Fair and equitable AI in biomedical research and healthcare: social science perspectives

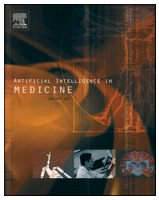
Baumgartner, R.; Arora, P.; Bath, C.; Burljaev, D.; Ciereszko, K.; Custers, B.H.M.; ... ; Williams, R.

Artificial Intelligence In Medicine 144 (2023) 102658

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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



Research paper

Fair and equitable AI in biomedical research and healthcare: Social science perspectives

Renate Baumgartner [a,b,*], Payal Arora [c], Corinna Bath [d], Darja Burljaev [a], Kinga Ciereszko [e], Bart Custers [f], Jin Ding [g], Waltraud Ernst [h], Eduard Fosch-Villaronga [f], Vassilis Galanos [i], Thomas Gremsl [j], Tereza Hendl [k,l], Cordula Kropp [m], Christian Lenk [n,1], Paul Martin [g], Somto Mbelu [o], Sara Morais dos Santos Bruss [p], Karolina Napiwodzka [e], Ewa Nowak [e], Tiara Roxanne [q], Silja Samerski [r], David Schneeberger [s], Karolin Tampe-Mai [m], Katerina Vlantoni [t], Kevin Wiggert [u], Robin Williams [i]

[a] Center of Gender- and Diversity Research, University of Tübingen, Wilhelmstrasse 56, 72074 Tübingen, Germany
[b] Athena Institute, Vrije Universiteit Amsterdam, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands
[c] Erasmus School of Philosophy, Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3062 PA Rotterdam, The Netherlands
[d] Gender, Technology and Mobility, Institute for Flight Guidance, TU Braunschweig, Hermann-Blenk-Str. 27, 38108 Braunschweig, Germany
[e] Department of Philosophy, Adam Mickiewicz University in Poznan, Szamarzewski Street 89C, 60-569 Poznan, Poland
[f] eLaw - Center for Law and Digital Technologies, Leiden University, Steenschuur 25, 2311 ES Leiden, Netherlands
[g] iHuman and Department of Sociological Studies, University of Sheffield, ICOSS, 219 Portobello, Sheffield S1 4DP, United Kingdom
[h] Institute for Women's and Gender Studies, Johannes Kepler University Linz, Altenberger Strasse 69, 4040 Linz, Austria
[i] Science, Technology and Innovation Studies, School of Social and Political Science, University of Edinburgh, Old Surgeons' Hall, High School Yards, Edinburgh EH1 1LZ, United Kingdom
[j] Institute of Ethics and Social Teaching, Faculty of Catholic Theology, University of Graz, Heinrichstraße 78b/2, 8010 Graz, Austria
[k] Professorship for Ethics of Medicine, University of Augsburg, Stenglinstraße 2, 86156 Augsburg, Germany
[l] Institute of Ethics, History and Theory of Medicine, Ludwig-Maximilians-University in Munich, Lessingstr. 2, 80336 Munich, Germany
[m] Center for Interdisciplinary Risk and Innovation Studies (ZIRIUS), University of Stuttgart, Seidenstraße 36, 70174 Stuttgart, Germany
[n] Institute of the History, Philosophy and Ethics of Medicine, Ulm University, Parkstraße 11, 89073 Ulm, Germany
[o] Erasmus School of Philosophy, Erasmus University Rotterdam, 10A Ademola Close off Remi Fani Kayode Street, GRA Ikeja, Lagos, Nigeria
[p] Haus der Kulturen der Welt (HKW), John-Foster-Dulles-Allee 10, 10557 Berlin, Germany
[q] Data & Society Institute, 228 Park Ave S PMB 83075, New York, NY 10003-1502, United States of America
[r] Fachbereich Soziale Arbeit und Gesundheit, Hochschule Emden/Leer, Constantiaplatz 4, 26723 Emden, Germany
[s] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Auenbruggerplatz 2, 8036 Graz, Austria
[t] Department of History and Philosophy of Science, School of Science, National and Kapodistrian University of Athens, Panepistimioupoli, Ilisia, Athens 15771, Greece
[u] Institute of Sociology, Department Sociology of Technology and Innovation, Technical University of Berlin, Fraunhoferstraße 33-36, 10623 Berlin, Germany

ARTICLE INFO

Keywords:
Inequalities
Health equity
Medicine
Discrimination
Bias

ABSTRACT

Artificial intelligence (AI) offers opportunities but also challenges for biomedical research and healthcare. This position paper shares the results of the international conference "Fair medicine and AI" (online 3–5 March 2021). Scholars from science and technology studies (STS), gender studies, and ethics of science and technology formulated opportunities, challenges, and research and development desiderata for AI in healthcare. AI systems and solutions, which are being rapidly developed and applied, may have undesirable and unintended consequences including the risk of perpetuating health inequalities for marginalized groups. Socially robust development and implications of AI in healthcare require urgent investigation. There is a particular dearth of studies in human-AI interaction and how this may best be configured to dependably deliver safe, effective and equitable healthcare. To address these challenges, we need to establish diverse and interdisciplinary teams equipped to develop and apply medical AI in a fair, accountable and transparent manner. We formulate the importance of including social science perspectives in the development of intersectionally beneficent and equitable AI for biomedical research and healthcare, in part by strengthening AI health evaluation.

* Corresponding author.
E-mail address: r.baumgartner@vu.nl (R. Baumgartner).
[1] Sadly, Christian Lenk has passed away after the submission of this paper.

https://doi.org/10.1016/j.artmed.2023.102658
Received 27 May 2022; Received in revised form 30 June 2023; Accepted 1 September 2023
Available online 4 September 2023

1. Introduction

The quest for artificial intelligence (AI) has a long history, which can be traced back to myths of human-like machines and artificers creating moving automata [1,127]. It has been marked by several "summers", times of euphoric activity, and "winters", where the scientific development (seemingly) stagnated. Historically, important foundations for AI were already laid in the 1940s–1970s (e.g., the development of the term AI; the work of Turing on computation or of McCulloch and Pitts on artificial neurons). Practical achievements since the 1970s included knowledge-based or expert systems, which tried to mimic the human reasoning process by building upon human domain-specific expert knowledge. Then in the 1990s, the usefulness of neural networks, which had been developed earlier, was "rediscovered". With the availability of big data (sets), there was a paradigm shift to systems using machine learning and deep learning, i.e., the induction of rules ("models") from large (training) data sets instead of relying on explicit rules programmed by humans [2–4].

These systems are being increasingly adopted across the private sector (e.g. in financial services, manufacturing, farming, engineering, telecommunications, retail, travel, transport and logistics) [5], and in the public sector, e.g. public administration (virtual agents, adaptive delivery of public services, case-management), public transportation (autonomous transportation, predictive maintenance, traffic planning) research and public health [6,7]. Some of the key challenges of AI systems deployment have already been explored and documented, allowing us to extend this documentation to the present agenda setting for healthcare AI. Industry and manufacturing are two key areas that we would still like to address as they have been covered more extensively in literature than those in healthcare, while they also act as entry points to better understand healthcare AI challenges.

Industrial and manufacturing applications of AI face several challenges that extend problematic aspects of computing, information and communication technologies (ICTs), and digital automation more broadly. The challenges of AI in manufacturing industry chiefly revolve around: (1) data quality and availability: AI systems heavily rely on large volumes of high-quality, that is well-defined, cleaned, and clustered, and updated training data [8]; (2) data privacy and security: industrial environments often deal with sensitive data, including proprietary information, trade secrets, and personally identifiable information, potentially vulnerable to unauthorized access, cyber threats, and breaches as well as workplace surveillance [9]; (3) explainability and interpretability are key in industrial applications, especially those involving safety-critical systems or compliance requirements. Deep neural network AI algorithms, can be considered "black boxes" as they provide recommendations without clear explanations [10]; (4) integration complexities: integrating AI into existing industrial systems and workflows can be complex due to legacy IT systems, diverse data formats, and incompatible interfaces [11]; (5) shifting expertise requirements in relation to both industry-specific domain knowledge and AI/technical expertise leading to workforce de- or reskilling [11]; (6) ethical considerations: undesired, perpetuated social bias in industrial applications can have serious consequences, such as discriminatory recruitment or treatment practices (from pay-gaps to parental leaves) or safety risks when AI applications on one industrial domain extend to others (such as military) [12–14]; (7) regulation and compliance: integrating AI technologies may require compliance with existing industry-specific regulations and standards which may be at odds with recent AI-specific regulations which industry must comply with [15]. Addressing these challenges requires collaboration among industry stakeholders, policymakers, researchers, and AI practitioners – we see no reason that these lessons are not applicable to AI in healthcare. For each of these domains and sectors, specific opportunities and challenges, limitations, barriers, shortcomings and risks exist e.g., the impact of AI on employment, privacy concerns including the risk of mass surveillance or of persuasion by tailored information flows, the possibility of biased decision-making, the safety of critical applications like AI systems regulating the water supply and cybersecurity concerns [3,16,17]. Each of these domains and sectors requires an independent analysis. In this paper we focus therefore on a domain which holds a special status: AI in healthcare.

This position paper is the result of the international conference "Fair Medicine and AI: Chances, Challenges, Consequences" hosted by the Center of Gender and Diversity Research of the University of Tübingen (Germany) that took place online on March 3.-5.2021. Participants at the conference included social scientists, ethicists and gender studies scholars. This paper critically synthesizes the key findings of the conference. AI claims to hold considerable opportunities in advancing healthcare in the fields of telemedicine, assessment, and biomedical research, as long as technical and organizational challenges are addressed, including, for example, robust infrastructures to support responsible innovation and effective post-market surveillance [18]. From supporting clinical decision-making and image analysis (e.g., pattern recognition for cancer diagnosis) to assisting with the whole patient lifecycle management (e.g., diagnosis, treatment, and aftercare), algorithmic tools are already being used [19,20]. Much discussed examples can be found in the fields of radiology, pathology, dermatology, ophthalmology, cardiology, mental health, and other sub-disciplines of medicine and tools can be used by healthcare professionals, patients, and others. Certain experts, such as Eric Topol [20], praise AI as a remedy against health-related discrimination. Others warn that it may reproduce and exacerbate existing inequalities and therefore argue that various forms of bias, axes of discrimination and foundational flaws within practical medicine should be addressed and equity through AI should be promoted [21–25]. AI can augment inequalities by overlooking and discriminating against whole population groups, such as women, racialized populations, LGBTQIA+ patients, and people from socioeconomically disadvantaged backgrounds, whose health may not be accurately supported by machine learning (ML) based tools [25–28]. Neural networks or inductionist ML approaches may incorrectly detect risk factors that happen to be associated with demographic disadvantages and falsely attribute health risks to these instead of the unidentified cause. Noisy data and gaps in the evidence base increase the risk of spurious associations and incorrect inferences. AI can leave some people marginalized from private insurance-based care systems by segmenting risks very accurately or by wrongly attributing risks. Indeed, a growing body of research provides evidence of cases in which the implementation of AI and digital technologies in healthcare magnified racial and gender inequalities and generated unequal health outcomes [24,29–33]. An article about AI in healthcare must proceed with some working definition of AI. This is proven to be quite challenging. A variety of AI experts note that AI resists definition for numerous reasons.

AI is not a static or monolithic entity defined by specific attributes, but rather a set of versatile capabilities that can be applied in various contexts. It is an umbrella term encompassing a wide range of evolving tools and techniques, enhanced through iterative cycles of social interaction, technical development, utilization and reinvention by users [34]. Historically, AI was conceived as a field of scientific inquiry studying intelligent behaviour in human and nonhuman animals and machines, exploring whether the latter can be constructed in a way to imitate the former, and whether this accomplishment can shed light into the very concept of intelligence. As a field, it borrows from and contributes to

engineering, computer science, cognitive psychology, linguistics, mathematics, and philosophy, among others [1]. This research field is chiefly operated via digital computational tools, and an array of techniques have been proposed and developed – including the vast majority of computer rule-based languages and operating systems that work on the basis of manipulation of digital symbols. This form of "symbolic" or "rule-based AI" has now been embedded in most problem-solving and heuristic computer programs of today. Parallel to the development of this strand of rule-based AI ("give instructions to a machine"), a statistical, data-based strand was also developed ("let the machine learn from many examples"). While this approach was thought of as potentially useful, yet unfeasible due to lack of sufficient data, long-term acquisition and accumulation of large datasets based on internet user behaviour, governmental and police demographics, industrial and military applications documentation, education assessments, or medical histories, have enabled a resurgence of this "machine learning" approach in the late 2000s [35].

This allowed the assembly of a number of technical configurations, currently broadly understood as AI, to deliver fascinating results within the aforementioned sectors, through applications such as: chatbots and virtual agents customer interaction and entertainment, the reuse of massively produced data for generation of statistically novel arrangements, transcriptions based on natural language processing, predictive analytics applied from future workflow performance to medical diagnosis relating to pattern recognition and insights, emotional and facial categorisation and recognition/detection and advanced biometrics used from phone unlocking to policing, novel encryption methods for information cybersecurity, peer-to-peer networking, and heuristics/problem solving based on multiple examples. As known to AI scholars as the "AI effect," often these applications do not bear the label "AI" if treated in isolation or for less rhetorical or persuasive purposes.

The operational potentials of AI are different from human intelligence and intuitive capacities and expand the classical evidence and experience-based medical knowledge, among other things, by insights based on statistical correlations. In medicine (and elsewhere) AI can be used to identify relationships within large heterogeneous data sets such as published research outcomes or biomedical datasets. This is expected to open up new opportunities for discovering novel treatments with shortened R&D time and reduced costs. Particular salience for AI is anticipated with precision medicine, promising a tailored medical approach to healthcare according to each person's specific genotype and phenotype that can help address drug intolerances as well as group-specific risks and individual differences [36]. Digital health applications and 'digital assistants' promise to support individual monitoring and treatment of diseases, and tailored and healthy lifestyles, but also enable practices of permanent self-monitoring [20,37].

Some scholars have suggested that AI offers opportunities to identify and counteract discrimination directly, e.g., through discrimination-aware data mining, data cleaning designed for fairness, data quality measures, and AI impact assessments [38–44]. However, others have argued that discrimination goes beyond issues of data collection and quality control, i.e., that it is grounded in unequal social structures [45,46]. Hence, if we want to address discrimination in/with AI, anti-discrimination research needs to go beyond mathematical correction and systematically explore so-called biases in data sets. AI and statistical analysis at the same time can be useful tools to shed light on prevailing inequalities [21,24,46]. AI can be used to calculate which social determinants affect individual and public health and disease patterns, thereby aiming to contribute to more suitable tailored treatments and better health. In the context of the discovery and development of new treatments, AI powered drug repurposing can reduce the time and cost of drug development making it economic to treat rare diseases where there has historically been an unmet medical need and poor access to therapy. AI has also recently been applied in the search for drug and vaccine development against COVID-19 [47]. Yet, further concerns remain regarding how to best integrate AI with the broader quest for addressing social inequalities and facilitating equitable health outcomes.

In this position paper, we will first address the challenges and risks surrounding adoption of AI systems, including biased data/models, discrimination/structural injustice and more technical features of AI (its 'black-box' opacity, its apparent objectivity), before moving to ethical and legal challenges and offering ways to advance AI in biomedical research and healthcare from social science perspectives.

2. Challenges and risks of AI systems

2.1. Biased data and models

Inequalities in healthcare have been a major challenge for public health for a long time. The introduction of smart systems has the potential to deliver a range of benefits including improving efficiency and knowledge management and broadening access. New communicative practices and data analytics could be used to make healthcare more patient-centered and through greater citizen involvement and attending to patient experience, building empathy and communicative practices into healthcare [48–51]. However, these systems have not (yet) alleviated inequalities but have indeed created new problems, such as the perpetuation of inequalities due to biased data and defective theoretical models [52]. The misattribution of risks rooted in demographic factors has caused controversies, e.g., in the field of law enforcement. In healthcare, however, the bigger issue is exclusion. Knowledge and decisions in relation to minority and marginalized segments might be less accurate in identifying and mitigating health risks for these groups and thereby exacerbate existing inequalities. For instance, Parikh et al. [53] warn of the risk of falling below professional standards and overlooking the medical needs of diverse and multiethnic populations. Different forms of bias exist. Some distinction between people's needs may even be desirable, as in the case of precision medicine, where categorical information can be relevant for a precise diagnostic [24,43]. However, stochastic data evaluation, drawing on training sets which exclude whole population groups or represent diversity unevenly, risks reproducing undesirable bias within data which can lead to unintended discrimination against groups of people, also called "inequitable bias" [43]. Yet, most algorithms deployed in the healthcare context do not consider these aspects and do not account for bias detection [24,54].

Another crucial problem is posed by inaccurate and imbalanced datasets which better represent some social groups than others. These imbalances result from both existing inequalities in access to healthcare and more generally from differences in access of various demographic groups to those institutions that have digital infrastructures and collect data for datasets. For example, data from middle-class and rather high-income demographics and from affluent countries may be collected and used in datasets more frequently than data from people with lower socioeconomic status and from middle- and low-income countries. Although the exact amount of missing or inaccurate data about socially relevant categories in electronic health records and other records is unclear, we can speak of data gaps such as a "gender data gap" and gaps in information about race and ethnicity within health data [21,29,46]. As a result of these "signal problems", big-data sets are beset by invisible lacunae whereby "some citizens and communities are overlooked or underrepresented" [55]. For this reason, algorithms trained on these

data sets may not accurately detect or treat their health risks.

It is crucially important to take these social differences into account in medical research and healthcare. Empirical data show that the socioeconomic background of patients, such as their profession, education, and income, has a significant impact on healthcare and the treatment of individual patients [56]. Similarly, considering sex and gender in health data is crucial because they affect individuals' health and illness [24]. Characterizing and monitoring the training data sets and the collection of heterogeneous, intersectionally[2] disaggregated and accurate data representing a diverse population are prerequisites for ensuring that AI in healthcare contributes to healthcare equity and justice [30]. However, 'eliminating' bias is itself a challenge as all medical and scientific knowledge is also situated in historical and social settings and therefore needs to be discussed not only by specialists but also interrogated by those affected by claims without medical evidence [57]. The challenge is to train systems in a way that does not compromise the safety and privacy of users,[3] does not perpetuate harm and discrimination, and facilitates benefits for the whole target population [59]. Research is beginning to address these challenges. However, we do not, for example, currently have an effective means of characterizing, let alone standardizing, ethnic diversity in training data sets.

2.2. Structural injustice in medicine and society

While more diverse data is necessary, it should be considered that discrimination, exclusion, prejudice and stereotyping go well beyond the matter of data collection. They are rooted in persistent social inequalities and find their way into data through processing, labelling and classification e.g., when choosing which sub-populations are part of the training data, for which group of people an algorithmic tool should provide the most accurate result, or failing to ensure intersectional benefits across the whole target population [21,29,60]. Therefore, an overarching approach is needed to counteract all types of injustice more efficiently, including structural racism and sexism in medicine as a multifaceted problem (e.g., that patients of color's pain is less likely to be taken seriously and treated with medication or that women's myocardial infarctions remain undiagnosed because their symptoms might be different from "typical" male ones [61,62]). The countering of discrimination through AI platforms is related to questions of equity in the healthcare system. Thus, it needs to be combined with requests for greater justice in medicine, healthcare, the tech industry and society. Urgently research is needed to figure out how to design new technologies and/or transform existing technologies to incorporate intersectional justice and cater for more diverse and inclusive and anti-colonial standpoints. This requires a commitment to a justice-oriented design of AI algorithms and AI-based support systems and independent global oriented auditing overseers as well as workers trained in STS to assist the evaluation of frameworks, datasets, and epistemological decisions given the changing nature of these systems [63]. Such an approach goes beyond solutions that are solely aimed at fixing bias in particular technologies towards strategies that mitigate discriminatory social practices, norms, and structures. Debiasing technologies (often considered as impossible in ML) will not suffice to counter these social and health inequalities and challenges. Based on the above, we need a society that is biased in favor of social justice [64].

2.3. AI as a black box and the multidimensionality of health

AI can process huge amounts of data, going beyond human information processing capabilities. How machines learn is often opaque to outsiders, who may be deliberately excluded from knowledge by intellectual property protection, but also to insiders, with no one clearly holding comprehensive explanatory knowledge about their functioning [65,66]. This black-boxed nature of AI, which arises from both technical circumstances (e.g., the difficulty of establishing why an algorithm trained on particular data sets reached a specific output) and the protection of proprietary models used, makes it hard to audit AI systems and to guarantee a transparent process that can be explained (to allow informed consent for instance), audited (by competent authorities), and traceable (in cases of harm). Although the issue of explainability of algorithmic decisions is currently being addressed, there is a trade-off between the explanatory power and performance of a machine learning model and its ability to produce explainable and interpretable predictions [67–69]. Users of AI systems, including physicians and patients, may have little opportunity to interrogate and challenge the operation of algorithmic systems and their outcomes [70–72]. One possibility is to shift the issue from the much invoked "trust in AI systems" to building accountability ("responsible AI systems"), for example, by introducing post-market surveillance and audits of medical care delivery and outcomes [73–76]. Evidence of increasing accuracy of AI models and their robust performance in real-world care delivery may offset concerns about explainability – as we explore below [77]. However, AI needs algorithms designed in accordance with justice principles that consider the multidimensionality of health which means taking into account physical, mental, emotional, social, spiritual, vocational and other dimensions of health [78,79]. Automatisms which do not account for concrete, specific and individual situations are risky, misleading and contradict ethical guidelines in many countries and will not be trusted, particularly by marginalized groups.

2.4. The presumed objectivity of AI

Issues of reliability and fairness[4] as well as counteractive measures against bias have become hot topics in the computer science community [45,80]. The supposed ideal of the machine's workings as objective, however, is now being questioned more and more [81–83]. Most critical research focuses on distorted data and human error in interpreting data and results [84]. However, a critical academic stance requires us to go beyond naive understandings and claims about the objectivity of scientific facts and performance of technological artifacts [85,86]. Consider, for instance, the belief that, through objective methods we can identify subjective, internal states of users [87]. Gender classification systems classify users as 'male' and 'female' because they use 'sex' as a parameter. However, neither sex nor *gender* is binary [88–90].

STS have demonstrated empirically how technology and society mutually shape each other [86,91]. Technologies are shaped by and risk perpetuating the power structures and social order of their time and context: they fuse immense amounts of information from the past to predict an outcome in the future. The sociotechnical systems of health and AI have a long history of objectivation generated through statistical classification and analysis. Classifications used to provide data for AI are "powerful technologies" that frequently increase social inequality by valorizing hegemonic viewpoints while silencing others [92]. Moreover, the field of AI in medicine and healthcare is embedded within a powerful promissory environment. STS encourages critical interrogation of how expectations and claims are mobilized and may become performative in shaping technical and policy choices [93–95]. This however remains a

[2] The concept of intersectionality describes the ways in which systems of inequality based on gender, race, ethnicity, sexual orientation, gender identity, disability, class and other forms of discrimination "intersect" to create unique dynamics and effects.

[3] AI offers several ethical and legal challenges, especially when considering data that could de-anonymize the data owners. However, some solutions offered by Blockchain technologies could be helpful against tracing and tracking [58].

[4] "Fairness in AI" is understood as the exclusion of discrimination, the protection of patients' rights and interests and the adequate participation in medical progress.

relatively understudied area in terms of medical AI adoption and regulation.

2.5. AI ethics, legal requirements, and approval

Regarding the ethical perspective on medical AI, it has proven a challenge to categorize the primary ethical risks of medical AI. As meta-analysis has shown, "justice and fairness" are often a core part of ethical guidelines, but there is a plethora of different viewpoints on what "fairness" constitutes, reaching from addressing biased data to inclusion and equality, making it harder to find common ground [96].

Several literature reviews have made it easier to identify the main risks: For example, Morley et al. [97] undertook a literature review which identified three main categories of concerns regarding medical AI (epistemic, normative, traceability) at different levels of abstraction (e.g., reaching from the individual to the societal level). These categories (partly) overlap with the challenges discussed above, e.g., "unfair outcomes" is categorized as a normative concern while "misguided evidence", i.e., biased data is seen as an epistemic concern. The black box problem is reflected by (lacking) traceability.

Therefore, meta-analysis and literature reviews have made it easier to find common ground about ethical risks, which has led to a process of concretization. For example, the Ethical Guidelines of the High-Level Expert Group on AI [98], which name fairness as one of the four foundational ethical principles, are a starting point of the proposed Artificial Intelligence Act (hereafter AIA).

Regarding the legal perspective, most current regulatory instruments at the national or European level were not written with AI in mind. Therefore (national) liability regimes or even the relatively new European Medical Devices Regulation (hereafter MDR) or the General Data Protection Regulation are applicable to medical AI systems but they do not sufficiently address the specificities of AI products (for example, the black boxed nature of recommendations can make it hard to prove liability; the performance of medical devices powered by AI on the market can vary between adoption settings) [99–101]. Therefore, even though legal instruments are often designed to be "technology neutral", the current legal framework does not always neatly fit AI systems.

Addressing these deficiencies of the current state of law, the European Commission proposed an "Artificial Intelligence Act" in 2021. This AIA follows a risk-based model: Some practices like social scoring modeled on the example of China are banned (Art. 5 AIA) while the majority of the AIA concerns so-called high-risk-systems (which are defined by a list in Annex III or by reference to other EU legislation; Art. 6 AIA) [102,103]. Medical devices powered by AI will in most cases automatically be considered high-risk. The AIA introduces an approval procedure for high-risk AI systems. AI specific conformity assessment will be integrated into the already existing approval procedures for medical devices required by the MDR (Art. 43 seq. AIA).

The AIA addresses some of the challenges that have been addressed above. Art. 10 AIA concerning data governance requires that datasets "shall have the appropriate statistical properties […] as regards the persons or groups of persons on which the high-risk AI system is intended to be used" and that data sets "shall take into account […] the characteristics or elements that are particular to the specific geographical, behavioural or functional setting". This addresses the problems discussed above regarding biased data sets. Regarding the problem of black boxed recommendations, Art. 13 AIA requires that high-risk systems "shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately" (though we note that mandating such transparency legally is not the same as delivering it technologically/in practice). In addition to AIA, the European Commission also proposed to amend product and civil liability addressing specific risks of AI, especially regarding the problem of the burden of proof and the lack of access to information for opaque, black-boxed solutions [104].

2.6. From explainable AI models to accountable AI-based systems

The critical attention currently being given to the development and deployment of AI in biomedical research and health may contribute to more responsible and accountable innovation and use of AI that may mitigate though never entirely prevent unintended harms [77,105]. This includes, for example, attempts to improve explainability (offset the opacity) of AI models; critical appraisals of training data biases and of the operation of algorithms; risk governance and clinical governance assessments; and, audits and evaluations of performance of AI tools in use.

What is at stake, ultimately, is the performance of AI-based tools in use. There is however a dearth of studies of human-AI interaction in health care delivery even though AI model performance can vary considerably between settings, depending, inter-alia on the integration of AI tools into workflows, including the level of clinical expertise and the distribution of tasks between human and machine intelligences. And above all, how can the different strengths and frailties of human and machine intelligence be most reliably combined?[5]

3. Moving forward

Social science perspectives provide manifold ways to address issues regarding the development and adoption of AI in medicine and healthcare encompassing both the development of the tools and their application, interpretation and risk analysis (Table 1). Interdisciplinary, transdisciplinary, diversified and participatory research regarding the development of AI should be established because *diversity* creates inclusive healthcare systems [107].

Furthermore, research should investigate the underlying and implicit assumptions about medical work held by technology developers and their beliefs, norms and implications regarding diversity, intersectionality, and justice to understand how these may shape the design and implementation of new technologies [108–110]. We also need new training datasets that are more reflective of diverse concerns/issues and auditing institutions moderated by humans and machines that can continuously work at ensuring these systems are checked and steered in the right directions. There is a need for detailed investigations and comparative case studies about the specific application and actual usage of AI in research and health service to understand its potential impacts (including unintended and undesired impacts) on medical decision-making and treatment. Studies need to pay attention to the dynamics within particular settings and to the local aspects of healthcare systems including differences and disadvantaged groups within nations and also international differences (e.g., for Japan see Brucksch and Sasaki [111]) including economically and technologically disadvantaged countries of the North and especially the Global South. Additionally, studies about economic inequalities and exploitative relations between countries and around the world are needed in this regard. This would also allow exploring how expectations from AI in medicine relate to the use of AI in the clinic and its effects on healthcare systems in a comparative manner.

There is a need to identify the stakeholders involved in designing AI systems and healthcare regulation, including policymakers, users, data providers (e.g., patients) non-governmental organizations, civil society organizations, and tech companies responsible for designing these platforms. These stakeholders influence the quality of AI in medicine and healthcare. Therefore, it is crucial to study how they reconfigure health, illness and patients and what criteria they apply to optimize algorithmic decision-making [72]. It is important to better understand how a wide range of direct and indirect users (including various health professionals, patients, carers and others such as procurers and

[5] We note that patient and carer engagement and responses to health AI are even less well-studied, which poses a further set of challenges for research and policy [106].

Table 1
Challenges and proposals for fair and equitable medical AI.
This tentative list captures a set of emerging challenges and potential mitigations that have not yet been systematically categorised. Yet, these are tendencies that can be explored with a variety of sources of expertise as simultaneous points of departure. While the table is provided in a sequential manner, there is no hierarchization of priorities implied. Given our emphasis on the structural nature of the challenges underpinning AI in healthcare, these items are to be simultaneously explored and interrogated. It would be premature to depict how these challenges overlap or interrelate.

Challenges	Proposals
Technical • Undesirably biased datasets and models • The apparent objectivity of AI	• Ethnographic investigation of underlying and implicit assumptions on behalf of stakeholders • Attending to performance of 'objectivity' with interrogation of epistemic and normative competence to verify decisions, evidence and reasons that justify decisions • New training datasets that are more reflective of diverse concerns/issues and auditing institutions, e.g., semi-synthetic equity oriented datasets
Trustworthiness • AI as a black box, and the multidimensionality of health	• Close relationship between technology-providers and users, especially actors of chronically discriminated communities, which should be engaged from an early stage of research and development to gauge impact and establish feedback loops with tech providers • Development of robust AI/data platforms, with easy-to-operate toolkits • Implementation of educational opportunities for medical users
Siloing • Lack of a diverse body of interdisciplinary, transdisciplinary research	• Investments in institutional building that legitimate stakeholder collaboration with effective incentives, and knowledge capture • Innovative collaboration including patient and user groups as well as industry, regulators and clinicians
Biographical and multi-level • Understand the entire lifecycle of AI in healthcare, its extension and longer-term evolution over time and across multiple settings	• Detailed stakeholder identification, from technical designers and vendors to data providers (patients) • Investigation of local/situated aspects of healthcare systems at national and international level • Building of synergies across levels (sectoral, state, and other nodal points of relevance) • Understanding meso- and macro-deployment of algorithms and data-driven technology
Foresight • Mitigation of unintended and undesired impacts of AI in healthcare	• Comparative case studies on specific application and actual usage of AI in research and health service
Inclusion and equity • Structural injustice in medicine and society	• Understanding how AI may exacerbate the vulnerability of oppressed/misrepresented populations and how this discrimination works • Intersectional research attending to vulnerable populations, and power asymmetries at local level • Investigation at international level of economic inequalities and exploitative relations between countries
Governance, ethical, legal • Role of policymakers and non-governmental civil organizations in the shaping of AI	• Reflective process required from all researchers involved, disclosing and interrogating underlying assumptions not only of technical designers, clinicians, and data providers, but also policymakers, NGOs, and social scientists involved in the shaping of AI in healthcare • Moving from explainable AI models to accountable AI-based systems

regulators), can be sufficiently well-educated and informed about the functioning of those tools to critically evaluate their effectiveness and recommendations. More research is needed to investigate what roles non-governmental and civil society organizations, alongside formal regulatory bodies, can play in strengthening the development of AI in medicine and healthcare systems.

Research has shown that AI can be fundamentally shaped by social power asymmetries and inequalities and, hence, generate unequal outcomes and effects [21,29,31,45,46]. It is important to better understand ways in which AI may exacerbate the vulnerability of situationally oppressed/misrepresented human populations, how discrimination due to a representation gap (e.g., women, LGBTQIA+ people, racialized people, differently abled people and people of different ages) works and how the gap is related to individual and systemic discrimination against specific population groups [33,46,112–114]. More research is needed to better understand which sub-populations are most at risk of harm, why and how they are made vulnerable and with which consequences. Simultaneously, it is important to act on the growing pool of research that is already available [29,30,32,46,113–115]. We should investigate what the most effective short and long-term strategies to address the risk of harm may be and what are appropriate strategies to ensure health justice for members of persistently oppressed and misrepresented populations. It is crucial to identify how we can ensure equitable and just health benefits while respecting and supporting the agency of vulnerable populations [115–117]. Considering that the roots of discrimination through AI are structural, what are the most viable systemic solutions in design and implementation of AI tools? And how can we strive for equity and at the same time keep social categories open and flexible for change?

Of particular interest is how the move towards AI-based decision support systems changes the production and application of knowledge about disease and health of practitioners and patients [118,119]. These questions require ethnographic research to explore the micro-decisions made by computer and data scientists as well as misleading and scientifically unsubstantiated assumptions and expectations of medical professionals and patients as well. Attention should focus further on the hidden role of social meso- and macro structures embedded in the deployment of algorithms and data-driven models. Empirical research is required on ways in which 'objectivity' is performed through socio-technical practices in the context of AI-based health apps and how agency and responsibility are re-distributed and accounted for in those human-machine interactions. It is important to better identify which actors and factors should be considered to make transparent, accurate, scientifically valid and ethically justifiable decisions about health and illness/classifications and how these decisions are constructed and justified. Who or what is given epistemic and normative competence to verify decisions, evidence and reasons that justify decisions?

Critical issues around trust and trustworthiness also need to be addressed, such as how direct or indirect users can be assured that AI technology in healthcare will be safe, effective, privacy respecting, ethically governed and employed for social good [46,73,120,121]. One way to improve the trustworthiness and usability of AI solutions might be to build a close relationship between technology-providers and users and engage them from an early stage of research and development

[122,123]. An alternative route might involve developing robust AI/data platforms, with easy-to-operate toolkits that clinicians can deploy and demonstrate their robustness in use. Another approach could be to implement educational opportunities for medical users of AI-based technologies to provide knowledge of how they work, how to handle results, and to enable them with informed assessments of limitations. To maximize the potential benefits of AI, new forms of innovative collaboration including patient and user groups as well as industry, regulators and clinicians will need to be created. Innovative public policy can help support such initiatives. AI in medicine also requires greater responsibilities and more sophisticated understanding of persistent social inequalities and medical knowledge on the part of the developing engineers. At the same time, AI-based decision-making/recommendation processes must be understood by humans making medical decisions, as recommended, for example, in earlier or recent assessments of AI in radiology [124]. Consequently, digital competences in dealing with digital health data, automated diagnoses and AI-based therapy suggestions need to be improved on all levels. Critical knowledge about AI must be anchored in educational concepts, otherwise we risk splitting society into digital literate/informed and uninformed citizens [74]. Beyond that, developing AI for health equality demands interdisciplinary and diverse teams that integrate – under just working conditions – members of historically marginalized populations groups and their perspectives [46,114,125,126]. Social science research is not immune from the risk of bias and reinforcement of dominant social hierarchies and inequalities. Therefore, a reflective process is required from all researchers involved, disclosing and interrogating underlying assumptions, power inequalities and intersectional impacts of AI with respect to gender, racial inequalities, class and other relevant axes of discrimination. This reflective and awareness raising process is crucial to prevent the embedding of inequalities into AI training, design and implementation models, thereby magnifying discriminatory effects through the adoption of autonomous systems [33]. Such critical research simultaneously opens up opportunities to clarify the criteria for equity in health research more generally, which have been ill-defined to date, as well as questions about the patterns and manifestations of inequalities in concrete socio-political contexts, and to find effective responses to mitigate them. It is important to consider in each situation who profits and benefits from the research and for which purposes the research is used. The ultimate goal should be to develop win-win models of human and machine intelligence that maximize benefits of both having intersectional justice in mind and being used for social good.

Approaches from the social sciences, gender studies, critical race and data studies, STS and medical ethics show the risk of a perpetuation of medical and social inequalities by current AI algorithms and solutions. The conference has identified numerous problem cases of this type and voiced an emerging need of clinical post-market surveillance of AI supported decisions. This calls for joined scientific and public monitoring, deliberation and seeking normative tools (e.g., in terms of medical AI humanities, ELSI projects and citizen science) to deal with disparities and risks generated by AI technologies applied in biomedical research and healthcare. These findings and concerns need to be disseminated to relevant stakeholders within the computer sciences in tech industry and academia, funding bodies, healthcare professional and communities of practice, who are becoming increasingly aware of the potential for autonomous systems to reinforce existing contours of inequality and who are keen to explore ways to monitor and mitigate. The findings are being disseminated in collaborations with stakeholders, through conferences, and further publications.

Declaration of competing interest

None.

Acknowledgements

Funding: This work was supported by the Wellcome Trust [grant number 219875/Z/19/Z]; the BMBF [grant number FKZ 01GP1791]; acatech NATIONAL ACADEMY OF SCIENCE AND ENGINEERING and Körber Stiftung; the FWF [project P-32554 “A reference model of explainable Artificial Intelligence for the Medical Domain”]; the United Kingdom Research and Innovation: Trusted Autonomous Systems Programme [grant number EP/V026607/1]. EFV would like to acknowledge that this collaborative paper is part of the Safe and Sound project, a project that has received funding from the European Union’s Horizon-ERC program Grant Agreement No. 101076929. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] Nilsson NJ. The quest for artificial intelligence: a history of ideas and achievements. New York: Cambridge University Press; 2010.
[2] Joint Research Center AI Watch. Historical evolution of artificial intelligence: analysis of the three main paradigm shifts in AI. https://op.europa.eu/en/publication-detail/-/publication/6264ac29-2d3a-11eb-b27b-01aa75ed71a1/language-en; 2020.
[3] Russell S, Norvig P. Artificial intelligence: a modern approach. 4th Ed. Global ed. Harlow: Pearson; 2021.
[4] Shortliffe EH. Artificial intelligence in medicine: weighing the accomplishments, hype, and promise. Yearb Med Inform 2019;28(1):257–62. https://doi.org/10.1055/s-0039-1677891.
[5] Ebers M, Standardizing AI. In: DiMatteo LA, Poncibò C, Cannarsa M, editors. The Cambridge handbook of artificial intelligence. Cambridge/New York/Port Melbourne/New Delhi/Singapore: Cambridge University Press; 2022. p. 321–44.
[6] Microsoft EY. Artificial intelligence in the public sector: European outlook for 2020 and beyond. https://info.microsoft.com/rs/157-GQE-382/images/EN-CNTNT-eBook-artificial-SRGCM3835.pdf; 2020.
[7] Joint Research Center AI Watch. Artificial intelligence in public services: overview of the use and impact of AI in public services in the EU. https://publications.jrc.ec.europa.eu/repository/handle/JRC120399; 2020.
[8] Kotliar DM. The return of the social: algorithmic identity in an age of symbolic demise. New Media Soc 2020;22(7):1152–67. https://doi.org/10.1177/1461444820912535.
[9] Krzywdzinski M, Pfeiffer S, Evers M, Gerber C. Measuring work and workers: wearables and digital assistance systems in manufacturing and logistics. Berlin: Wissenschaftszentrum Berlin für Sozialforschung; 2022. PID: http://hdl.handle.net/10419/251912.
[10] Mezgár I, Váncza J. From ethics to standards: a path via responsible AI to cyber-physical production systems. Annu Rev Control 2022;53:391–404. https://doi.org/10.1016/j.arcontrol.2022.04.002.
[11] Belloc F, Burdin G, Cattani L, Ellis W, Landini F. Coevolution of job automation risk and workplace governance. Res Policy 2022;51(3):104441. https://doi.org/10.1016/j.respol.2021.104441.
[12] Damioli G, Van Roy V, Vertesy D, Vivarelli M. AI technologies and employment: micro evidence from the supply side. Appl Econ Lett 2022;30(6):816–21. https://doi.org/10.1080/13504851.2021.2024129.
[13] Goyal A, Aneja R. Artificial intelligence and income inequality: do technological changes and worker’s position matter? J Public Aff 2020;20(4):e2326. https://doi.org/10.1002/pa.2326.
[14] Kim J, Heo W. Artificial intelligence video interviewing for employment: perspectives from applicants, companies, developer and academicians. Inf Technol People 2021;35(3):861–78. https://doi.org/10.1108/itp-04-2019-0173.
[15] Shneiderman B. Human-centered AI. Oxford: Oxford University Press; 2022.
[16] Soleimani M, Intezari A, Taskin N, Pauleen D. Cognitive biases in developing biased Artificial Intelligence recruitment system. In: Proceedings of the 54th Hawaii international conference on system sciences; 2021. p. 5091–9.
[17] Soleimani M. Developing unbiased artificial intelligence in recruitment and selection: a processual framework: a dissertation presented in partial fulfilment of the requirements for the degree of doctor of philosophy in management at Massey University, Albany, Auckland, New Zealand. Albany/Auckland: Massey University; 2022.
[18] Panch T, Mattie H, Celi LA. The “inconvenient truth” about AI in healthcare. NPJ Digit Med 2019;2(77):1–3. https://doi.org/10.1038/s41746-019-0155-4. PMID: 31453372; PMCID: PMC6697674.
[19] Garvin E. Ethical concerns of AI in healthcare: Can AI do more harm than good?. https://hitconsultant.net/2019/08/06/ethical-concerns-of-ai-in-healthcare-can-ai-do-more-harm-than-good/#.YPQQtOhKg2w/; 2019.
[20] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44–56. https://doi.org/10.1038/s41591-018-0300-7.

[21] Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018;178(11):1544–7. https://doi.org/10.1001/jamainternmed.2018.3763.

[22] Nordling L. Mind the gap. Nature 2019;573:103–5. https://doi.org/10.1038/d41586-019-02872-2.

[23] Straw I. The automation of bias in medical Artificial Intelligence (AI): decoding the past to create a better future. Artif Intell Med 2020:110. https://doi.org/10.1016/j.artmed.2020.101965.

[24] Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digit Med 2020;3(81):1–10. https://doi.org/10.1038/s41746-020-0288-5.

[25] Fosch-Villaronga E, Drukarch H, Khanna P, Verhoef T, Custers B. Accounting for diversity in AI for medicine. Comput Law & Secur Rev 2022;47:105735.

[26] Barbee H, Deal C, Gonzales G. Anti-transgender legislation—a public health concern for transgender youth. JAMA Pediatr 2022;176(2):125–6.

[27] Nielsen MW, Stefanick ML, Peragine D, Neilands TB, Ioannidis J, Pilote L, et al. Gender-related variables for health research. Biol Sex Differ 2021;12(23):1–16.

[28] Baumgartner R, Ernst W. Künstliche Intelligenz in der Medizin? Intersektionale queerfeministische Kritik und Orientierung. Gender 2023;1:11–25.

[29] Perez CC. Invisible women: exposing data bias in a world designed for men. New York: Abrams Press; 2019.

[30] Figueroa CA, Luo T, Aguilera A, Lyles CR. The need for feminist intersectionality in digital health. Lancet Digit Health 2021;3(8):e526–33. https://doi.org/10.1016/S2589-7500(21)00118-7.

[31] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366:447–53. https://doi.org/10.1126/science.aax2342.

[32] Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial bias in pulse oximetry measurement. N Engl J Med 2020;383(25):2477–8. https://doi.org/10.1056/nejmc2029240.

[33] Ledford H. Millions of black people affected by racial bias in health-care algorithms. Nature. 2019;574:7780.

[34] Williams R. European perspectives on the anticipatory governance of AI. In: Shi Q, editor. AI Governance 2019: a Year in Review: Observations of 50 Global Experts. Shanghai: Institute for Science of Science; 2019. p. 27–8. https://www.aigovernancereview.com/static/AI-Governance-in-2019-7795369fd451da49ae4471ce9d648a45.pdf. [Accessed 29 June 2023].

[35] High-Level Expert Group on Artificial Intelligence (HLEGAI). A definition of AI: main capabilities and scientific disciplines. Brussels: European Commission; 2019. https://www.aepd.es/sites/default/files/2019-09/ai-definition.pdf [accessed 29 June 2023].

[36] Baumgartner R. Precision medicine and digital phenotyping: digital medicine's way from more data to better health? Big Data Soc. 2021;8(2):1–12. https://doi.org/10.1177/20539517211066452.

[37] Swan M. The quantified self: fundamental disruption in big data science and biological discovery. Big Data 2013;1(2):85–99. https://doi.org/10.1089/big.2012.0002.

[38] Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. ACM Comput Surv 2009;41(3):1–52. https://doi.org/10.1145/1541880.1541883.

[39] Custers BHM, Calders T, Schermer B, Zarsky T. Discrimination and privacy in the information society: data mining and profiling in large databases. Heidelberg: Springer; 2013.

[40] Kiourtis A, Mavrogiorgou A, Manias G, Kyriazis D. Ontology-driven data cleaning towards lossless data compression. In: Séroussi B, et al., editors. Challenges of trustable AI and added-value on health. IOS Press; 2022. p. 421–2.

[41] Mavrogiorgou A, Kiourtis A, Manias G, Kyriazis D. Adjustable data cleaning towards extracting statistical information. In: Mantas J, et al., editors. Public health and informatics. IOS Press; 2021. p. 1013–4.

[42] Pedreshi D, Ruggieri S, Turini F. Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining; 2008. p. 560–8. https://doi.org/10.1145/1401890.1401959.

[43] Pot M, Kieusseyan N, Prainsack B. Not all biases are bad: equitable and inequitable biases in machine learning and radiology. Insights Imaging 2021;12(13):1–10. https://doi.org/10.1186/s13244-020-00955-7.

[44] Tae KH, Roh Y, Oh YH, Kim H, Whang SE. 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. https://doi.org/10.1145/3329486.3329493.

[45] Holzmeyer C. Beyond 'AI for Social Good' (AI4SG): social transformations—not tech-fixes—for health equity. Interdiscip Sci Rev 2021;46(1–2):94–125. https://doi.org/10.1080/03080188.2020.1840221.

[46] Benjamin R. Race after technology: abolitionist tools for the new Jim code. Cambridge: Polity Press; 2019.

[47] Arshadi KA, Webb J, Salem M, Cruz E, Calad-Thomson S, Ghadirian, et al. Artificial intelligence for COVID-19 drug discovery and vaccine development. Front Artif Intell 2020;3:4–13. https://doi.org/10.3389/frai.2020.00065.

[48] Ceccaroni L, Bibby J, Roger E, Flemons P, Michael K, Fagan L, et al. Opportunities and risks for citizen science in the age of artificial intelligence. Citiz Sci: Theory Pract 2019;4(1).

[49] Wiggins A, Wilbanks J. The rise of citizen science in health and biomedical research. Am J Bioeth 2019;19(8):3–14.

[50] Insel TR. How algorithms could bring empathy back to medicine. Nature. 2019;567(7747):172–4.

[51] Alabdulatif A, Khalil I, Saidur Rahman M. Security of blockchain and AI-empowered smart healthcare: application-based analysis. Appl Sci 2022;12(21):11039.

[52] Hagendorff T, Wezel K. 15 challenges for AI: or what AI (currently) can't do. AI & Soc. 2020;35:355–65. https://doi.org/10.1007/s00146-019-00886-y.

[53] Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. Algorithms must meet regulatory standards of clinical benefit. Science. 2019;363(6429):810–2. https://doi.org/10.1126/science.aaw0029.

[54] Cabitza F, Ciucci D, Rasoini R. A giant with feet of clay: on the validity of the data that feed machine learning in medicine? In: Cabitza F, Magni M, Batini C, editors. Organizing for the digital world: lecture notes in information systems and organisation. Cham: Springer; 2018. p. 121–36.

[55] Crawford K. Think again: big data. https://foreignpolicy.com/2013/05/10/think-again-big-data/. [Accessed 9 May 2023].

[56] WHO Health Commission. Final report of the CSDH. Closing the gap in a generation: health equity through action on the social determinants of health. Geneva: World Health Organization; 2008.

[57] Haraway D. Situated knowledges: the science question in feminism and the privilege of partial perspective. Fem Stud 1988;14(3):575–99. https://doi.org/10.2307/3178066.

[58] Azaria A, Ekblaw A, Vieira T, Lippman A. MedRec: using blockchain for medical data access and permission management. In: 2016 2nd international conference on open and big data. Vienna, Austria: IEEE; 2016. p. 25–30. https://doi.org/10.1109/OBD.2016.11.

[59] Neyland D. Bearing account-able witness to the ethical algorithmic system. Sci Technol Hum Values 2016;41(1):50–76. https://doi.org/10.1177/0162243915598056.

[60] Baumgartner R. Künstliche Intelligenz in der Medizin: Diskriminierung oder Fairness? In: Bauer G, Kechaja M, Engelmann S, Haug L, editors. Diskriminierung und Antidiskriminierung: Beiträge aus Wissenschaft und Praxis. Bielefeld: transcript; 2021. p. 147–62. https://doi.org/10.1515/9783839450819-009.

[61] Meghani SH, Byun E, Gallagher RM. Time to take stock: a meta-analysis and systematic review of analgesic treatment disparities for pain in the United States. Pain Med 2012;13(2):150–74. https://doi.org/10.1111/j.1526-4637.2011.01310.x.

[62] Mehta LM, Beckie TM, DeVon HA, Grines CL, Krumholz HM, Johnson MN, et al. Acute myocardial infarction in women: a scientific statement from the American Heart Association. Circulation. 2016;133(9):916–47. https://doi.org/10.1161/cir.0000000000000351.

[63] Domínguez Hernández A, Galanos V. A toolkit of dilemmas: beyond debiasing and fairness formulas for responsible AI/ML. In: IEEE International Symposium on Technology and Society; 2022. https://doi.org/10.1109/ISTAS55053.2022.10227133.

[64] Mitchell TM. The need for biases in learning generalizations. New Jersey: Department of Computer Science, Laboratory for Computer Science Research; 1980.

[65] Felzmann H, Fosch-Villaronga E, Lutz C, Tamò-Larrieux A. Towards transparency by design for artificial intelligence. Sci Eng Ethics 2020;26:3333–61. https://doi.org/10.1007/s11948-020-00276-4.

[66] Quinn TP, Jacobs S, Senadeera M, Le V, Coghlan S. The three ghosts of medical AI: can the black-box present deliver? Artif Intell Med 2022:124. https://doi.org/10.48550/arXiv.2012.06000.

[67] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy. 2021;23(1):1–45. https://doi.org/10.3390/e23010018.

[68] Molnar C. Interpretable machine learning. A guide for making black box models interpretable. https://christophm.github.io/interpretable-ml-book/. [Accessed 9 May 2023].

[69] Ursin F, Lindner F, Ropinski T, Salloch S, Timmermann C. Levels of explicability for medical artificial intelligence: what do we normatively need and what can we technically reach? Ethik Med 2023;35(2):173–99.

[70] MacKenzie D. The certainty trough. In: Williams R, Faulkner W, Fleck J, editors. Exploring expertise: issues and perspectives. London: Palgrave Macmillan; 1998. p. 325–9. https://doi.org/10.1007/978-1-349-13693-3_15.

[71] Watson D, Bruce IN, McInnes IB, Floridi L. Clinical applications of machine learning algorithms: beyond the black box. BMJ. 2019;364:1886. https://doi.org/10.1136/bmj.l886.

[72] Acatech, Körber-Stiftung, ZIRIUS TechnikRadar. Zukunft der Gesundheit. Stakeholderperspektiven; 2021. https://www.zirius.uni-stuttgart.de/dokumente/Langfassung-TechnikRadar-2021-Einzelseiten.pdf. [Accessed 9 May 2023].

[73] High-level expert group on artificial intelligence (HLEGAI). The assessment list for trustworthy artificial intelligence (ALTAI). https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment; 2020.

[74] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17(195):1–9. https://doi.org/10.1186/s12916-019-1426-2.

[75] Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. Lancet Digit Health 2022;4(5):E384–97.

[76] Sujan M, Smith-Frazer C, Malamateniou C, Connor J, Gardner A, Unsworth H, et al. Validation framework for the use of AI in healthcare: overview of the new British standard BS30440. BMJ Health Care Inform 2023;30:e100749.

[77] Pierce RL, Van Biesen W, Van Cauwenberge D, Decruyenaere J, Sterckx S. Explainability in medicine in an era of AI-based clinical decision support systems. Front Genet 2022:13. https://doi.org/10.3389/fgene.2022.903600.

[78] Eberst RM. Defining health: a multidimensional model. J Sch Health 1984;54(3):99–104.
[79] La Fors K, Custers BHM, Keymolen E. Reassessing values for emerging big data technologies: integrating design-based and application-based approaches. Ethics Inf Technol 2019;21:209–26. https://doi.org/10.1007/s10676-019-09503-4.
[80] ACM conference on fairness, accountability, and transparency (ACM FAccT). https://facctconference.org/index.html; 2022.
[81] O'Neil C. Weapons of math destruction: how big data increases inequality and threatens democracy. New York: Crown; 2016.
[82] Moreau JT, Baillet S, Dudley RW. Biased intelligence: on the subjectivity of digital objectivity. BMJ Health Care Inform 2020;27(3):e100146. https://doi.org/10.1136/bmjhci-2020-100146.
[83] Beaulieu A, Leonelli S. Data and society: a critical introduction. Los Angeles, CA: SAGE; 2021.
[84] Zweig K, Fischer S, Lischka K. Wo Maschinen irren können. In: Fehlerquellen und Verantwortlichkeiten in Prozessen algorithmischer Entscheidungsfindung. Kaiserslautern: Bertelsmann Stiftung; 2018. https://doi.org/10.11586/2018006.
[85] Bath C, Meißner H, Trinkaus S, Völker S. Verantwortung und Un/Verfügbarkeit: impulse und Zugänge eines (neo)materialistischen Feminismus. Münster: Westfälisches Dampfboot; 2017.
[86] Gillespie T. The relevance of algorithms. In: Gillespie T, Boczkowski PJ, Foot KA, editors. Media Technologies: Essays on Communication, Materiality, and Society. Cambridge: The MIT Press; 2014. p. 167–94. https://doi.org/10.7551/mitpress/9780262525374.001.0001.
[87] Fosch-Villaronga E. "I love you," said the robot. Boundaries of the use of emotions in human-robot interaction. In: Ayanoglu H, Duarte E, editors. Emotional design in human robot interaction: theory, methods, and application. Springer, Human-Computer Interaction Series; 2019. p. 93–110. https://doi.org/10.1007/978-3-319-96722-6_6.
[88] Fausto-Sterling A. Sexing the body: gender politics and the construction of sexuality. New York: Basic Books; 2018.
[89] Ainsworth C, Sex redefined.. The idea of two sexes is simplistic. Biologists now think there is a wider spectrum than that. Nature. 2015;518(7539):288–91. https://doi.org/10.1038/518288a.
[90] Fosch-Villaronga E, Poulsen A, Søraa RA, Custers BHM. A little bird told me your gender: gender inferences in social media. Inf Process Manag 2021;58:102541. https://doi.org/10.1016/j.ipm.2021.102541.
[91] MacKenzie D, Wajcman J. The social shaping of technology. Buckingham: Open University Press; 1999.
[92] Bowker GC, Star SL. Sorting things out. Classification and its consequences. Cambridge: The MIT Press; 2000. https://doi.org/10.7551/mitpress/6352.001.0001.
[93] Pollock N, Williams R. The business of expectations: how promissory organizations shape technology and innovation. Soc Stud Sci 2010;40(4):525–48. https://doi.org/10.1177/0306312710362275.
[94] Sismondo S. Ghost-managed medicine: big pharma's invisible hands. Manchester: Mattering Press; 2018. https://doi.org/10.28938/9780995527775.
[95] Van Lente H. Navigating foresight in a sea of expectations: lessons from the sociology of expectations. Technol Anal Strateg Manag 2012;24(8):769–82. https://doi.org/10.1080/09537325.2012.715478.
[96] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell 2019;1:389–99.
[97] Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. Soc Sci & Med 2020;260:1–14. https://doi.org/10.1016/j.socscimed.2020.113172.
[98] High-level expert group on artificial intelligence ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai; 2019 [Accessed 9 May 2023].
[99] Schneeberger D, Stöger K, Holzinger A. The European legal framework for medical AI. In: Holzinger A, Kieseberg P, Tjoa AM, Weippl E, editors. CD-MAKE 2020: machine learning and knowledge extraction. Cham: Springer; 2020. p. 209–26. https://doi.org/10.1007/978-3-030-57321-8_12.
[100] Jabri S. Artificial intelligence and healthcare: products and procedures. In: Wischmeyer T, Rademacher T, editors. Regulating artificial intelligence. Cham: Springer; 2020. p. 307–35.
[101] Molnár-Gábor F. Artificial intelligence in healthcare: doctors, patients and liabilities. In: Wischmeyer T, Rademacher T, editors. Regulating artificial intelligence. Cham: Springer; 2020. p. 337–60.
[102] Ebers M, Hoch VRS, Rosenkranz F, Ruschemeier H, Steinrötter B. The European Commission's proposal for an artificial intelligence act: a critical assessment by members of the robotics and AI law society (RAILS). J ;4:589–603. https://doi.org/10.3390/j4040043.
[103] Veale M, Zuiderveen Borgesius F. Demystifying the Draft EU Artificial Intelligence Act: analysing the good, the bad, and the unclear elements of the proposed approach. Comput Law Rev Int 2021;22(4):97–112. https://doi.org/10.9785/cri-2021-220402.
[104] Hacker P. The European AI liability directives: critique of a half-hearted approach and lessons for the future. https://arxiv.org/abs/2211.13960. [Accessed 9 May 2023].
[105] Kerasidou C, Kerasidou A, Buscher M, Wilkinson S. Before and beyond trust: reliance in medical AI. J Med Ethics 2022;48(11):852–6.
[106] Collins PH, Bilge S. Intersectionality. Medford: Polity Press; 2020.
[107] Rock D, Grant H. Why diverse teams are smarter. Harv Bus Rev 2016;4:2–5.
[108] Weingarten R. Die Aushandlung von Praktiken: Kommunikation zwischen Fachexperten und Medieningenieuren. In: Rammert W, Schlese M, Wagner G, Wehner J, Weingarten R, editors. Wissensmaschinen. Soziale Konstruktion eines technischen Mediums. Das Beispiel Expertensysteme. Frankfurt/New York: Campus; 1998. p. 129–88.
[109] Wiggert K. The role of scenarios in scripting (the use of) medical technology. The case of data-driven clinical decision support systems. Berlin: Institutional Repository DepositOnce; 2021. https://doi.org/10.14279/depositonce-11441.
[110] Hyysalo S. Health technology development and use: from practice-bound imagination to evolving impacts. New York: Routledge; 2010. https://doi.org/10.4324/9780203849156.
[111] Brucksch S, Sasaki K. Humans and devices in medical contexts. Case studies from Japan. Singapore: Springer Verlag; 2021. https://doi.org/10.1007/978-981-33-6280-2.
[112] Cave S, Dihal K. The whiteness of AI. Philos Technol 2020;33:685–703. https://doi.org/10.1007/s13347-020-00415-6.
[113] Costanza-Chock S. Design justice - community-led practices to build the worlds we need. Cambridge: The MIT Press; 2020. https://doi.org/10.7551/mitpress/12255.001.0001.
[114] Roxanne T. Digital territory, digital flesh: decoding the indigenous body. APRJA 2019;8(1):70–80. https://doi.org/10.7146/aprja.v8i1.115416.
[115] Carbonell V, Liao SY. Materializing systemic racism, materializing health disparities. Am J Bioeth 2021;21(9):16–8. https://doi.org/10.1080/15265161.2021.1952339.
[116] Chung R. Structural health vulnerability: health inequalities, structural and epistemic injustice. J Soc Philos 2021;52(2):201–16. https://doi.org/10.1111/josp.12393.
[117] Hendl T, Roxanne T. Digital surveillance in a pandemic response: what bioethics ought to learn from indigenous perspectives. Bioethics 2022;36(3):305–12. https://doi.org/10.1111/bioe.13013.
[118] Kaplan B. Objectification and negotiation in interpreting clinical images: implications for computer-based patient records. Artif Intell Med 1995;7(5):439–54. https://doi.org/10.1016/0933-3657(95)00014-w.
[119] Stefanelli M. The socio-organizational age of artificial intelligence in medicine. Artif Intell Med 2001;23(1):25–47. https://doi.org/10.1016/s0933-3657(01)00074-4.
[120] Mhlambi S. From rationality to relationality: Ubuntu as an ethical and human rights framework for artificial intelligence governance. https://carrcenter.hks.harvard.edu/publications/rationality-relationality-ubuntu-ethical-and-human-rights-framework-artificial. [Accessed 9 May 2023].
[121] Chun WHK. Discriminating data: correlation, neighborhoods, and the new politics of recognition. Cambridge: The MIT Press; 2021.
[122] Martinho A, Kroesen M, Chorus C. A healthy debate: exploring the views of medical doctors on the ethics of artificial intelligence. Artif Intell Med 2021:121. https://doi.org/10.1016/j.artmed.2021.102190.
[123] Korb W, Geißler N, Strauß G. Solving challenges in inter- and trans-disciplinary working teams: lessons from the surgical technology field. Artif Intell Med 2015;63(3):209–19. https://doi.org/10.1016/j.artmed.2015.02.001.
[124] Tang A, Tam R, Cadrin-Chênevert A, Guest W, Chong J, Barfett J, et al. Canadian association of radiologists white paper on artificial intelligence in radiology. Can Assoc Radiol J 2018;69(2):120–35. https://doi.org/10.1016/j.carj.2018.02.002.
[125] Roxanne T. Refusing re-presentation. In: Wenn KI, dann feministisch – Impulse aus Wissenschaft und Aktivismus, editor. Netzforma* e.V.- Verein für feministische Netzpolitik; 2021. p. 1–13.
[126] Turner K, Wood D, D'Ignazio C. The abuse and misogynoir playbook. In: Gupta A, Ganapini M, Butalid R, editors. The state of AI ethics report. Montreal: Montreal Ethics Institute; 2021. p. 15–34.
[127] Manolis S, Konstantinos K, Konstantinos S, Tympas A. 'AI can be analogous to steam power' or from the 'Postindustrial Society' to the 'Fourth Industrial Revolution': An intellectual history of artificial intelligence. ICON: J Int Committee Hist. Technol. 2022;1:97–116. https://www.icohtec.org/wp-content/uploads/2022/09/27-1-97.pdf.