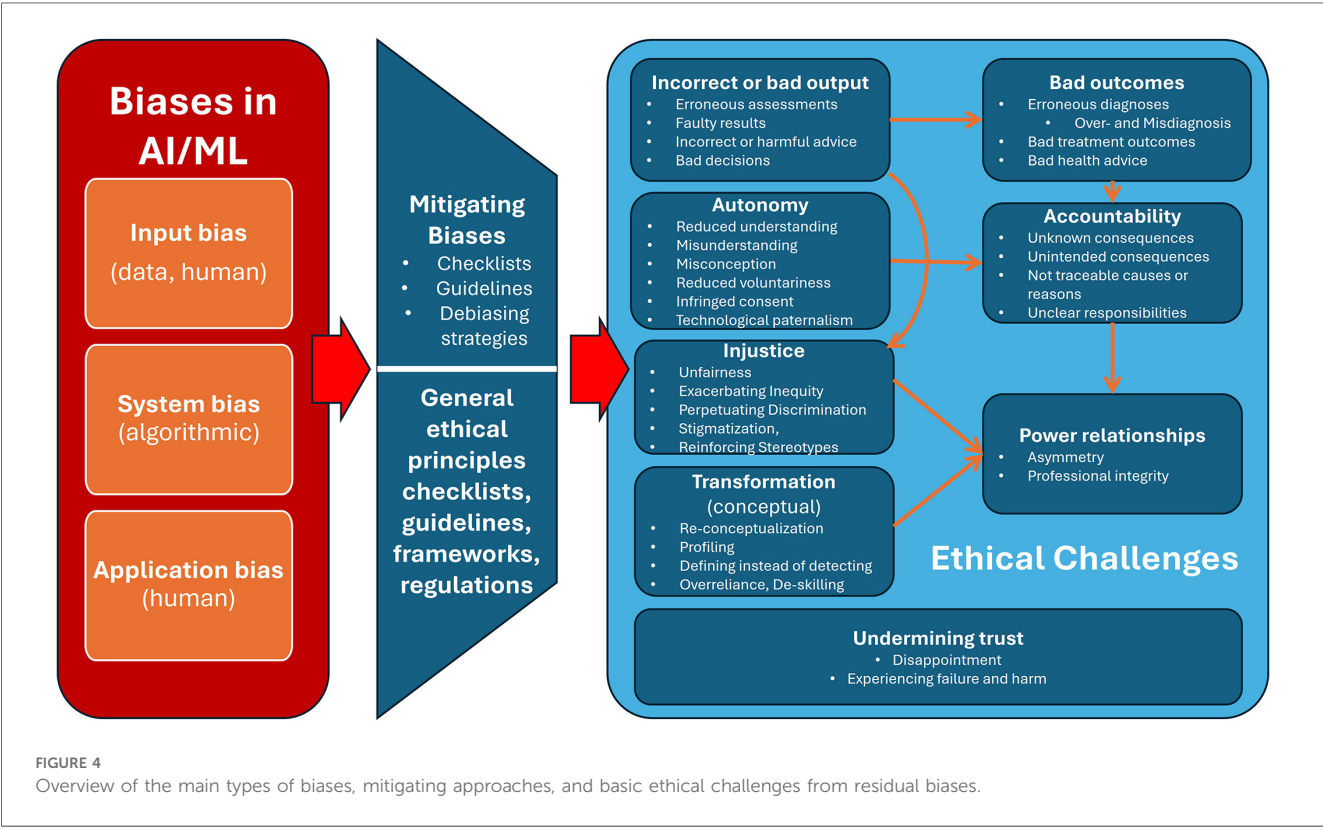
FIGURE 4
Overview of the main types of biases, mitigating approaches, and basic ethical challenges from residual biases.

Instead of believing that the ethical issues can be avoided or handled by the application of ethical principles or perspectives we have to learn how to face with and live with them. Biases add a new type of uncertainty with ethical burdens, that we have to learn to live with (134). Importantly, the biases make it challenging to ascertain that the benefits from AI applied in healthcare outweigh the negative implications. They call for modesty and measures to harness the hype.

TABLE 5 Summary of the practical implications and guidance for the basic ethical issues of bias in AI systems.

Injustice	Bad output/outcome	Autonomy	Transformation	Accountability
Mitigate bias and its implications as much as possible Be transparent about potential injustice due to bias Make trade-offs between fairness and efficiency transparent	Monitor and audit outcomes from AI-based systems Be transparent about potential differences in outcomes Relate outcomes to basic values, such as reduced pain, dysfunction, and suffering	Be transparent about uncertainty and ignorance about bias and the implications thereof (disclosure) Be aware of potential paternalism due to bias, e.g., in decision support	Be creative with respect to envisioning transformative effects of AI systems and their biases Look for conceptual changes and looping-effects	Clarify responsibility for outcomes Address over-reliance and de-skilling Avoid bias-numbing

This indicates that despite great scientific efforts (bias mitigation) and ethical endeavors (AI ethics) we must expect and live with some unknown or residual biases from AI systems. Rather than scaring us off, this should sharpen our attention and inspire our efforts to address biases in AI systems both scientifically and ethically. Even more, it requires a close collaboration between scientists and ethicists.

Discussion

This article started by briefly reviewing the main types of biases in AI systems and identified a series of basic ethical issues from these biases. Then it examined some of the many ways to identify, measure, and mitigate these biases. While acknowledging these great efforts, there are (yet) no measures to eliminate all biases in AI systems. Residual biases pose inevitable epistemic challenges with profound ethical implications and issues. The article then briefly scrutinized whether the general principles, checklists, guidelines, frameworks, or regulations of ethics in AI systems could address the identified ethical issues. However, due to the unresolved epistemic challenges, it is (yet) unlikely that these general approaches will address the ethical issues of biases. Accordingly, we have to acknowledge and live with the ethical issues listed in Table 1 and Figure 4. A host of approaches in ethics and moral psychology offer support to do so. An important lesson from this study is that we have to take biases and their basic ethical issues into account when assessing and implementing AI systems.

It is important to notice that I do not claim or promote any kind of AI exceptionalism. Biases occur with all types of health decisions, and epistemic challenges with ethical implications result from very many technologies, including AI systems (167, 168). However, the hype of AI, its widespread, and partially uncritical implementation makes the ethics of biases in AI highly pertinent.

Moreover, I have not argued that biases will never be eradicated or that ethical principles or frameworks will not ever be able to address the ethical issues. I have only argued that, yet they do not.

Additionally, I have ignored a range of issues, such as global sustainability of developing algorithms. Furthermore, I have not addressed aspects like "the danger of a malevolent artificial general intelligence, machine consciousness, the reduction of social cohesion by AI ranking and filtering systems on social networking sites, the political abuse of AI systems, a lack of

diversity in the AI community, links to robot ethics, the dealing with trolley problems, the weighting between algorithmic or human decision routines, "hidden" social and ecological costs of AI, to the problem of public–private-partnerships and industry-funded research" (18). These are issues for further work.

The implications listed in Table 1 are neither exhaustive nor exclusive. The ethical implications of bias in AI systems can interact and overlap. For example, the overreliance may stem from transformative and conceptual changes. Nonetheless, I believe that the categories are relevant for addressing the ethical issues of bias in AI systems. Future work and development may refine this typology.

Moreover, the review is not exhaustive when it comes to bias mitigation measures or ethical principles and frameworks for AI. The latter has more than 3,680,000 references in Google Scholar. Many more relevant references could have been added, e.g., on intersectionality frameworks applied to AI bias and emerging algorithmic auditing standards (88, 169). Regulatory measures could also have been included, such as the EU AI Act, which in Article 15 addresses bias and refers to the ethical principles of fairness, accountability, transparency, and privacy (170–172). However, they do not provide specific measures and practices to address the ethical issues of biases in AI.

As acknowledged in the introduction, biases may have morally good effects. It is argued that bias can be helpful or contribute to balance injustice (52, 173, 174) and that biases may be corrective: "bias itself might be used to counter the effects of certain other types of bias risks" (173) e.g., in order to reduce risk. There are also ways that AI can be used to omit or reduce human biases. For example, AI can be used for preference-identification and predictions, as humans are bad at anticipating and deliberate on future events due to various biases (175, 176). Even if some biases are good, we need to differentiate the good from the bad, i.e., we need to identify the negative implications of biases in AI (and balance them against the positive ones as well as the benefits from the AI systems as such). Reviewing the morally good aspects of bias in AI is beyond the scope of this study and warrants a separate investigation.

Moreover, the scope of this study, and its examples, has been limited to healthcare. While the findings may be relevant for other fields of AI applications or for AI bias in general, further studies are needed to investigate its generalizability and transferability. Such studies can benefit from comparative studies with and within other fields, such as criminal justice (85, 177, 178).

As acknowledged, the study of bias may itself be biased. The literature has identified stakeholders to be "bias apologists" and "bias deniers" (179). This may challenge the work on acknowledging and addressing the biases and its ethical implications.

Conclusion

The brief review of the vast number of biases in AI systems identified three main types of bias: input bias, system bias, and application bias. These biases pose a series of basic ethical challenges: injustice, bad output/outcome, loss of autonomy, transformation of basic concepts and values, and loss of accountability (IBATA). Reviewing the many ways to identify, measure, and mitigate these biases demonstrated great efforts to reduce biases and their ethical implications. However, at present they are not able to eliminate all biases. Some biases remain unknown, and residual biases pose inevitable epistemic challenges with profound ethical implications and issues. Investigating whether the general principles, checklists, guidelines, frameworks, or regulations of AI ethics could address the identified ethical issues with bias ends negative as the ethical issues are profound, diverse, and complex. Instead, it is suggested that we have to live with the (residual) ethical issues of biases in AI systems. A host of approaches in ethics and moral psychology offer support to do so.

Few technologies are flawless. Avoiding all ethical issues of AI is impossible. However, the task is to maximize the benefits and minimize the harms—and to provide so much knowledge about both benefits and harms as possible. Therefore, we must identify and mitigate as many biases as possible and strive to reveal the consequences of those that cannot be avoided.

The epistemic and ethical challenges with biases in AI systems both should sharpen our attention and inspire our efforts both scientifically and ethically. Even more, it requires a close collaboration between scientists and ethicists.

Overall, we must strive to use this powerful tool to obtain our goals instead of letting it dictate our values.

Author contributions

BH: Project administration, Formal analysis, Writing – review & editing, Methodology, Visualization, Writing – original draft, Investigation, Conceptualization.

Funding

Acknowledgments

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2025.1614105/full#supplementary-material

References

1. Chen F, Wang L, Hong J, Jiang J, Zhou L. Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. *J Am Med Inform Assoc*. (2024) 31(5):1172–83. (In eng). doi: 10.1093/jamia/ocae060

2. Gray M, Samala R, Liu Q, Skiles D, Xu J, Tong W, et al. Measurement and mitigation of bias in artificial intelligence: a narrative literature review for regulatory science. *Clin Pharmacol Ther*. (2024) 115(4):687–97. doi: doi: 10.1002/cpt.3117

3. Varsha P. How can we manage biases in artificial intelligence systems–A systematic literature review. *Int J Infm Manag Data Insights*. (2023) 3(1):100165. doi: 10.1016/j.jjimei.2023.100165

4. DeCamp M, Lindvall C. Mitigating bias in AI at the point of care. *Science*. (2023) 381(6654):150–2. (In eng). doi: 10.1126/science.adh2713

5. Koçak B, Ponsiglione A, Stanzione A, Bluethgen C, Santinha J, Ugga L, et al. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagn Interv Radiol*. (2024) 31:75. doi: 10.4274/dir.2024.242854

6. Mensah GB. Artificial intelligence and ethics: a comprehensive review of bias mitigation, transparency, and accountability in AI systems. Preprint, November 2023;10. (2023).

7. Nazer LH, Zatarah R, Waldrip S, Ke JXC, Moukheiber M, Khanna AK, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLoS Digit Health*. (2023) 2(6):e0000278. doi: 10.1371/journal.pdig. 0000278

8. Pareek C. Unmasking bias: a framework for testing and mitigating AI bias in insurance underwriting models. *J Artif Intell Mach Learn & Data Sci*. (2023) 1(1):1736–41. doi: 10.51219/JAIMLD/Chandra-shekhar-pareek/377

9. Sasseville M, Ouellet S, Rhéaume C, Sahlia M, Couture V, Després P, et al. Bias mitigation in primary health care artificial intelligence models. Scoping review. *J Med Internet Res*. (2025) 27:e60269. doi: 10.2196/60269

10. Van Giffen B, Herhausen D, Fahse T. Overcoming the pitfalls and perils of algorithms: a classification of machine learning biases and mitigation methods. *J Bus Res*. (2022) 144:93–106. doi: 10.1016/j.jbusres.2022.01.076

11. Benzinger L, Ursin F, Balke W-T, Kacprowski T, Salloch S. Should artificial intelligence be used to support clinical ethical decision-making? A systematic review of reasons. *BMC Med Ethics* 2023;24(1):48. doi: 10.1186/s12910-023-00929-6

12. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). *NPJ Digit Med*. (2024) 7(1):183. doi: 10.1038/s41746-024-01157-x

13. Karimian G, Petelos E, Evers SM. The ethical issues of the application of artificial intelligence in healthcare: a systematic scoping review. *AI Ethics*. (2022) 2(4):539–51. doi: 10.1007/s43681-021-00131-7

14. Khan AA, Badshah S, Liang P, Waseem M, Khan B, Ahmad A, et al. Ethics of AI: a systematic literature review of principles and challenges. *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering*. New York (2022). p. 383–92

15. Morley J, Floridi L. The ethics of AI in health care: an updated mapping review. Available online at SSRN 4987317 (2024).

16. Murphy K, Di Ruggiero E, Upshur R, Willison DJ, Malhotra N, Cai JC, et al. Artificial intelligence for good health: a scoping review of the ethics literature. *BMC Med Ethics*. (2021) 22(1):1–17. doi: 10.1186/s12910-021-00577-8

17. Tang L, Li J, Fantus S. Medical artificial intelligence ethics: a systematic review of empirical studies. *Digit Health*. (2023) 9:20552076231186064. doi: 10.1177/20552076231186064

18. Hagendorff T. The ethics of AI ethics: an evaluation of guidelines. *Minds Mach*. (2020) 30(1):99–120. doi: 10.1007/s11023-020-09517-8

19. Bakiner O. What do academics say about artificial intelligence ethics? An overview of the scholarship. *AI Ethics*. (2023) 3(2):513–25. doi: 10.1007/s43681-022-00182-4

20. WHO. *Ethics and Governance of Artificial Intelligence for Health*. Geneva: World Health Organization (2021).

21. Bostrom N, Yudkowsky E. The ethics of artificial intelligence. In: Yampolskiy, Roman V, editor. *Artificial Intelligence Safety and Security*. Boca Raton, FL: Chapman and Hall/CRC (2018). p. 57–69. doi: 10.1201/9781351251389-4

22. Di Nucci E. Should we be afraid of medical AI? *J Med Ethics*. (2019) 45(8):556–8. doi: 10.1136/medethics-2018-105281

23. Floridi L. *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford: Oxford University Press (2023).

24. French SE, Lee K, Kibben M, Rose S. Are we ready for artificial ethics: aI and the future of ethical decision making. *Int J Ethical Leadership*. (2019) 6(1):24–53. https://scholarlycommons.law.case.edu/ijel/vol6/iss1/4

25. Hagendorff T. A virtue-based framework to support putting AI ethics into practice. *Philos Technol*. (2022) 35(3):55. doi: 10.1007/s13347-022-00553-z

26. Huang C, Zhang Z, Mao B, Yao X. An overview of artificial intelligence ethics. *IEEE Trans Artif Intell*. (2022) 4(4):799–819. doi: 10.1109/TAI.2022.3194503

27. Khalil OE. Artificial decision-making and artificial ethics: a management concern. *J Bus Ethics*. (1993) 12:313–21. doi: 10.1007/BF01666535

28. Liao SM. *Ethics of Artificial Intelligence*. Oxford: Oxford University Press (2020).

29. Müller VC. Ethics of artificial intelligence and robotics. (2020).

30. O'Neill E, Klincewicz M, Kemmer M. Ethical issues with artificial ethics assistants. (2021).

31. Panǎ LL. Artificial Ethics. In: Khosrow-Pour M, editor. *Encyclopedia of Information Science and Technology*, Fourth Edition. New York: IGI Global Scientific Publishing (2018). p. 88–97. doi: 10.4018/978-1-5225-2255-3.ch008

32. Pereira LM, Lopes AB. Machine ethics. *Studies in Applied Philosophy, Epistemology and Rational Ethics*. Cham: Springer (2011). p. 53.

33. Wallach W, Asaro P. *Machine Ethics and Robot Ethics*. London: Routledge (2020).

34. Murphy MA. Using structured ethical techniques to facilitate reasoning in technology ethics. *AI Ethics*. (2025) 5(1):479–88. doi: 10.1007/s43681-023-00371-9

35. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: mapping the debate. *Big Data Soc*. (2016) 3(2):2053951716679679. doi: 10.1177/2053951716679679

36. Hoche M, Mineeva O, Rätsch G, Vayena E, Blasimme A. What makes clinical machine learning fair? A practical ethics framework. *PLoS Digit Health*. (2025) 4(3): e0000728. doi: 10.1371/journal.pdig.0000728

37. Caliskan A. Artificial intelligence, bias, and ethics. IJCAI2023:7007-7013.

38. Oyeniran C, Adewusi AO, Adeleke AG, Akwawa LA, Azubuko CF. Ethical AI: addressing bias in machine learning models and software applications. *Comput Sci IT Res J*. (2022) 3(3):115–26. doi: 10.51594/csitrj.v3i3.1559

39. Yapo A, Weiss J. Ethical implications of bias in machine learning. (2018).

40. Modi TB. Artificial intelligence ethics and fairness: a study to address bias and fairness issues in AI systems, and the ethical implications of AI applications. *Rev Review Index J Multidiscip*. (2023) 3(2):24–35. doi: 10.31305/rrijm2023.v03.n02.004

41. Green BN, Johnson CD, Adams A. Writing narrative literature reviews for peer-reviewed journals: secrets of the trade. *J Chiropr Med*. (2006) 5(3):101–17. doi: 10.1016/S0899-3467(07)60142-6

42. Chaney MA. So you want to write a narrative review article? *J Cardiothorac Vasc Anesth*. (2021) 35(10):3045–9. doi: 10.1053/j.jvca.2021.06.017

43. Agarwal S, Charlesworth M, Elrakhawy M. How to write a narrative review. *Anaesthesia*. (2023) 78(9):1162–6. doi: 10.1111/anae.16016

44. Toet A, Brouwer A-M, van den Bosch K, Korteling J. Effects of personal characteristics on susceptibility to decision bias: a literature study. *Int J Humanit Soc Sci*. (2016) 5:1–17. https://karelvandenbosch.nl/documents/2016_Toet_etal_IJHSS_Effects_of_personal_characteristics_on_susceptibility_to_decision_bias_a_literature_study.pdf

45. Schwartz R, Schwartz R, Vassilev A, Greene K, Perine L, Burt A, et al. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. US Department of Commerce, Gaithersburg, MD: National Institute of Standards and Technology (NIST) (2022).

46. Hanna M, Pantanowitz L, Jackson B, Palmer O, Visweswaran S, Pantanowitz J, et al. Ethical and bias considerations in artificial intelligence (AI)/machine learning. *Mod Pathol*. (2024) 38:100686. doi: 10.1016/j.modpat.2024.100686

47. Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal M, et al. Bias in data-driven artificial intelligence systems—an introductory survey. *WIREs Data Mining Knowl Discov*. (2020) 10(3):e1356. doi: doi: 10.1002/widm.1356

48. Ratwani RM, Sutton K, Galarraga JE. Addressing AI algorithmic bias in health care. *JAMA*. (2024) 332(13):1051–2. doi: 10.1001/jama.2024.13486

49. Flores L, Kim S, Young SD. Addressing bias in artificial intelligence for public health surveillance. *J Med Ethics*. (2024) 50(3):190–4. doi: 10.1136/jme-2022-108875

50. Abdul Razak MS, Nirmala CR, Sreenivasa BR, Lahza H, Lahza HFM. A survey on detecting healthcare concept drift in AI/ML models from a finance perspective. *Front Artif Intell*. (2023) 5:955314. doi: 10.3389/frai.2022.955314

51. Kumar A, Aelgani V, Vohra R, Gupta SK, Bhagawati M, Paul S, et al. Artificial intelligence bias in medical system designs: a systematic review. *Multimed Tools Appl*. (2024) 83(6):18005–57. doi: 10.1007/s11042-023-16029-x

52. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? *AMA J Ethics*. (2019) 21(2):167–79. doi: 10.1001/amajethics.2019.167

53. Daneshjou R, Vodrahalli K, Liang W, Novoa RA, Jenkins M, Rotemberg V, et al. Disparities in dermatology ai: assessments using diverse clinical images. arXiv [Preprint] arXiv:211108006. (2021). doi: 10.48550/arXiv.2111.08006

54. Lo Piano S. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Hum Soc Sci Commun*. (2020) 7(1):9. doi: 10.1057/s41599-020-0501-9

55. Russo F, Schliesser E, Wagemans J. Connecting ethics and epistemology of AI. *AI Soc*. (2023) 39:1585–603. doi: 10.1007/s00146-022-01617-6

56. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. (2019) 366(6464):447–53. doi: 10.1126/science.aax2342

57. DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. *J Am Med Inform Assoc*. (2020) 27(12):2020–3. doi: 10.1093/jamia/ocaa094

58. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med*. (2021) 4(1):31. doi: 10.1038/s41746-021-00385-9

59. Sætra HS. *Technology and Sustainable Development: The Promise and Pitfalls of Techno-solutionism*. New York, NY: Taylor & Francis (2023).

60. Berendt B. AI for the common good?! pitfalls, challenges, and ethics pen-testing. *Paladyn, J Behav Robot*. (2019) 10(1):44–65. doi: 10.1515/pjbr-2019-0004

61. Broussard M. *Artificial Unintelligence: How Computers Misunderstand the World*. Boston: MIT Press (2018).

62. König PD, Wenzelburger G. Between technochauvinism and human-centrism: can algorithms improve decision-making in democratic politics? *Eur Polit Sci*. (2022) 21:132–49. doi: 10.1057/s41304-020-00298-3

63. Barger-Lux MJ, Heaney RP. For better and worse: the technological imperative in health care. *Soc Sci Med*. (1986) 22(12):1313–20. doi: 10.1016/0277-9536(86)90094-8

64. Hofmann B. Is there a technological imperative in health care? *Int J Technol Assess Health Care*. (2002) 18(3):675–89. doi: 10.1017/S0266462302000491

65. Koenig BA. The technological imperative in medical practice: the social creation of a "routine" treatment. In: Lock M, Gordon D, editors. *Biomedicine Examined*. Dordrecht: Springer (1988). p. 465–96. doi: 10.1007/978-94-009-2725-4_18

66. Rothman D. *Beginnings Count: The Technological Imperative in American Health Care*. New York: Oxford University Press (1997).

67. Voinea C, Wangmo T, Vică C. Paternalistic AI: the case of aged care. *Hum Soc Sci Commun*. (2024) 11(1):824. doi: 10.1057/s41599-024-03282-0

68. Hofmann B. Technological paternalism: on how medicine has reformed ethics and how technology can refine moral theory. *Sci Eng Ethics*. (2003) 9(3):343–52. (Vitenskapelig artikkel) (In Engelsk). doi: 10.1007/s11948-003-0031-z

69. Milian RD, Bhattacharyya A. Artificial intelligence paternalism. *J Med Ethics*. (2023) 49(3):183–4. doi: 10.1136/jme-2022-108768

70. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. (2019) 380(14):1347–58. doi: 10.1056/NEJMra1814259

71. Floridi L. Digital's cleaving power and its consequences. *Philos Technol*. (2017) 30(2):123–9. doi: 10.1007/s13347-017-0259-1

72. Floridi L. *Information: A very Short Introduction*. Oxford: OUP Oxford (2010).

73. Hofmann BM. My biomarkers are fine, thank you": on the biomarkerization of modern medicine. *J Gen Internal Med*. (2024)40:453–7. (Vitenskapelig artikkel) (In Engelsk). doi: 10.1007/s11606-024-09019-8

74. Fricker M. *Epistemic Injustice. Power and the Ethics Of Knowing*. Oxford: Oxford University Press (2007).

75. Coeckelbergh M. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Sci Eng Ethics*. (2020) 26(4):2051–68. doi: 10.1007/s11948-019-00146-8

76. Peinelt N. *Detecting Semantic Similarity: Biases, Evaluation and Models*. Warwick: University of Warwick (2021).

77. Lee H-PH, Sarkar A, Tankelevitch L, Drosos I, Rintel S, Banks R, et al. The impact of generative AI on critical thinking: self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. (2025).

78. Farhan A. The impact of artificial intelligence on human workers. *J Commun Educ*. (2023) 17(2):93–104. doi: 10.58217/joce-ip.v17i2.350

79. Hasanzadeh F, Josephson CB, Waters G, Adedinsewo D, Azizi Z, White JA. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *NPJ Digit Med*. (2025) 8(1):154. doi: 10.1038/s41746-025-01503-7

80. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv (CSUR)*. (2021) 54(6):1–35. doi: 10.1145/3457607

81. Siddique S, Haque MA, George R, Gupta KD, Gupta D, Faruk MJH. Survey on machine learning biases and mitigation techniques. *Digital*. (2024) 4(1):1–68. doi: 10.3390/digital4010001.

82. Seyyed-Kalantari L, Zhang H, McDermott MB, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med*. (2021) 27(12):2176–82. doi: 10.1038/s41591-021-01595-0

83. Wu W, Huang T, Gong K. Ethical principles and governance technology development of AI in China. *Engineering*. (2020) 6(3):302–9. doi: 10.1016/j.eng.2019.12.015

84. Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. *IBM J Res Dev*. (2019) 63(4/5):4:1–4:15. doi: 10.1147/JRD.2019.2942287

85. Barocas S, Hardt M, Narayanan A. *Fairness and Machine Learning: Limitations and Opportunities*. Boston: MIT Press (2023).

86. Puyol-Antón E, Ruijsink B, Piechnik SK, Neubauer S, Petersen SE, Razavi R, et al. Fairness in cardiac MR image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer (2021) p. 413–423.

87. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling lime and shap: adversarial attacks on *post hoc* explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020). p. 180–6

88. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020). p. 33–44.

89. Knight FH. *Risk, Uncertainty and Profit*. Boston: Houghton Mifflin Company (1921).

90. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ*. (2020) 368:132–49. doi: 10.1136/bmj.l6927

91. Rathkopf C, Heinrichs B. Learning to live with strange error: beyond trustworthiness in artificial intelligence ethics. *Camb Q Healthc Ethics*. (2023) 33:333–45. doi: 10.1017/S0963180122000688

92. Gursoy F, Kennedy R, Kakadiaris I. A critical assessment of the algorithmic accountability act of 2022. Available online at SSRN 4193199. (2022). doi: 10.2139/ssrn.4193199

93. Mökander J, Juneja P, Watson DS, Floridi L. The US algorithmic accountability act of 2022 vs. The EU artificial intelligence act: what they can learn from each other? *Minds Mach*. (2022) 32(4):751–8. doi: 10.1007/s11023-022-09612-y

94. Al-Hwsali A, Alsaadi B, Abdi N, Khatab S, Alzubaidi M, Solaiman B, et al. Scoping review: legal and ethical principles of artificial intelligence in public health. *Healthcare Transform Inform Artif Intell*. (2023) 305:640–3.

95. Möllmann NR, Mirbabaie M, Stieglitz S. Is it alright to use artificial intelligence in digital health? A systematic literature review on ethical considerations. *Health Informatics J*. (2021) 27(4):14604582211052391. doi: 10.1177/14604582211052391

96. Siala H, Wang Y. SHIFTing artificial intelligence to be responsible in healthcare: a systematic review. *Soc Sci Med*. (2022) 296:114782. doi: 10.1016/j.socscimed.2022.114782

97. Stahl BC, Stahl BC. Ethical issues of AI. *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*. Dordrecht: Springer Nature (2021). p. 35–53.

98. Buolamwini J, Unmasking AI. *My Mission to Protect What is Human in a World of Machines*. New York: Random House (2024).

99. Sætra HS, Danaher J. To each technology its own ethics: the problem of ethical proliferation. *Philos Technol*. (2022) 35(4):93. doi: 10.1007/s13347-022-00591-7

100. Whittlestone J, Nyrup R, Alexandrova A, Cave S. The role and limits of principles in AI ethics: towards a focus on tensions. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019). p. 195–200

101. Mittelstadt B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell*. (2019) 1(11):501–7. doi: 10.1038/s42256-019-0114-4

102. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—an Ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach*. (2018) 28:689–707. doi: 10.1007/s11023-018-9482-5

103. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell*. (2019) 1(9):389–99. doi: 10.1038/s42256-019-0088-2

104. Siau K, Wang W. Artificial intelligence (AI) ethics: ethics of AI and ethical AI. *J Database Manage*. (2020) 31(2):74–87. doi: 10.4018/JDM.2020040105

105. Canca C. Operationalizing AI ethics principles. *Commun ACM*. (2020) 63(12):18–21. doi: 10.1145/3430368

106. Association IS. The IEEE global initiative on ethics of autonomous and intelligent systems. Recuperado de Standards ieee org. (2018).

107. Pekka A, Bauer W, Bergmann U, Bieliková M, Bonefeld-Dahl C, Bonnet Y, et al. The european commission's high-level expert group on artificial intelligence: ethics guidelines for trustworthy AI. *Working Document for Stakeholders' Consultation Brussels*. Brussels: European Commission (2018). p. 1–37.

108. Holdren JP, Bruce A, Felten E, Garris M, Lyons T. *Preparing for the Future of Artificial Intelligence*. Washington, DC: Springer (2016).

109. Organisation for Economic Co-operation and Development. *Recommendation of the Council on Artificial Intelligence*. Paris: OECD Legal Instruments (2019). Available online at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

110. Institute FoL. Asilomar AI principles. Available online at: https://futureoflife.org/open-letter/ai-principles/ (Accessed March 04, 2025).

111. Crawford K, Dobbe R, Dryer T, Fried G, Green B, Kaziunas E, et al. *AI Now 2019 Report*. New York, NY: AI Now Institute (2019).

112. Abrassart C, Bengio Y, Chicoisne G, de Marcellis-Warin N, Dilhac M-A, Gambs S, et al. *Montréal Declaration for Responsible Development of Artificial Intelligence*. Montreal: University of Montreal (2018). Available online at: https://umontreal.scholaris.ca/items/bb7296ff-9042-4cf9-a4b7-dfcba8265b53 (Accessed March 03, 2025).

113. OpenAI. *OpenAI Charter*. OpenAI. Available online at: https://openai.com/charter/ (Accessed March 05 2025).

114. Council. ITI. ITI AI policy principles. Available online at: https://www.itic.org/public-policy/ITIAI PolicyPrin ciplesFINA L.pdf (Accessed February 02, 2025).

115. DeepMind G. AI Principles. Available online at: https://ai.google/responsibility/principles/ (Accessed June 06, 2025).

116. Google. Perspectives on issues in AI governance. (2018). Available online at: https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf (Accessed May 03, 2025).

117. Cutler A, Pribić M, Humphrey L, Rossi F, Sekaran A, Spohrer J, et al. *Everyday Ethics for Artificial Intelligence*. Armonk, NY: IBM Corporation (2022).

118. AI Po. About us. Available online at: https://partnershiponai.org/about/ (Accessed July 05, 2025).

119. Raab CD. Information privacy, impact assessment, and the place of ethics. *Comput Law Secur Rev*. (2020) 37:105404. doi: 10.1016/j.clsr.2020.105404

120. Hoffmann AL, Roberts ST, Wolf CT, Wood S. Beyond fairness, accountability, and transparency in the ethics of algorithms: contributions and perspectives from LIS. *Proc Assoc Inform Sci Technol*. (2018) 55(1):694–6. doi: 10.1002/pra2.2018.14505501084

121. Benjamins R. A choices framework for the responsible use of AI. *AI Ethics*. (2021) 1(1):49–53. doi: 10.1007/s43681-020-00012-5

122. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. In: Bohr A, Memarzadeh K, editors. *Artificial Intelligence in Healthcare*. Amsterdam: Elsevier (2020). p. 295–336.

123. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. (2019) 28(3):231–7. doi: 10.1136/bmjqs-2018-008370

124. McConnell T. Moral residue and dilemmas. In: Mason H, editor. *Moral Dilemmas and Moral Theory*. Oxford: Oxford University Press (1996). p. 36–47.

125. Epstein EG, Hamric AB. Moral distress, moral residue, and the crescendo effect. *J Clin Ethics*. (2009) 20(4):330–42. doi: 10.1086/JCE200920406

126. Webster G, Bayliss F. Moral residue. In: Rubin S, Zoloth L, editors. *Margin of Error: The Ethics of Mistakes in the Practice of Medicine*. Hagerstown: MdUniversity Publishing Group (2000). p. 217–30.

127. Makins N. *Essays on Moral Doubt, Meta-Ethics, and Choice*. London: London School of Economics and Political Science (2022).

128. Čartolovni A, Stolt M, Scott PA, Suhonen R. Moral injury in healthcare professionals: a scoping review and discussion. *Nurs Ethics*. (2021) 28(5):590–602. doi: 10.1177/0969733020966776

129. Griffin BJ, Purcell N, Burkman K, Litz BT, Bryan CJ, Schmitz M, et al. Moral injury: an integrative review. *J Trauma Stress*. (2019) 32(3):350–62. doi: 10.1002/jts.22362

130. Shay J. Moral injury. *Psychoanal Psychol*. (2014) 31(2):182. doi: 10.1037/a0036090

131. Morley G, Bradbury-Jones C, Ives J. The moral distress model: an empirically informed guide for moral distress interventions. *J Clin Nurs*. (2022) 31(9-10):1309–26. doi: 10.1111/jocn.15988

132. Lamiani G, Borghi L, Argentero P. When healthcare professionals cannot do the right thing: a systematic review of moral distress and its correlates. *J Health Psychol*. (2017) 22(1):51–67. doi: 10.1177/1359105315595120

133. Grenier S, Barrette A-M, Ladouceur R. Intolerance of uncertainty and intolerance of ambiguity: similarities and differences. *Pers Individ Dif*. (2005) 39(3):593–600. doi: 10.1016/j.paid.2005.02.014

134. Han PK. *Uncertainty in Medicine: a Framework for Tolerance*. Oxford: Oxford University Press (2021).

135. Hillen MA, Gutheil CM, Strout TD, Smets EM, Han PK. Tolerance of uncertainty: conceptual analysis, integrative model, and implications for healthcare. *Soc Sci Med*. (2017) 180:62–75. doi: 10.1016/j.socscimed.2017.03.024

136. Haas M, Stojan JN. Uncertainty about uncertainty tolerance: the elephants in the room. *Med Educ*. (2022) 56(12):1152–4. doi: 10.1111/medu.14926

137. Ilgen JS, Watsjold BK, Regehr G. Is uncertainty tolerance an epiphenomenon? *Med Educ*. (2022) 56(12):1150–2. doi: 10.1111/medu.14938

138. Patel P, Hancock J, Rogers M, Pollard SR. Improving uncertainty tolerance in medical students: a scoping review. *Med Educ*. (2022) 56(12):1163–73. doi: 10.1111/medu.14873

139. Platts-Mills TF, Nagurney JM, Melnick ER. Tolerance of uncertainty and the practice of emergency medicine. *Ann Emerg Med*. (2020) 75(6):715–20. doi: 10.1016/j.annemergmed.2019.10.015

140. Reis-Dennis S, Gerrity MS, Geller G. Tolerance for uncertainty and professional development: a normative analysis. *J Gen Intern Med*. (2021) 36(8):2408–13. doi: 10.1007/s11606-020-06538-y

141. Rosen NO, Ivanova E, Knäuper B. Differentiating intolerance of uncertainty from three related but distinct constructs. *Anxiety, Stress Coping*. (2014) 27(1):55–73. doi: 10.1080/10615806.2013.815743

142. Strout TD, Hillen M, Gutheil C, Anderson E, Hutchinson R, Ward H, et al. Tolerance of uncertainty: a systematic review of health and healthcare-related outcomes. *Patient Educ Couns*. (2018) 101(9):1518–37. doi: 10.1016/j.pec.2018.03.030

143. Brashers DE. Communication and uncertainty management. *J Commun*. (2001) 51(3):477–97. doi: 10.1111/j.1460-2466.2001.tb02892.x

144. Ghosh AK, Joshi S. Tools to manage medical uncertainty. *Diabetes Metab Syndr Clin Res Rev*. (2020) 14(5):1529–33. doi: 10.1016/j.dsx.2020.07.055

145. Han PK, Strout TD, Gutheil C, Germann C, King B, Ofstad E, et al. How physicians manage medical uncertainty: a qualitative study and conceptual taxonomy. *Med Decis Making*. (2021) 41(3):275–91. doi: 10.1177/0272989X21992340

146. Ilgen JS, Teunissen PW, de Bruin AB, Bowen JL, Regehr G. Warning bells: how clinicians leverage their discomfort to manage moments of uncertainty. *Med Educ*. (2021) 55(2):233–41. doi: 10.1111/medu.14304

147. Rains SA, Tukachinsky R. Information seeking in uncertainty management theory: exposure to information about medical uncertainty and information-processing orientation as predictors of uncertainty management success. *J Health Commun*. (2015) 20(11):1275–86. doi: 10.1080/10810730.2015.1018641

148. Van den Bos K. The social psychology of uncertainty management and system justification. *Soc Psychol Bases Ideol Syst Justif*. (2009) 80(6):185–209. doi: 10.1093/acprof:oso/9780195320916.003.008

149. Walker WE, Harremoës P, Rotmans J, van der Sluijs JP, van Asselt MBA, Janssen P, et al. Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integr Assess*. (2003) 4(1):5–17. doi: 10.1076/iaij.4.1.5.16466

150. Leverenz A, Hernandez RA. Uncertainty management strategies in communication about urinary tract infections. *Qual Health Res*. (2023) 33(4):321–33. doi: 10.1177/10497323231156370

151. Maguire P, Faulkner A. Communicate with cancer patients: 2. Handling uncertainty, collusion, and denial. *BMJ Br Med J*. (1988) 297(6654):972. doi: 10.1136/bmj.297.6654.972

152. Stolper E, Van Royen P, Jack E, Uleman J, Olde Rikkert M. Embracing complexity with systems thinking in general practitioners' clinical reasoning helps handling uncertainty. *J Eval Clin Pract*. (2021) 27(5):1175–81. doi: 10.1111/jep.13549

153. Wells G, Williams S, Davies SC. The department of health perspective on handling uncertainties in health sciences. *Philos Trans R Soc A*. (2011) 369(1956):4853–63. doi: 10.1098/rsta.2011.0123

154. Ilgen JS, Eva KW, de Bruin A, Cook DA, Regehr G. Comfort with uncertainty: reframing our conceptions of how clinicians navigate complex clinical situations. *Adv Health Sci Educ*. (2019) 24(4):797–809. doi: 10.1007/s10459-018-9859-5

155. Hofmann B. Moral obligations towards human persons' wellbeing versus their suffering: an analysis of perspectives of moral philosophy. *Health Policy*. (2024) 142:105031. doi: 10.1016/j.healthpol.2024.105031

156. Cantarelli P, Belle N, Belardinelli P. Behavioral public HR: experimental evidence on cognitive biases and debiasing interventions. *Rev Public Pers Adm*. (2018) 40(1):56–81. doi: 10.1177/0734371X1877

157. Scott IA, Soon J, Elshaug AG, Lindner R. Countering cognitive biases in minimising low value care. *Med J Aust*. (2017) 206(9):407–11. (In eng). doi: 10.5694/mja16.00999

158. Bland S. An interactionist approach to cognitive debiasing. *Episteme*. (2020) 19:66–88. doi: 10.1017/epi.2020.9

159. Herman MH. Towards enhancing moral agency through subjective moral debiasing. *APA E, ed* (2020).

160. Ludolph R, Schulz PJ. Debiasing health-related judgments and decision making: a systematic review. *Med Decis Making*. (2018) 38(1):3–13. doi: 10.1177/0272989X17716672

161. Tung A, Melchiorre M. Debiasing and educational interventions in medical diagnosis: a systematic review. *Univ Toronto Med J*. (2023) 100(1):48–57. doi: 10.33137/utmj.v100i1.38937

162. Bigman YE, Wilson D, Arnestad MN, Waytz A, Gray K. Algorithmic discrimination causes less moral outrage than human discrimination. *J Exp Psychol Gen*. (2023) 152(1):4. doi: 10.1037/xge0001250

163. Lima G, Grgić-Hlača N, Cha M. Blaming humans and machines: what shapes people's reactions to algorithmic harm. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). p. 1–26

164. Lima G, Grgić-Hlača N, Cha M. Human perceptions on moral responsibility of AI: a case study in AI-assisted bail decision-making. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021). p. 1–17.

165. Hofmann BM. *Biomarking life. A Pragmatic Approach to Conceptualization of Health and Disease*. Dordrecht: Springer Netherlands (2024). p. 162–8.

166. Hacking I. The looping effects of human kinds. (1995).

167. Hofmann BM. Biases and imperatives in handling medical technology. *Health Policy Technol*. (2019) 8:377–85. (Vitenskapelig artikkel) (In Engelsk). doi: 10.1016/j.hlpt.2019.10.005

168. Hofmann BM. Artificial intelligence – the emperor's new clothes? *Digit Health*. (2024) 10:20552076241287370. (Vitenskapelig artikkel) (In Engelsk). doi: 10.1177/20552076241287370

169. Costanza-Chock S, Raji ID, Buolamwini J. Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022). p. 1571–83

170. Butt J. Analytical study of the world's first EU artificial intelligence (AI) act. *Int J Res Publ Rev*. (2024) 5(3):7343–64. doi: 10.55248/gengpi.5.0324.0914

171. Musch S, Borrelli M, Kerrigan C. The EU AI Act: a comprehensive regulatory framework for ethical AI development. Available online at SSRN 4549248 (2023).

172. Wachter S. Limitations and loopholes in the EU AI act and AI liability directives: what this means for the European union, the United States, and beyond. *Yale J Law Technol*. (2023) 26:671. doi: 10.2139/ssrn.4924553

173. Nwebonyi N, McKay F. Exploring bias risks in artificial intelligence and targeted medicines manufacturing. *BMC Med Ethics*. (2024) 25(1):113. doi: 10.1186/s12910-024-01112-1

174. Pot M, Kieusseyan N, Prainsack B. Not all biases are bad: equitable and inequitable biases in machine learning and radiology. *Insights Imaging*. (2021) 12(1):13. doi: 10.1186/s13244-020-00955-7

175. Ferrario A, Gloeckler S, Biller-Andorno N. AI Knows best? Avoiding the traps of paternalism and other pitfalls of AI-based patient preference prediction. *J Med Ethics*. (2023) 49(3):185–6. doi: 10.1136/jme-2023-108945

176. Ferrario A, Gloeckler S, Biller-Andorno N. Ethics of the algorithmic prediction of goal of care preferences: from theory to practice. *J Med Ethics*. (2023) 49(3):165–74. doi: 10.1136/jme-2022-108371

177. Chouldechova A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data*. (2017) 5(2):153–63. doi: 10.1089/big.2016.0047

178. Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A. Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd acm Sigkdd International Conference on Knowledge Discovery and Data Mining* (2017). p. 797–806

179. Aquino YSJ, Carter SM, Houssami N, Braunack-Mayer A, Win KT, Degeling C, et al. Practical, epistemic and normative implications of algorithmic bias in healthcare artificial intelligence: a qualitative study of multidisciplinary expert perspectives. *J Med Ethics*. (2023) 51:420–8. doi: 10.1136/ jme-2022-108850

180. Reddy S. Algorithmic medicine. In: Reddy S, editor. *Artificial Intelligence*. New York: Productivity Press (2020). p. 1–14.