frontiers | Research Topics

Data-intensive medicine and healthcare: Ethical and social implications in the era of artificial intelligence and automated decision making

Edited by
Gabriele Werner-Felmayer, Jusaku Minari, Silke Schicktanz, Aviad Raz and Tamar Sharon

Published in
Frontiers in Genetics

FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-3534-9
DOI 10.3389/978-2-8325-3534-9

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Data-intensive medicine and healthcare: Ethical and social implications in the era of artificial intelligence and automated decision making

Topic editors

Gabriele Werner-Felmayer — Innsbruck Medical University, Austria
Jusaku Minari — Kyoto University, Japan
Silke Schicktanz — University of Göttingen, Germany
Aviad Raz — Ben-Gurion University of the Negev, Israel
Tamar Sharon — Radboud University, Netherlands

Citation

Werner-Felmayer, G., Minari, J., Schicktanz, S., Raz, A., Sharon, T., eds. (2023). *Data-intensive medicine and healthcare: Ethical and social implications in the era of artificial intelligence and automated decision making*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3534-9

Table of contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 07 September 2023
DOI 10.3389/fgene.2023.1280344

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Dov Greenbaum,
Yale University, United States

*CORRESPONDENCE
Aviad Raz,
aviadraz@bgu.ac.il

RECEIVED 20 August 2023
ACCEPTED 04 September 2023
PUBLISHED 07 September 2023

CITATION
Raz A, Minari J, Schicktanz S, Sharon T and Werner-Felmayer G (2023), Editorial: Data-intensive medicine and healthcare: ethical and social implications in the era of artificial intelligence and automated decision-making.
Front. Genet. 14:1280344.
doi: 10.3389/fgene.2023.1280344

COPYRIGHT
© 2023 Raz, Minari, Schicktanz, Sharon and Werner-Felmayer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Editorial: Data-intensive medicine and healthcare: ethical and social implications in the era of artificial intelligence and automated decision-making

Aviad Raz[1]*, Jusaku Minari[2], Silke Schicktanz[3], Tamar Sharon[4] and Gabriele Werner-Felmayer[5]

[1]Department of Sociology and Anthropology, Ben-Gurion University of the Negev, Beer-Sheba, Israel, [2]Uehiro Research Division for iPS Cell Ethics, Center for iPS Cell Research and Application (CiRA), Kyoto University, Kyoto, Japan, [3]Department of Medical Ethics and History of Medicine, University Medical Center Göttingen, Göttingen, Germany, [4]Department of Ethics and Political Philosophy and Interdisciplinary Hub on Digitalization and Society, Radboud University, Nijmegen, Netherlands, [5]Institute of Biological Chemistry, and Bioethics Network Ethucation, Medical University Innsbruck, Innsbruck, Austria

KEYWORDS

genomic medicine, ethics, personalized medicine, automated decision-making, artificial intelligence

Editorial on the Research Topic
Data-intensive medicine and healthcare: ethical and social implications in the era of artificial intelligence and automated decision-making

Medical "big data" and artificial intelligence (AI) are a hyped duo. Promises include developing more personalised treatments, delegating medical decision-making to tireless and seemingly objective algorithms, improving preventive screening, and providing healthcare more efficiently through predictive risk scores. AI and big data, however, do not automatically transform into improved health outcomes. The practical and functional uses of AI in big data environments require integrating and interpreting a wide variety of medical data (e.g., from genomics or other omics, imaging, biomarker analyses) and other personal data. As a result, AI-driven technology bears various new challenges and risks at the societal, algorithmic, organizational, expert, and individual levels.

Scholarship on the ethical, legal, and social issues of using AI in data-intensive medicine and healthcare has highlighted numerous areas of contention, including regulation, explainability, privacy, data sharing and protection, trust, and biases, as well as how AI might affect the patient–doctor relationship and support interdisciplinary expert teams in their decisions. Aiming to extend this perspective, this Research Topic focuses on AI applications in various areas of innovative data-intensive medicine, such as genomics, neuroscience, and child and elderly care. The contributions explore how ethical and social considerations can/should be part of medical AI by considering issues of diversity, the significance of datafication and automation, public and patient participation, developing deliberative or open science approaches (such as open codes), and how to ensure interoperability among developers and users while preventing misuse, hacking, or manipulation. The Research Topic comprises 10 articles dealing with various aspects of

the prospects and perils of AI in healthcare, which can be grouped into several themes representing key concerns in this emerging field—especially regulation, data sharing, and explainability.

Rubeis et al. can be read as a prolegomenon to the Research Topic, as they introduce a useful typology of the various ways in which "democratizing AI" is used to hype the field of AI in healthcare. Their study highlights the ways in which the concept of "democratizing AI" tends to frame patients as consumers and focus on free-market solutions, while omitting the deliberative processes and modes of participation needed to ensure that those affected by AI in healthcare have a say on its development and use. These needs and lacunae are further highlighted in the other articles.

The required and/or missing regulation and ethical embedding of AI-assisted healthcare are discussed in four articles. Stake and Heinrichs examine the ethical aspects of e-health applications for child health screening. They propose to develop age-specific models that consider the vulnerability of children to balance their right to informational self-determination with medical needs. Meszaros et al. examine more generally the future directions of AI regulation in medical care implied by the proposed EU AI Act and the EU General Data Protection Regulation, analysing ways to harmonize the principles of data protection and ethical AI. Fritzsche et al. discuss the recent use of AI for polygenic risk scores (PRSs), which may enable higher prediction accuracy but also presents a range of increasingly complex ethical challenges regarding fairness, trust, and explainability, as well as regulatory uncertainties. The authors strongly advocate a proactive approach to embedding ethics in research and implementation processes for AI-driven PRSs. Raz and Minari expand this discussion by comparing AI-derived ethnicity-related PRSs and social scoring, both of which, while representing different applications, may reproduce biases. The authors argue that if AI-derived PRSs evaluate or classify the risks of natural persons based on their ethnic/racial self-designations, this will be akin to AI-derived social scoring based on previous social behaviours in multiple contexts or known or predicted personal or personality characteristics.

The challenges of data sharing are explored in two articles. Reer et al. review the requirements for useful data sharing in human neuroscience. They discuss international legal frameworks and the standardization of data and metadata organization and annotation. Bak et al. criticize the conventionally used "either/or" choice of the "consent or anonymize approach" and its challenge to balancing data privacy and data access. They argue that the "AI revolution" in healthcare can be realized only through transnational data sharing governance policies.

Two articles address the issue of explainability. Pierce et al. discuss the opacity problem of AI in clinical use by drawing a distinction between the function of explainability for the current patient and that for the future patient. They argue that in day-to-day clinical practice, accuracy is sufficient as an "epistemic warrant" for clinical decision-making and that the most compelling reason for requiring explainability in the sense of scientific or causal explanation is its potential to improve future care. Ott and Dabrock suggest that while transparency often follows an "all or nothing" logic, intelligibility offers the opportunity to uncover the essential elements of an AI system: Does the system provide an adequate basis for rendering people intelligible? And does it do so not only ex ante during data collection and algorithm design but continuously during implementation and adaptation and, finally, ex post after the actual use case?

Finally, Schicktanz et al. suggest a novel approach not only to embedding ethics into the development and use of medical AI (as all the articles discuss for their respective fields) but also to integrating AI into the development of ethical assessment. They argue for AI-assisted ethical simulation that can improve context-sensitive ethical analyses, as well as for thought experiments and future-oriented technology assessments—for example, applications catering for persons with dementia or cognitive impairment.

The diversity of the articles included in this Research Topic reminds us that under no circumstances should groups exclusively pursuing their own interest dominate the debate on medical AI. Rather, addressing the ethical challenges of medical AI requires interdisciplinary efforts involving computer scientists, ethicists, sociologists, policymakers, and domain experts (such as healthcare professionals) to address the multiple aspects of this debate which should be open to all the stakeholders involved.

Author contributions

AR: Writing–original draft, Writing–review and editing. JM: Writing–review and editing. SS: Writing–review and editing. TS: Writing–review and editing. GW-F: Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. AR and JM are supported by funding provided by the JSPS–ISF Joint Program, Grant 62/22 (ISF), JPJSBP120228404 (JSPS), "Biobanks for genomic medicine in Israel and Japan: an analysis of ethics and policy." JM is supported by the JSPS Grant-in-Aid for Scientific Research (B), No. JP21H03163. TS is supported by funding provided by the European Research Council, Grant No. 804985.

Acknowledgments

We thank the authors of the papers comprising this Research Topic for their valuable contributions and the referees for their rigorous review.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

PERSPECTIVE
published: 13 June 2022
doi: 10.3389/fgene.2022.929453

You Can't Have AI Both Ways: Balancing Health Data Privacy and Access Fairly

*Marieke Bak[1], Vince Istvan Madai[2,3], Marie-Christine Fritzsche[4], Michaela Th. Mayrhofer[5] and Stuart McLennan[4]**

[1]Department of Ethics, Law and Humanities, Amsterdam UMC, University of Amsterdam, Amsterdam, Netherlands, [2]QUEST Center for Responsible Research, Berlin Institute of Health (BIH), Charité Universitätsmedizin Berlin, Berlin, Germany, [3]School of Computing and Digital Technology, Faculty of Computing, Engineering and the Built Environment, Birmingham City University, Birmingham, United Kingdom, [4]Institute of History and Ethics in Medicine, TUM School of Medicine, Technical University of Munich, Munich, Germany, [5]ELSI Services and Research, Biobanking and BioMolecular Resources Research Infrastructure European Research Infrastructure Consortium (BBMRI-ERIC), Graz, Austria

Artificial intelligence (AI) in healthcare promises to make healthcare safer, more accurate, and more cost-effective. Public and private actors have been investing significant amounts of resources into the field. However, to benefit from data-intensive medicine, particularly from AI technologies, one must first and foremost have access to data. It has been previously argued that the conventionally used "consent or anonymize approach" undermines data-intensive medicine, and worse, may ultimately harm patients. Yet, this is still a dominant approach in European countries and framed as an either-or choice. In this paper, we contrast the different data governance approaches in the EU and their advantages and disadvantages in the context of healthcare AI. We detail the ethical trade-offs inherent to data-intensive medicine, particularly the balancing of data privacy and data access, and the subsequent prioritization between AI and other effective health interventions. If countries wish to allocate resources to AI, they also need to make corresponding efforts to improve (secure) data access. We conclude that it is unethical to invest significant amounts of public funds into AI development whilst at the same time limiting data access through strict privacy measures, as this constitutes a waste of public resources. The "AI revolution" in healthcare can only realise its full potential if a fair, inclusive engagement process spells out the values underlying (trans) national data governance policies and their impact on AI development, and priorities are set accordingly.

Keywords: digital health, data access, data privacy, ethics, artificial intelligence, fairness, resource allocation

OPEN ACCESS

Edited by:
Gabriele Werner-Felmayer, Innsbruck Medical University, Austria

Reviewed by:
Tatyana Novossiolova, Center for the Study of Democracy, Bulgaria
Marietjie Botes, University of Luxembourg, Luxembourg

****Correspondence:***
Stuart McLennan
stuart.mclennan@tum.de

Specialty section:
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Genetics

Received: *26 April 2022*
Accepted: *23 May 2022*
Published: *13 June 2022*

Citation:
Bak M, Madai VI, Fritzsche M-C, Mayrhofer MT and McLennan S (2022) You Can't Have AI Both Ways: Balancing Health Data Privacy and Access Fairly. Front. Genet. 13:929453. doi: 10.3389/fgene.2022.929453

INTRODUCTION

The growth of digital health data and increasing computational capabilities have created significant opportunities for the use of artificial intelligence (AI) technology in healthcare. With the ability to learn from large volumes of clinical, -omics, and other health data, AI has the potential to support a wide range of activities: diagnosis, clinical decision making, personalized medicine, clinical research, drug development, administrative processes, and the mitigation of health disparities (Shibata &

Wada, 2011; Fleming, 2018; Shortliffe & Sepúlveda, 2018; Davenport & Kalakota, 2019; Fiske, Henningsen, & Buyx, 2019; Liu et al., 2019; Schork, 2019; Woo, 2019). If data-intensive medicine can realize continuous improvement of healthcare quality and thereby reduce patient harm, improve health, empower personal decision making, and increase equity, it would fulfil the core ethical principles of healthcare (ABIM Foundation, 2002; Beauchamp & Childress, 2013; McLennan et al., 2018).

The potential opportunities of AI have led many countries, particularly in the European Union (EU), to invest significant financial and human resources in AI initiatives. In the past few years, previously unseen amounts of public and private investment have flowed into AI applications (KPMG, 2018; CB Insights, 2019). National AI strategies with large, dedicated budgets were published by many EU countries (Righi et al., 2022), e.g., the German federal government promised to allocate 3 billion EUR in funding between 2020–2025 (Die Bundesregierung, 2018). Funding for healthcare and medical AI-related research projects through the EU Horizon 2020 scheme increased between 2014–2020, although large differences in investments can be seen between Member States (around 80 million EUR was awarded to projects in each of the top-funded countries and around 100.000 EUR in countries receiving the lowest amount of funding) (De Nigris et al., 2020, p. 27). To guide the responsible design of these new AI systems in healthcare and beyond, several ethical and legal instruments were newly created by the European Commission (EC), such as the proposed Artificial Intelligence Act (EC, 2021), the Guidelines for Trustworthy AI (EC, 2019), and the updated Medical Device Regulation (EC, 2020), to complement the General Data Protection Regulation (GDPR) which remains the key legal instrument regarding data usage for AI development (EC, 2016).

The use of health data for AI development raises important data privacy concerns, both at an individual and group level (McLennan et al., 2018; Mittelstadt 2019). Thus, there is a tension between incentives and actions that promote AI and incentives and actions that limit access to the required data: "the data hunger of AI runs up against the norm of personal data minimization" (Sorell et al., 2022). This leads to complex dilemmas. All the resources and efforts currently devoted to AI development could go to waste if the issue of data access is not adequately addressed. In this context, it is noteworthy that the proposed EU AI Act requires, for example, the highest levels of data quality and quantity for sufficient training, validation, and testing as well as the necessary heterogeneity to cover relevant patient (sub) populations and variants in the intended clinical setting (Art. 10). This requires broad access to healthcare data, and tools not fulfilling these requirements would not be permitted. Countries must thus decide how to balance the positive goals of secondary-use activities like healthcare AI with mitigating associated privacy risks. These trade-offs raise issues of resource allocation and justice that have so far been largely neglected in policy debates and the scholarly literature. In this perspective article, we provide an overview of these macro-level ethical trade-offs related to data use for healthcare AI. While we remain neutral on how one should value data privacy and access, we conclude by providing procedural recommendations that allow this decision to be made in a fair manner.

VARIATION IN EUROPEAN UNION DATA GOVERNANCE

Health-related AI applications are in crucial need of patient data during the development of the AI model in the training, validation and test phases. These health data are often initially collected for a different purpose than AI development, and this secondary use requires a valid ethical and legal basis. In Europe, the central legal instrument in this domain remains the GDPR which is directly enforceable in all EU Member States and applies to all EU citizens. The GDPR has the dual aim of protecting personal data, meaning data that can be traced back to living individuals without unreasonable effort, and achieving a higher level of harmonization of data protection practices.

As a result of political compromises, however, the GDPR leaves it open in several places for Member States to issue derogations in their national law when it concerns public interest, scientific or historical research purposes or statistical purposes. (Heckmann and Scheurer, 2021). This may include deciding on what constitutes sufficient methods of pseudonymization, when data can be considered fully non-identifiable, what further restrictions should be imposed on processing sensitive data for research purposes, and what are sufficient safeguards and conditions for processing data under the research exemption (Shabani et al., 2018). In addition to the GDPR, national health and biobanking laws might also have implications for data protection requirements and ultimately access to health data and data governance. (Bak et al., 2020; Kindt et al., 2021; Slokenberga et al., 2021) As a result, there remains a wide variation of data governance approaches across Europe and the actual balance between data protection rules and access requirements is struck at country-level. In this regard more conservative Germany and more liberal Finland are examples of countries that differ in their approaches to data governance.

The Finnish approach to data access is evident in its Act on the Secondary Use of Health and Social Data (Ministry of Social Affairs and Health, 2019) which provides the basis for the national data permit authority FinData to facilitate access to and sharing of patient data. The country has adopted a national policy oriented towards big data and open data to transform the technical and governance infrastructure for AI and other computer science research (Aula, 2019). In Finland, consent is not legally required for including personal data in national health registries, but data access is controlled through detailed policies and security procedures (Vrijenhoek et al., 2021). Moreover, the Biobank Act (2012) which is currently undergoing further reform, allows samples and related data to be used for research purposes without (re-)consent for every research project, and biobank samples can be linked to health data from national registries. Being the frontrunner in developing a national AI strategy already in 2017, Finland is among the most digitally developed EU countries and provides an online service

which citizens use to view their health information from different sources (EC, 2019; Jormanainen et al., 2019). There is an explicit focus on public education and awareness, including a free online AI course. As in other Nordic countries, Finland´s national AI strategy generally reflects the core values of trust, openness, and transparency (Robinson, 2020).

This contrasts with the German approach that has traditionally been geared toward comprehensive control and where health data research is usually conducted with patient consent. For example, consent is *the* legal basis for any processing of data stored in the newly launched electronic patient record (elektronische Patientenakte or ePA) whose use is voluntary, and which gives patients full control over their data (Molnar-Gabor et al., Forthcoming 2021). Data processing for scientific research in the public interest might take place without consent, if organizational and technical provisions are met, as specified in the Federal Data Protection Act (Molnár-Gábor et al., 2018). In 2018, the German State Minister for Digitalization stated that the country's strict data protection laws block development in the healthcare sector (Kaiser, 23 December 2018). Indeed, in practice, this research exemption seems hardly ever used. A recent interview study with researchers, data protection officers and research ethics committee representatives in the state of Bavaria, found that German law was perceived as vague and was differently interpreted across federal states and institutions (McLennan et al., 2022a). This resulted in secondary health data research usually only taking place when consent had been obtained or data were fully anonymized.

TRADE-OFFS IN REALIZING THE POTENTIAL OF ARTIFICIAL INTELLIGENCE IN HEALTHCARE

Data Privacy Versus Data Access

This variation in data governance approaches can hamper (inter-) national data sharing and makes it difficult to create disease registries and to develop AI tools (De Lange et al., 2019; McLennan et al., 2019; Haneef et al., 2020). The disagreement over the interpretation of certain provisions in the GDPR, including research exemptions, is not easily solved as it stems from different viewpoints on how to balance foundational values like informational self-determination versus solidarity (Hoffman et al., 2012; van Veen, 2018). Whether (national) strategies should focus on data privacy or data access is a difficult question linked to various ethical dilemmas. Namely, what we might identify as a more liberal approach to data access might have in turn serious implications for fundamental rights to privacy. A restrictive approach, on the other hand, might undermine data-intensive medicine and in turn cause harm by biasing models and leading to wasted investments into AI development.

Governments and institutions taking a more liberal approach to data governance, i.e., interpreting the GDPR generously by focusing on its harmonization and data sharing aim, may face complex ethical issues and public resistance. For instance, the *care.data* program in the United Kingdom famously collected health data for secondary use without informed consent and with limited options for opt-out, which adversely affected public trust in health data initiatives (Vezyridis & Timmons, 2017). Innovations in AI may promise to improve the quality of care and lower costs, but the need for detailed personal information as input data exacerbates known concerns about issues like data privacy, bias and discrimination (Mittelstadt & Floridi, 2016; Price & Cohen, 2019).

Those with a more restrictive view on data governance generally use the "consent or anonymize" mind set: personal data may only be used if informed consent is obtained or the information is fully anonymized (Mostert et al., 2016). However, requiring (re-)consent can lead to significant administrative and financial hurdles that delay important activities or even make them unfeasible (Tu et al., 2004; Jansen et al., 2007). Requiring (re-)consent may also lead to major selection biases that undermine data representativeness, which can lead to biased AI models that in turn harm patients and exacerbate existing health inequalities (Vayena et al., 2018). In addition, while consent may protect the privacy of persons whose data are used to train and test AI models (that is, if the information is clear and unambiguously presented and the patient is in a position to make a reasoned decision), it does not protect the privacy of others who did not consent but can still have inferences drawn about them based on rules derived from a cohort of consenting individuals (Barocas et al., 2014).

Furthermore, although anonymized data is out of scope of the GDPR, data anonymization is not free of technical, legal and ethical challenges. Full anonymization has become increasingly difficult due to the potential of cross-linking datasets and the inclusion of highly personal data like genetic sequences (Gymrek et al., 2013). Further, irreversible anonymization may involve removing essential information needed to perform secondary activities like research. Additionally, some authors argue that anonymization is merely possible in a specific context for a short period of time and requires regular reassessments to determine whether the status of anonymization can still be upheld, making it equally resource intense as asking consent (Sariyar and Schlünder, 2016). Even if full anonymization was possible and/or feasible, it offers no guarantees that AI models based on such "anonymous" data do not harm the individuals who donated their data (Barocas et al., 2014).

In Europe, concerns have been raised for several years about the "overprotection" of personal data under (draft versions of) the GDPR, which are still relevant given the varying interpretations of the regulation (Ploem et al., 2013; Author Anonymous, 2015; Timmers et al., 2019). In a recent open letter by genetic researchers, a similar concern was voiced about access to digital sequence information that can be used for public health, as policy negotiations are feared to favour data sovereignty and limit data sharing under the Convention on Biological Diversity (DSI, 2022). The broader debate on informational self-determination versus scientific data research dates back well into the previous century. Yet, when it comes to AI, we sometimes seem to forget that data access is the most important prerequisite for any AI innovation. This omission may lead to a situation where some policies follow the current trend of pouring tremendous resources into health AI developments

when, at the same time, the success of the funded research is effectively made impossible due to the country's specific interpretation of the GDPR and relevant national law.

Overprotection or Overinvestment?

The potential of healthcare AI in Europe is limited when countries' data governance approaches are overly strict, ambiguous, or contradicting. Haneef et al. (2020) surveyed the use of AI by national public health institutes and found it limited in practice, reportedly due to the complexity of data regulation laws coupled with lack of human resources and the absence of a robust data governance framework in various countries and institutions. Enabling researchers to create AI applications that help improve care, requires giving them greater access to patient data, albeit conditional and within a secure environment. The EU and several Member States seek to achieve a win on all fronts, i.e., they want to become both a leading player in health AI as well as provide maximum protection regarding health data privacy. However, policy-makers must realize that hard choices are unavoidable to be able to strike the right balance in data governance.

Public resources are generally finite, so whether the right to health is best fulfilled by prioritizing investment in AI-driven technologies over data infrastructure development or other healthcare spending, is ethically relevant. As we indicated above, a country that takes a very restrictive approach to data access needs to take this into account when allocating funds. Future legislation such as the proposed EU AI act could essentially ban AI in healthcare applications if developers do not have broad access to relevant healthcare data and therefore cannot meet generalization and bias mitigation requirements. Thus, development of robust technological data management and governance structures, such as the proposed European Health Data Space (EHDS) and standards for interoperability of health records that promise to improve data access and usability (Shabani, 2022), should then be established prior or at least in parallel to the creation of specific AI tools. The European Investment Bank claims that the EU is limiting innovation by underinvesting in AI, quoting an investment gap of up to 10 billion EUR (Verbeek & Lundqvist, 2021), but we disagree with this general statement. Rather, investing in AI-driven healthcare technology that cannot prosper due to unresolved data governance issues would rather constitute an overinvestment, i.e., an unjust waste of resources.

Moreover, resources allocated to health AI may come at the expense of non-AI solutions. Since the value of AI remains uncertain and many health interventions in the field of AI are—thus far—of limited real-world effectiveness (D'Amour et al., 2020; Skorburg et al., 2021), it has been argued that policy-makers should not allocate resources to AI tools exclusively, especially when these resources could strengthen existing evidence-based solutions and help to overcome structural barriers to care (Skorburg et al., 2021). This dilemma is well-known in the field of public health. For instance, in the field of HIV prevention in low- and middle-income countries, the development of pharmaceutical PrEP (Pre-Exposure Prophylaxis) led to fears that funding for the free provision of condoms would be curtailed. However, PrEP, was never intended to be a stand-alone intervention and its combined use with condoms has proven to be more effective and acceptable than either intervention on its own (Bak et al., 2018). Similarly, the discussion around AI in medicine has shifted away from the complete replacement of physicians and their judgement to more synergistic uses of AI (i.e. doctors plus AI) (Mazzanti et al., 2018; Dos Santos et al., 2019). Thus, if actors decide to invest in health AI, this needs to be accompanied with investment into not only data access structures but also the surrounding healthcare system that interacts with the AI tool. Nonetheless, this might be difficult given resource constraints. How then should we decide what constitutes just resource allocation for health AI?

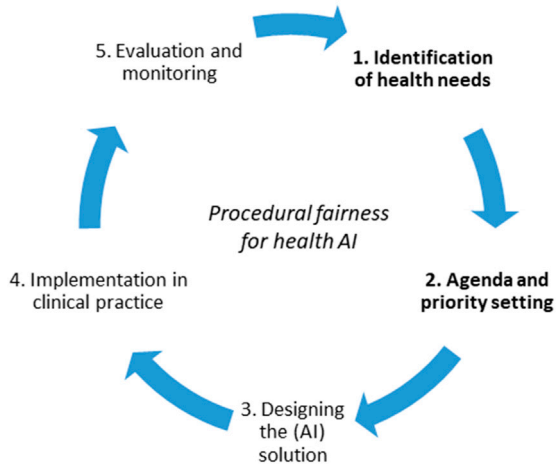


FIGURE 1 | Procedural fairness for priority-setting in health AI, with special attention for steps 1 and 2. Adapted from the Policy Cycle (Howlett & Giest, 2015).

TOWARDS A FAIR PRIORITIZATION FOR HEALTH ARTIFICIAL INTELLIGENCE

Most of the literature on AI ethics focuses only on the fairness concerns *inherent* to this upcoming technology (e.g., related to bias and discrimination in the models), rather than on the trade-offs between data privacy and access and the resulting questions of resource allocation. For example, in the high-level expert guidance on Trustworthy AI by the EC, seven key requirements are listed that should be implemented by model developers and about which end-users should be informed (EC, 2019). By emphasizing the requirements of the AI system itself, however, the EC narrows the ethical debate to the interaction with a specific application. While such principlist guidelines can help sensitize professionals to the built-in values of AI applications, they do not provide a solution to the wider moral dilemmas that arise from value conflicts and resource limitations (Bak, 2020).

Discussions about ethical requirements for AI should thus be preceded by a broader ethical debate about these priorities: rather than just holding AI to account, our public investments in AI

should be accountable. The policy and planning cycle of health intervention development helps illustrate our point (**Figure 1**). While the focus of most ethicists and policy-makers has been on step 3 (the design of the AI solution) and to a lesser extent steps 4 and 5 (implementation and evaluation), we want to refocus the debate on steps 1 and 2 of the cycle (identification of health needs and subsequent priority-setting). Our suggestion is in line with recommendations from the World Economic Forum that the creation of national AI strategies should start with a SWOT (strengths, weaknesses, opportunities, and threats) analysis, as was done in Finland, to keep policy goals in line with resource constraints and needs of citizens (Madzou et al., 2019). This is ultimately a political discussion, as is any debate on technology that involves choices between competing values.

The conditions of such societal debate can be found in the work of the American philosopher Norman Daniels (2007), who argues that when there is no consensus on substantive values, we should focus on procedural values. Fair process is important as it allows healthcare organizations to pursue their (research) policies with a mandate from society. This idea was formalized into a model known as Accountability for Reasonableness (A4R) which proposes key conditions for the legitimacy of decision-making in public health (Daniels & Sabin, 1997). It is beyond the scope of this paper to discuss the A4R framework in detail but it has been found valuable for the field of digital health (Wong, 2020) and was used for drafting the Montreal Declaration for Responsible Development of AI, which launched in 2017 after an extensive public deliberation process (Dilhac et al., 2018; Brall et al., 2019). We support the idea that A4R or similar procedural fairness frameworks should be used in deliberations about resource allocation for health AI.

Decision-makers in EU countries should structurally engage an inclusive group of researchers, data subjects, clinicians, and other relevant stakeholders, to deliberate the trade-offs between data privacy and the value of AI. We want to emphasize we do not suggest favouring any of the two approaches but propose that inclusive engagement or "data democracy" is needed to ensure that decisions empower affected communities and are sensitive to their specific needs, which in turn may help to promote public trust (Ienca et al., 2018; Kalluri, 2020; Nyrup, 2021). Ethicists may join the process to help explain and clarify complex moral questions (McLennan et al., 2022b). This of course requires transparent insight into the available budgets and competing needs. All in all, if such reflections lead to a country explicitly deciding to focus on a strict, conditional or liberal approach to data privacy and/or data access, that decision is morally legitimate if it fulfils conditions of procedural fairness, e.g. accountability and transparency.

CONCLUSION

The development and implementation of AI for healthcare comes with trade-offs: striving for all-embracing data privacy has proven incompatible with the desire to realize the full potential of AI for medical purposes. We have outlined that countries need to implement digital health strategies that are consistent, which requires an examination of the core values that underlie the national data governance frameworks. In a nutshell, they should deliberate with their citizens and be able to explain to them why they have set certain priorities, and the chosen balance between specific data privacy and data access conditions should be reflected in the national and ultimately European AI budgets. Failing to do so is leading to distributive justice concerns that should not be overlooked in debates on the ethical aspects of health-related AI.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

SM, MB, MF and VM conceived the initial idea of the paper. MB prepared the initial draft with the assistance of SM. MF, VM and MM reviewed the draft and critically revised it for important intellectual content. All authors read and approved the final version of the article.

REFERENCES

ABIM Foundation (2002). Medical Professionalism in the New Millennium: a Physician Charter. *Ann. Intern Med.* 136 (3), 243–246. doi:10.7326/0003-4819-136-3-200202050-00012

Aula, V. (2019). Institutions, Infrastructures, and Data Friction-Reforming Secondary Use of Health Data in Finland. *Big Data Society* 6 (2).

Author Anonymous (2015). Data Overprotection. *Nature* 522, 391–392. doi:10.1038/522391b

Bak, M. A. R. (2020). Computing Fairness: Ethics of Modeling and Simulation in Public Health. *Simulation* 98 (2), 103–111. doi:10.1177/0037549720932656

Bak, M. A. R., Ploem, M. C., Ateşyürek, H., Blom, M. T., Tan, H. L., and Willems, D. L. (2020). Stakeholders' Perspectives on the Post-mortem Use of Genetic and Health-Related Data for Research: a Systematic Review. *Eur. J. Hum. Genet.* 28 (4), 403–416. doi:10.1038/s41431-019-0503-5

Bak, M. A. R., van Dam, A., and Janssens, R. (2018). Awareness and Acceptability of Pre-exposure Prophylaxis (PrEP) Among Men Who Have Sex with Men in Kazakhstan: a Mixed Methods Study. *Central Asian J. Med. Sci.* 4 (2), 102–115.

Barocas, S., Nissenbaum, H., Lane, J, Stodden, V, Bender, S, and Nissenbaum, H. (2014). Big Data's End Run Around Anonymity and Consent. *Priv. big data, public good Fram. Engagem.* 1, 44–75. doi:10.1017/cbo9781107590205.004

Beauchamp, T. L., and Childress, J. F. (2013). *Principles of Biomedical Ethics*. 7th ed. Oxford: Oxford University Press.

Brall, C., Schröder-Bäck, P., and Maeckelberghe, E. (2019). Ethical Aspects of Digital Health from a Justice Point of View. *Eur. J. public health* 29, 18–22. doi:10.1093/eurpub/ckz167

Biobank Act (2012). *Ministry of Social Affairs and Health, Finland*. Biobank Act 688.

CB Insights (2019). Global Healthcare Report Q2 2019. Available at: https://www.cbinsights.com/research/report/healthcare-trends-q2-2019.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., et al. (2020). Underspecification Presents Challenges for Credibility in Modern

Machine Learning. arXiv [Preprint]. Available at: https://arxiv.org/abs/2011.03395.

Daniels, N. (2007). *Just Health: Meeting Health Needs Fairly*. New York: Cambridge University Press.

Daniels, N., and Sabin, J. (1997). Limits to Health Care: Fair Procedures, Democratic Deliberation, and the Legitimacy Problem for Insurers. *Philosophy Public Aff.* 26, 303–350. doi:10.1111/j.1088-4963.1997.tb00082.x

Davenport, T., and Kalakota, R. (2019). The Potential for Artificial Intelligence in Healthcare. *Future Healthc. J.* 6, 94–98. doi:10.7861/futurehosp.6-2-94

De Lange, D. W., Guidet, B., Andersen, F. H., Artigas, A., Bertolini, G., Moreno, R., et al. (2019). Huge Variation in Obtaining Ethical Permission for a Non-interventional Observational Study in Europe. *BMC Med. Ethics.* 20 (1). doi:10.1186/s12910-019-0373-y

De Nigris, S., Craglia, M., Nepelski, D., Hradec, J., Gómez-González, E., Gomez, E., et al. (2020). *AI Watch: AI Uptake in Health and Healthcare 2020*. Luxembourg: Publications Office of the European Union. 978-92-76-26936-6. EUR 30478 EN, (JRC122675. doi:10.2760/948860

Die Bundesregierung (2018). *Strategie Künstliche Intelligenz der Bundesregierung*. Berlin: German Government.

Dilhac, M., Abrassart, C., and Voarino, N. (2018). *Report On the Montréal Declaration for a Responsible Development of Artificial Intelligence. Univ. Montréal.* Available at: https://www.montrealdeclaration-responsibleai.com/.

DSI Scientific Network (2022). Open Letter. Available at: https://www.dsiscientificnetwork.org/open-letter/.

European Commission (2019). Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

European Commission (2021). Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.

European Commission (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance). OJ L 119/1.

European Commission (2020). Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on Medical Devices, Amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and Repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA Relevance). Text EEA relevance.

Fiske, A., Henningsen, P., and Buyx, A. (2019). Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. *J. Med. Internet. Res.* 21, e13216. doi:10.2196/13216

Fleming, N. (2018). How Artificial Intelligence Is Changing Drug Discovery. *Nature* 557, S55–S57. doi:10.1038/d41586-018-05267-x

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., and Erlich, Y. (2013). Identifying Personal Genomes by Surname Inference. *Science* 339, 321–324. doi:10.1126/science.1229566

Haneef, R., Delnord, M., Vernay, M., Bauchet, E., Gaidelyte, R., Van Oyen, H., et al. (2020). Innovative Use of Data Sources: A Cross-Sectional Study of Data Linkage and Artificial Intelligence Practices across European Countries. *Arch. Public Health.* 78 (1), 55–11. doi:10.1186/s13690-020-00436-9

Heckmann, D., and Scheurer, M. (2021). "Datenschutzrecht," in *Praxiskommentar Internetrecht. Juris*. Editors D. Heckmann and A. Paschke.

Hoffman, S., and Podgurski, A. (2012). Balancing Privacy, Autonomy, and Scientific Needs in Electronic Health Records Research. *SMUL Rev.* 65, 85.

Howlett, M., and Giest, S. (2015). "The Policy-Making Process," in *Routledge Handbook of Public Policy*. Editors E. Araral, S. Fritzen, M. Howlett, M. Ramesh, and X. Wu (London: Routledge), 17–18.

Jansen, T. C., Kompanje, E. J. O., Druml, C., Menon, D. K., Wiedermann, C. J., and Bakker, J. (2007). Deferred Consent in Emergency Intensive Care Research: what if the Patient Dies Early? Use the Data or Not? *Intensive Care Med.* 33, 894–900. doi:10.1007/s00134-007-0580-8

Jormanainen, V., Parhiala, K., Niemi, A., Erhola, M., Keskimäki, I., and Kaila, M. (2019). Half of the Finnish Population Accessed Their Own Data: Comprehensive Access to Personal Health Information Online Is a Corner-Stone of Digital Revolution in Finnish Health and Social Care. *Finn. J. eHealth Welf.* 11 (4). doi:10.23996/fjhw.83323

Kaiser, T. (2018). Dorothee Bär will Datenschutz für Patienten lockern. Die Welt. Available at: https://www.welt.de/wirtschaft/article186013534/Dorothee-Baer-will-Datenschutz-fuer-Patienten-lockern.html.

Kalluri, P. (2020). Don't Ask if Artificial Intelligence Is Good or Fair, Ask How it Shifts Power. *Nature* 583, 169. doi:10.1038/d41586-020-02003-2

Kindt, E., Fontanillo Lopez, C. A., Czarnocki, J., Kanevskaia, O., and Herveg, J. (2021). Study on the Appropriate Safeguards Required under Article 89(1) of the GDPR for the Processing of Personal Data for the Scientific Research. Prep. by Milieu under lead KU-Leuven under Contract No EDPS/2019/02-08 benefit EDPB.

KPMG Enterprise (2018). Venture Pulse: Q1'18 Global Analysis of Venture Funding. Available at: https://home.kpmg/xx/en/home/insights/2018/04/venture-pulse-q1-18-global-analysis-of-venture-funding.html.

Ienca, M., Ferretti, A., Hurst, S., Puhan, M., Lovis, C., and Vayena, E. (2018). Considerations for Ethics Review of Big Data Health Research: A Scoping Review. *PloS one* 13 (10), e0204937.

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., et al. (2019). A Comparison of Deep Learning Performance against Health-Care Professionals in Detecting Diseases from Medical Imaging: a Systematic Review and Meta-Analysis. *Lancet Digital Health* 1, e271–e297. doi:10.1016/s2589-7500(19)30123-2

Madzou, L., Shukla, P., Caine, M., Campbell, T. A., Davis, N., Firth-Butterfield, K., et al. (2019). A Framework for Developing a National Artificial Intelligence Strategy. *World Econ. Forum White Pap.* Available at: https://www3.weforum.org/docs/WEF_National_AI_Strategy.pdf.

Mazzanti, M., Shirka, E., Gjergo, H., and Hasimi, E. (2018). Imaging, Health Record, and Artificial Intelligence: Hype or Hope? *Curr. Cardiol. Rep.* 20 (6), 48–49. doi:10.1007/s11886-018-0990-y

McLennan, S., Fiske, A., Tigard, D., Müller, R., Haddadin, S., and Buyx, A. (2022a). Embedded Ethics: a Proposal for Integrating Ethics into the Development of Medical AI. *BMC Med. Ethics* 23 (1), 6–10. doi:10.1186/s12910-022-00746-3

McLennan, S., Kahrass, H., Wieschowski, S., Strech, D., and Langhof, H. (2018). The Spectrum of Ethical Issues in a Learning Health Care System: a Systematic Qualitative Review. *Int. J. Qual. Health Care.* 30 (3), 161–168. doi:10.1093/intqhc/mzy005

McLennan, S., Rachut, S., Lange, J., Fiske, A., Heckmann, D., and Buyx, A. (2022b). Practices and Attitudes of Bavarian Stakeholders Regarding the Secondary-Use of Health Data for Research Purposes during the COVID-19 Pandemic: a Qualitative Interview Study. *JMIR Prepr.* 14/04/2022:38754.

McLennan, S., Shaw, D., and Celi, L. A. (2019). The Challenge of Local Consent Requirements for Global Critical Care Databases. *Intensive Care Med.* 45 (2), 246–248. doi:10.1007/s00134-018-5257-y

Ministry of Social Affairs and Health, Finland (2019). Act on the Secondary Use of Health and Social Data.

Mittelstadt, B. D., and Floridi, L. (2016). The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts, *Sci. Eng. Ethics*, 22, 2, 445–480. The Ethics of Biomedical Big Data. Springer. doi:10.1007/978-3-319-33525-4_19

Mittelstadt, B. (2019). The Ethics of Biomedical 'Big Data' Analytics. *Philos. Technol.* 32 (1), 17–21. doi:10.1007/s13347-019-00344-z

Molnár-Gábor, F. (2018). Germany: a Fair Balance between Scientific Freedom and Data Subjects' Rights? *Hum. Genet.* 137 (8), 619–626.

Molnar-Gabor, F., Sellner, J., Pagil, S., Slokenberga, S., Tzortzatou, O., and Nyström, K. (Forthcoming 2021). Harmonization after the GDPR? Divergences in the Rules for Genetic and Health Data Sharing in Four Member States and Ways to Overcome Them by EU Measures: Insights from Germany, Greece, Latvia and Sweden. *Seminars in Cancer Biology*.

Mostert, M., Bredenoord, A. L., Biesaart, M. C. I. H., and Van Delden, J. J. M. (2016). Big Data in Medical Research and EU Data Protection Law: Challenges to the Consent or Anonymise Approach. *Eur. J. Hum. Genet.* 24 (7), 956–960. doi:10.1038/ejhg.2015.239

Nyrup, R. (2021). From General Principles to Procedural Values: Responsible Digital Health Meets Public Health Ethics. *Front. Digit. Health* 3, 690417. doi:10.3389/fdgth.2021.690417

Pinto dos Santos, D., Giese, D., Brodehl, S., Chon, S. H., Staab, W., Kleinert, R., et al. (2019). Medical Students' Attitude towards Artificial Intelligence: a Multicentre Survey. *Eur. Radiol.* 29 (4), 1640–1646. doi:10.1007/s00330-018-5601-1

Ploem, M. C., Essink-Bot, M. L., and Stronks, K. (2013). Proposed EU Data Protection Regulation Is a Threat to Medical Research. *BMJ* 346, f3534. doi:10.1136/bmj.f3534

Price, W. N., and Cohen, I. G. (2019). Privacy in the Age of Medical Big Data. *Nat. Med.* 25 (1), 37–43. doi:10.1038/s41591-018-0272-7

Righi, R., Pineda, C., Cardona, M., Soler Garrido, J., Papazoglou, M., and Samoili, S. (2022). *AI Watch Index 2021*. Luxembourg: Publications Office of the European Union: Joint Research Centre (JRC), European Commission.

Robinson, S. C. (2020). Trust, Transparency, and Openness: How Inclusion of Cultural Values Shapes Nordic National Public Policy Strategies for Artificial Intelligence (AI). *Technol. Soc.* 63, 101421. doi:10.1016/j.techsoc.2020.101421

Sariyar, M., and Schlünder, I. (2016). Reconsidering Anonymization-Related Concepts and the Term "Identification" against the Backdrop of the European Legal Framework. *Biopreservation Biobanking* 14 (5), 367–374. doi:10.1089/bio.2015.0100

Schork, N. J. (2019). Artificial Intelligence and Personalized Medicine. *Cancer Treat. Res.* 178, 265–283. doi:10.1007/978-3-030-16391-4_11

Shabani, M., and Borry, P. (2018). Rules for Processing Genetic Data for Research Purposes in View of the New EU General Data Protection Regulation. *Eur. J. Hum. Genet.* 26, 149–156. doi:10.1038/s41431-017-0045-7

Shabani, M. (2022). Will the European Health Data Space Change Data Sharing Rules? *Science* 375 (6587), 1357–1359. doi:10.1126/science.abn4874

Shibata, T., and Wada, K. (2011). Robot Therapy: A New Approach for Mental Healthcare of the Elderly – A Mini-Review. *Gerontology* 57, 378–386. doi:10.1159/000319015

Shortliffe, E. H., and Sepúlveda, M. J. (2018). Clinical Decision Support in the Era of Artificial Intelligence. *JAMA* 320, 2199–2200. doi:10.1001/jama.2018.17163

Skorburg, J. A., and Yam, J. (2021). Is There an App for that?: Ethical Issues in the Digital Mental Health Response to COVID-19. *AJOB Neurosci.*, 1–14. doi:10.1080/21507740.2021.1918284

Slokenberga, S., Tzortzatou, O., and Reichel, J. (2021). *GDPR and Biobanking: Individual Rights, Public Interest and Research Regulation across Europe*. Cham, Switzerland: Springer Nature, 434.

Sorell, T., Rajpoot, N., and Verrill, C. (2022). Ethical Issues in Computational Pathology. *J. Med. Ethics.* 48 (4), 278–284. doi:10.1136/medethics-2020-107024

Timmers, M., Van Veen, E.-B., Maas, A. I. R., and Kompanje, E. J. O. (2019). Will the EU Data Protection Regulation 2016/679 Inhibit Critical Care Research? *Med. Law Rev.* 27 (1), 59–78. doi:10.1093/medlaw/fwy023

Tu, J. V., Willison, D. J., Silver, F. L., Fang, J., Richards, J. A., Laupacis, A., et al. (2004). Impracticability of Informed Consent in the Registry of the Canadian Stroke Network. *N. Engl. J. Med.* 350, 1414–1421. doi:10.1056/nejmsa031697

Van Veen, E.-B. (2018). Observational Health Research in Europe: Understanding the General Data Protection Regulation and Underlying Debate. *Eur. J. Cancer* 104, 70–80. doi:10.1016/j.ejca.2018.09.032

Vayena, E., Blasimme, A., and Cohen, I. G. (2018). Machine Learning in Medicine: Addressing Ethical Challenges. *PLoS Med.* 15 (11), e1002689. doi:10.1371/journal.pmed.1002689

Verbeek, A., and Lundqvist, M. (2021). Artificial Intelligence, Blockchain and the Future of Europe. Innovation Finance Advis. part Eur. Invest. Bank's Advis. Serv., https://www.eib.org/en/publications/artificial-intelligence-blockchain-and-the-future-of-europe-report.

Vezyridis, P., and Timmons, S. (20172017). Understanding the care.Data Conundrum: New Information Flows for Economic Growth. *BD&S* 4, 2053951716688490. doi:10.1177/2053951716688490

Vrijenhoek, T., Tonisson, N., Kääriäinen, H., Leitsalu, L., and Rigter, T. (2021). Clinical Genetics in Transition-A Comparison of Genetic Services in Estonia, Finland, and the Netherlands. *J. Community Genet.* 12 (2), 277–290. doi:10.1007/s12687-021-00514-7

Wong, P.-H. (2020). Democratizing Algorithmic Fairness. *Philos. Technol.* 33, 225–244. doi:10.1007/s13347-019-00355-w

Woo, M. (2019). An AI Boost for Clinical Trials. *Nature* 573, S100–S102. doi:10.1038/d41586-019-02871-3

Author Disclaimer: Where authors are identified as personnel of the Biobanking and BioMolecular resources Research Infrastructure (BBMRI-ERIC), the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of BBMRI-ERIC.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Bak, Madai, Fritzsche, Mayrhofer and McLennan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 08 July 2022
doi: 10.3389/fgene.2022.902631

Ethical Implications of e-Health Applications in Early Preventive Healthcare

Mandy Stake[1]* *and Bert Heinrichs*[1,2]

[1]Institute for Neuroscience and Medicine: Brain and Behaviour (INM-7), Jülich Research Center, Jülich, Germany, [2]Institute of Science and Ethics (IWE), University of Bonn, Bonn, Germany

As a means of preventive medicine early detection and prevention examinations can identify and treat possible health disorders or abnormalities from an early age onwards. However, pediatric examinations are often widely spaced, and thus only snapshots of the children's and adolescents' developments are obtained. With e-health applications parents and adolescents could record developmental parameters much more frequently and regularly and transmit data directly for ongoing evaluation. AI technologies could be used to search for new and previously unknown patterns. Although e-health applications could improve preventive healthcare, there are serious concerns about the unlimited use of big data in medicine. Such concerns range from general skepticism about big data in medicine to specific challenges and risks in certain medical areas. In this paper, we will focus on preventive health care in pediatrics and explore ethical implications of e-health applications. Specifically, we will address opportunities and risks of app-based data collection and AI-based data evaluation for complementing established early detection and prevention examinations. To this end, we will explore the principle of the best interest of the child. Furthermore, we shall argue that difficult trade-offs need to be made between group benefit on the one hand and individual autonomy and privacy on the other.

Keywords: e-health, AI, pediatrics, preventive health care, early health examinations, ethics, best interest of the child, group benefit

OPEN ACCESS

Edited by:
Aviad Raz,
Ben-Gurion University of the Negev, Israel

Reviewed by:
Clémence Pinel,
University of Copenhagen, Denmar
Consolato M. Sergi,
Children's Hospital of Eastern Ontario (CHEO), Canada

***Correspondence:**
Mandy Stake
m.stake@fz-juelich.de

Specialty section:
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Genetics

Received: *23 March 2022*
Accepted: *17 June 2022*
Published: *08 July 2022*

Citation:
Stake M and Heinrichs B (2022) Ethical Implications of e-Health Applications in Early Preventive Healthcare. Front. Genet. 13:902631. doi: 10.3389/fgene.2022.902631

1 E-HEALTH IN PREVENTIVE HEALTH CARE IN GENERAL AND IN PEDIATRICS IN PARTICULAR

According to advocates, big data and AI can dramatically improve preventive healthcare, help establish networks linking patients' experiences and experts' knowledge, and bridge the gap between research and individual therapy (Ehrich et al., 2018). Yet at the same time, there are serious concerns about the unlimited use of big data in medicine. Such concerns range from general skepticism about big data in medicine to specific challenges and risks in certain medical areas (Summa et al., 2020). In this paper, we will focus on preventive health care in pediatrics and explore ethical implications of e-health applications.[1] Specifically, we will address opportunities and risks of app-based data

[1]For the sake of simplicity, we will refer to e-health in what follows, even though this category is broad and not very specific. We use e-health as an umbrella term for a wide range of means for data collection and data analysis. In particular, this includes so-called mobile health (m-health) applications as well as deep learning technologies for searching for unknown patterns.

collection and AI-based data evaluation for complementing established early detection and prevention examinations. To this end, we will explore the principle of the best interest of the child. Furthermore, we shall argue that difficult trade-offs need to be made between group benefit on the one hand and individual autonomy and privacy on the other.

Big data and AI have long since reached medicine (Yang et al., 2021). This is no more than a truism. Nevertheless, implementation in everyday medical practice is only just beginning and many questions–including ethical ones–are still unanswered. Policymakers are strongly promoting e-health because they see it as a unique opportunity to improve medical care and because they hope to reduce costs in the medium and for long term. A case in point is the European Commission's *e-Health Action Plan 2012–2020* which describes e-health as a more personalized, targeted healthcare that can be more effective and efficient, while also facilitating equality and patient empowerment (European Commission 2012). In a similar vein, the World Health Organization underlines the important role of digital technologies for the achievement of universal health coverage and for reaching the Sustainable Development Goals (WHO 2011). The hope for better health care does not seem to come out of thin air. Evidence shows that health apps can improve the efficiency and quality of health care while reducing costs (Bates et al., 2018). Ethical concerns must not be ignored, however, but should be taken into account from an early stage on in order to find appropriate solutions that ultimately increase the quality of medical care and perhaps even reduce costs. The high relevance of e-health applications as part of an increased interconnectivity and availability of medical data is supported by a political and social agenda. However, it also points to the interest of other actors, such as app providers and medical institutions, in the health-related data market, seeking potential monetary gains and possibly power through surveillance (Zuboff, 2019; Sadowski, 2020). When data is used as capital, in particular in the medical context, specific ethical concerns arise. Ensuring informational self-determination and data protection is certainly among the greatest challenges of e-health approaches. However, other ethical principles with which medical ethics has long operated should also be considered (Beauchamp and Childress, 2019). An attempt to ethically evaluate e-health applications in the context of big data also needs to bear in mind the political and social dimensions as well as the theoretical concepts of health, disease and normality. Moreover, power relations and interests of particular organizations, corporations, social groups (children, parents, physicians/researchers), and of other stakeholders like politicians or lobbyists are relevant. Again, the commercialization of medical data and the "technocratic power" (Sadowski 2020) over values, social goods and decisions about what ways of data extraction, data gathering, and data evaluation are acceptable, is of critical relevance in this context. In this study, although important, these dimensions can only be addressed on the sideline; they are, however, discussed more thoroughly, for example, in Deborah Lupton's *Digital Health* (Lupton, 2018) or more recently in Jathan Sadowski's *Too Smart* (2020).

The idea of using AI in medicine is older than one might think. Discussions about the implementation of AI can be traced back at least to Paycha (1968). In the specific context of pediatrics, one of the first approaches date back to 1984 when Kohachiro Sugiyama and Yasuhiro Hasegawa introduced the computer assisted medical decision-making system SHELP. Despite this history, pediatrics has received comparatively little attention in e-health initiatives so far. One reason for this could be purely practical, as e-health applications are not yet very pediatrician-friendly and require specific knowledge and information technologies that have yet to be deployed (cf. Kokol et al., 2017: 4). This is in line with the typical pattern that an increase of medical knowledge usually first leads to practical improvements for adults and is only later implemented in the field of pediatrics. Another reason could be that medical care of minors always involves special ethical and juridical challenges. Minors are considered a "vulnerable group" for whom particularly high levels of protection apply. However, the status of a vulnerable group can also be used as an argument that medical care needs to be improved particularly urgently. Children and adolescents should certainly not be deprived of possible improvements in medical care out of excessive caution. At first sight, the approach to improve mobile health (m-health) data collection via apps supported by mobile communication devices like mobile phones, tablets, personal digital assistants (PDAs) or smart watches seems to be promising. The data collected by these means could be analyzed in combination with AI algorithms. In fact, there already is a growing number of apps for monitoring children's health. Caregivers have the choice to use apps for a variety of topics, including infant care issues, mental health information and support, oral health knowledge, diabetes control, asthma monitoring, management of acute pain, overweight management, or oncologic symptom monitoring (cf. Radovic et al., 2016; Alqarni et al., 2018; Chatzakis et al., 2019; Seidman et al., 2019; Hsia et al., 2020; Martínez et al., 2020; Tragomalou et al., 2020). For monitoring development parameters, parents can choose from a number of apps as well. A search in app stores leads to several apps offered by universities, startups or multinational electronics companies with varying ratings, costs, and features. Although 58% of mobile phone users already downloaded a health-related mobile app as of 2015 (Krebs and Duncan 2015) and an ever-increasing demand is being noted (Carroll et al., 2017; Stewart 2021), several reviews show that a huge number of poor-quality apps, especially information apps and tracking apps that parents use for their children, makes choices difficult (Richardson et al., 2019; Virani et al., 2020): The outcome and the quality of apps depend on the task or goal that they were created for. Generally, m-health applications can be (i) used as data collection platforms and (ii) the collected data can be used for informational purposes in medical practice and healthcare. While there are apps that serve only one of these two purposes, in practice they are often intertwined (a point we address below in **Section 4**). More specifically, these applications also differ in their purpose for or effect on the user: for instance, some apps can influence the user's choices about what to do (e.g. symptom tracking apps, tracking medication usage), the user's moods (e.g. mental health apps), or the user's general experience of the interface with the

app (e.g. chatting with bots, tracking certain parameters, checking health status). However, many m-health applications lack reliance for right symptom tracking and evaluation, which opens up the possibility of incorrect diagnosis, but also potentially endangers users by not mirroring and even trivializing a given health problem, as for instance chatbots or so-called "conversational agents" in mental health apps with their repetitive and scripted responses.[2] Low quality also shows whenever an identification of sources is not available or vague, or when there is a lack of current information which lowers the credibility of the information provided (Richardson et al., 2019). Moreover, the majority of the applications are not tested by official regulatory bodies or a patient community, which should be taken as a reason not to rely on them too much at present. The fact that user groups provide data in an uncontrolled and unsystematic manner would also be problematic and should be seen as a lack of quality. Mobile health app usage has shown to differ largely regarding age, education, and e-health literacy skills (Bol et al., 2018), which again can heavily influence the evaluation of the collected data. Thus, it is important to keep in mind that quality assessment is a necessary step for the implementation of m-health on a broad level.

However, when we think about the possibility to use m-health applications in a controlled way and in collaboration with given in-person early examinations, it could still be particularly promising to complement the screening of children, which is carried out on a regular basis in many countries, with e-health solutions. As a means of preventive medicine, early detection and prevention examinations could thus identify and treat possible health disorders or abnormalities from an early age on. Nevertheless, pediatric examinations are often widely spaced, and thus only snapshots of the children's and adolescents' developments are obtained. This is one reason why the amount of data in pediatrics is very limited overall. With the current resources of e-health applications, parents and adolescents could record developmental parameters such as weight, height, social interactivity, language usage, or behavior patterns much more frequently and regularly, and transmit these data for ongoing evaluation. In addition, AI technologies could be used to identify previously unknown correlations which, in turn, could lead to improved diagnosis and treatment.

2 EARLY DETECTION AND PREVENTION EXAMINATIONS IN PEDIATRICS

Regular health screenings are an essential component of pediatrics providing important information about children's and adolescents' status of health and development, and thus providing early detection of diseases but also cases of neglect, maltreatment, and abuse. Many countries around the world have child health screening programs that provide primary health care, preventive screenings and immunizations. Looking at the European Union, there are such programs, for example, in the Netherlands centrally provided by institutions of the youth health authority, the "Consultatiebureau" (cf. NL Ministry of Health 2022); in Austria (cf. KBGG (2021): § 3; MuKiPass 2002: § 2), and in Germany. For a better understanding of these programs, we describe the situation in Germany in more detail.

In Germany, institutionalized early detection and prevention examinations in pediatrics has existed since 1971. All children are entitled by law for regular screening examinations known as "U-Untersuchungen" (U-examinations) until the age of 18. These examinations serve the early detection of diseases that pose a significant risk to the physical, mental, or psychosocial development of the child and are regulated in the *Guideline of the Federal Joint Committee on the Early Detection of Diseases in Children*, or short: Children's Guideline ("Kinder-Richtlinie") (cf. Kinder-Richtlinie 2022: §1 (1), p. 6). They are usually performed by a pediatrician or family doctor and take place at fixed time intervals. The U-examinations include physical examinations as well as assessments of the child's cognitive, social and emotional competencies, including a variety of parameters depending on the child's age, as well as a consultation with the parents. In addition, special screenings are conducted for specific diseases. Moreover, a child's vaccination status is assessed. (cf. BMG 2021). The examination results and vaccination status are registered in a standardized child examination booklet, which contains a removable card so that parents can prove to third parties, such as kindergartens, that their child regularly attends the U-examinations without disclosing confidential information (cf. BMG 2021). However, the screenings fall under the regulatory purview of the states and are only mandatory in some states (e.g., in Bavaria, Hesse, and Baden-Württemberg since 2008/2009),[3] while voluntary in other states (e.g., in Berlin, Saxony, or North Rhine-Westphalia).[4]

Early preventive health examinations are an important health reporting tool that was designed to gather more relevant medical data in pediatrics. This was a first step to reduce the lack of data

[2]According to preliminary evidence, chatbots have been found to be potentially beneficial, enjoyable and helpful when connected to proper research and in-person treatment; yet the study reviews are highly heterogenic and more research with standardized outcomes is required for a proper assessment (Vaidyam et al., 2019).

[3]In these states, laws ensure participation in screenings through data transmission from the registration authorities and pediatricians. If the screenings do not take place, a written invitation is sent, and if this is not followed, the youth or health office can be informed.

[4]However, some of these states now have more far-reaching structures, as for example North Rhine-Westphalia: The notification procedure may provide the local public youth welfare agency with additional indications as to whether and which families may need support services to ensure the best interests of the child (cf. UTeilnahmeDatVO: § 1 (2)). The data can be provided by the physician who performed a health examination in a secured written form via secured data transmission channels to the *North Rhine-Westphalia State Center for Health*. If necessary, the latter may inform the local public youth welfare organizations (cf. UTeilnahmeDatVO: §§ 2–4). Regardless whether or not screenings are mandatory or voluntary, there is a country-wide free reminder service provided by the *German Association of Pediatricians and Adolescents* (*Berufsverband der Kinder-und Jugendärzte e. V.*; short: BVKJ) in order to help parents keeping their children's vaccination and screening appointments in good time by e-mail. Independently of the U-examinations, there are compulsory school entry examinations in all German states.

that persists in pediatrics overall. But although preventive services for children and adolescents are provided in various forms throughout Europe, the amount of pediatric health data is still limited and scattered, i.e., there are data gaps due to often widely dispersed studies. This fact significantly limits not only pediatric health care, but also pediatric research. To be sure, there are attempts to fill these data gaps by scientific studies and regular repeat surveys. In Germany, the most comprehensive study of this type is the "Study on the Health of Children and Adolescents in Germany" (KiGGS, 2018) conducted by the Robert Koch Institute (RKI). This study is carried out over a period of many years and aims at gaining nationally valid, representative data on the health situation of children and adolescents. In addition, other national and international studies and surveys provide insights into children's health and development. The Information System of Federal Health Reporting (IS-GBE) provides a constantly growing data pool in the form of an online database (cf. BZgA 2022).

3 THE POTENTIAL OF E-HEALTH APPLICATIONS TO COLLECT CHILD HEALTH PARAMETERS

As mentioned above, the use of e-health applications is on the rise. Health apps can improve the efficiency and quality of health care while also reducing costs (cf. Bates et al., 2018: 1975–6). In particular, such applications can help to collect and analyze medical data. Therefore, the use of e-health applications in pediatrics seems very appropriate. Many people of today's parent generation are tech-savvy, which makes the collection and transmission of data via smart phones or internet-based software potentially easy to implement. In general, such e-health approaches offer an opportunity to move away from treatments based purely on pattern-based decision making and summary statistics to more individualized approaches and to make more accurate decisions based on more comprehensive data sets (cf. Mayer-Schönberger and Ingelsson 2017: 428). In pediatrics, this would mean that therapeutic measures for individual children could be initiated much earlier and easier than today. Moreover, such approaches could ensure that priorities for epidemiological and health policy measures are identified and surveyed more quickly and studies on child health in all fields could be intensified (cf. Ehrich et al., 2018: 488). In addition, new ways of data collection would allow for a better monitoring of changes in individual parameters and more regular time intervals. AI technologies could then be used to search for new and previously unknown patterns (cf. Ehrich et al., 2018: 491). Eventually, a new data collection could evolve, such as a "Wikipediatrics" where patients' experiences and experts' knowledge ranging from clinical research to care research and individual therapy could be represented (cf. Ehrich et al., 2018: 495). This would be a new way of storing and using knowledge for pediatricians enabling them to quickly look up simple parameters, illness factors, correlations, or diagnosis suggestions. In conclusion, the use of e-health applications in pediatric screening seems to have great potential.

4 ETHICAL CONSIDERATIONS

Regardless of the possible benefits described above, there are serious ethical challenges to be considered. They range from general concerns about big data in medicine to more specific issues related to minors. The idea of using app-based methods to monitor the development of children and adolescents, to use predictive knowledge, to monitor health, and to provide data on development and social status faces difficult trade-offs. In general, there are severe ethical issues concerning data extraction, data usage and data safety which we will come back to in the following section. Yet in particular, questions arise about the best interests of the individual child and his or her informational self-determination: If it turns out that the use is not or not always in the best interest of the individual child, then e-health applications could possibly be justified by reference to a group benefit. In this case, balancing issues would arise. We shall discuss these concerns in turn after the outline of some general problems.

4.1 e-Health and Big Data in the Medical Context

e-health, and more specifically: m-health, is part of a big data policy in the medical context promoting unique opportunities and efficient improvements in medical care while reducing costs (Bates et al., 2018). They are introduced as a means to collecting and evaluating additional health data as well as giving advice for preventive measures. Yet, as already mentioned above, ensuring informational self-determination and data protection is among the greatest challenges of this approach.

The m-health applications already available serve different purposes and goals. The large number of these applications shows the economic relevance: Data can be used to generate profits. Yet, any data acquired from or by the user can eventually generate profits. Moreover, there is a fair chance that the possibilities to *understand* procedures and to participate in decision making are even more impaired in the medical context than in other contexts: With regard to Big Data, the various purposes for data usage are diffuse and often mix without clear boundaries so that previously separate areas can merge and link information to a health context that was previously not considered relevant (cf. Summa, 2020: 98; Braun and Dabrock 2016: 326). These merges could arise, for instance, by linking health data to lifestyle choices or social environment data from social media, forums, blogs, or specialized communities (Rüping 2015: 794; Krüger-Brand 2015: A1026f.; Müller and Samerski 2016: A1749). Furthermore, the interconnectivity of the data on platforms and devices can make all personal data potentially health related (Bächle 2019: 48). Given these interconnected structures, the chance that data could be re-identified (even if properly anonymized before) increases, so that in turn breaches in data security can hardly be excluded. This results in an enhanced risk for informational self-determination because such cross-data connections may lead to possible discriminatory factors and individualization based on personal background information, as for instance capital assets, lifestyle, or living situation, affecting predictions, recommendations, therapy

suggestions as well as the access to and quality of health care services. Especially data transfers have a higher potential for data transgressions which again can lead to the danger of "surveillance capitalism" (Sharon 2018; Zuboff 2019; Tsakiliotes 2021), lower credibility and lower quality of the provided applications.

But when data is (also) used to generate economic profits, particular ethical concerns arise: Not only could data extraction, especially in the money-spinning medicine market, be another stabilizer of the much-discussed problem of a "digital capitalism" (Sadowski 2020) since personal and sensitive health data could be used as currency to create profit for the app providers. What is more, the content of the data evaluation based on data gathering in large data pools can be exploitative when provided and used by corporate actors (Sadowski 2019), and can breach data safety and personal consent, if passed on to other parties as, for instance, to insurance companies that already use data to assess risks and profits and thus could gain even more regulating power and authority in the private lives of the concerned persons (Sadowski 2020: ch. 6).

As was shown, most citizens–and this applies already to adults–do not have explicit knowledge of how their data is being used and how related decision processes take place (Summa, 2021: 113; Sadowski 2020: thesis 4); this is even less the case for children. Thus, it seems that the pure collection of more health data is not enough to argue for better early preventive health care. To the contrary, the pure collection of data without evaluation is not of any value for the app users and thus does not fairly compensate them, which makes this practice at least ethically questionable, Sadowski would even say "exploitative" (2020; thesis 4). Rather the data and the analysis of the data need to be critically appraised (Brault and Saxena 2021: 514), interpreted and evaluated to be valuable for the individual data provider. However, thus far it is not certain if these are feasible tasks in the context of app-based AI in general. In addition, it is unclear how this would influence and shape the scope for the concept of the child's best interest.

4.2 e-Health Applications and the Best Interest of the Child

The concept of the best interest of a person is complex and encompasses aspects of both physical and psychological well-being. In the context of medical and research ethics, the concept can serve as a normative standard for justifying decisions affecting individuals (e.g., Taylor 2016). While being able to live a self-determined life may be seen as a core element of a person's best interests, there can also be a conflict between subjective desires and what is objectively best for a person. Self-determination can be viewed as an ideal that consists of the "freedom to think, choose, and act on one's own life path" (Akbar 2019: 9). This ideal implies that a person's well-being is expressed, among other things, in living their life as they see fit and has value in the larger context of social well-being and equality (Krutzinna 2022: 129). However, medical needs may sometimes not comply with a person's wishes in order to serve his or her best interests. Nevertheless, major interventions in the self-determination of adults are today generally rejected as paternalistic. This is to say that the best interest of adults today is usually interpreted in individualistic terms and thus dissolved into self-determination. With children, the situation is more complex. The concept of "best interest" plays a more important role here, as their capacity for self-determination is only gradually developing, so that what is in the child's best interest cannot generally be identified with the child's own wishes and ideas, i.e. what lies in their self-interest. Often, fulfilling children's wishes is clearly not in their best interest.

In determining what is in the best interest of the child, parents or guardians play a key role. They have a wide scope for decision-making, which is, however, limited by objective factors. Especially with young children, parents alone must decide what is best to do. As they grow older, the views of the minors themselves become increasingly important. It can be particularly difficult to resolve the tension between the right to informational self-determination of children on the one hand, and measures to protect the child's health on the other. At the same time, a parent's refusal to take medical action may cause harm to a child and therefore be considered a violation of custodial duties and a lack of responsibility. This tension corresponds to the inherent conflict between the basic ethical principles of beneficence (or non-maleficence) on the one hand and autonomy on the other.

A thoughtful understanding of a child's best interest is presented in a recent paper by Jenny Krutzinna (2022). She argues that "despite a bona fide belief that we are assessing a child as a unique individual, with individual needs, traits and preferences, we continue to make many generalizations and category-based assumptions in determining the child's best interests." (Krutzinna 2022: 121) According to Krutzinna, a way out of this oversimplification and categorization of "the child" as a homogenous group is a concept that she calls the "model of the individual child" (MIC) that highlights the individuality and uniqueness of a child. This model does not dismiss universal and group-specific characteristics about and comparisons between children, but complements these approaches with an even more child-specific point of view that takes into account the specific character, background, likes and dislikes of the child who is thus seen as the individual person he/she is. In contrast to other approaches, this focus can help to prevent serious misjudgments about what is in the best interest of a particular child (cf. Krutzinna 2022: 123, 127, 141).

What follows from such an approach for the use of e-health applications for child screening? On the one hand, one could draw the conclusion that the interests of children would be particularly protected and supported by e-health applications in child screening since the main goal of their use is precisely a more individualistic approach based on the individual parameters. However, whether such an individual benefit exists and, if so, how big it is, is yet an open question. On the other hand, there is a further restrictive conclusion, since the feasibility of a child specific screening supported by e-health applications would have to be examined and evaluated for each individual case, i.e. whether this approach would be in the child's best interest, whether the benefits outweigh the disadvantages, and what the short-, medium- and long-term effects on the child's informational self-determination are. Such detailed examination

would arguably render the use of e-health applications in child screening impossible, because they can only be operated effectively if they are applied on a large scale. There is also reason to fear that the vertical asymmetry between adults and children is initially reinforced by such applications, as children are unlikely to be able to understand how they work and what their benefits are at first. This is certainly especially true for young children and may change with age.

Thus, in order to balance the right to informational self-determination on the one hand and medical needs on the other, as envisioned by the concept of the best interest of the child, we suggest that it is essential to develop age-dependent models that take special account of the vulnerability of children. Whenever possible, children should be involved in the use of apps, and they should have the opportunity to have a say in what data is collected and with whom it is shared, of course depending on age. As they get older, children should be allowed to determine more and more for themselves the extent to which such applications are used. These ethical requirements should already be considered when designing such applications.

If one assumes that the benefit for the individual child is rather small, does this automatically mean that the use of e-health applications for early diagnosis is ethically unjustifiable? This conclusion would be premature, as there are other areas where moderate violations of the best interest of the individual child are justified by an overriding group benefit. Therefore, this line of reasoning will now be examined.

4.3 Individual Benefit Versus Group Benefit

Originally, the concept of group benefit was introduced in the context of clinical trials. The difficulties and the extent of inclusion of children in research have been discussed broadly (Binik 2018; Kantin 2020). It was particularly difficult to justify the enrollment of minors according to established standards, at least if no direct benefit for participating children was foreseeable. However, to completely prohibit participation in studies without direct benefit to minors would have significantly impaired pediatric research. A way to avoid this consequence was that under certain conditions, group benefit can be a justification for accepting risk or some harm to individuals. For instance, group benefits can be used in addition to individual child protection to justify mandatory vaccinations for children attending kindergartens or schools (see Summa, 2020: 87; Xafis et al., 2019: 235, 238, 247; Winkler 2017: 27). This is a classic trade-off between security for the many on the one side and autonomy for the individual on the other side. Considering research involving minors, the concept of group benefit allows for more flexible trade-offs in certain situations than the strict consideration of the authenticity of every child (cf. Radenbach 2006; Löschke and Heinrichs, 2015).

In the case of an app-based approach in pediatrics, more comprehensive data collection and data evaluation could also be justified with reference to an overwhelming group benefit. For example, children often continue to receive medications "off-label" and the dosage is often based on the dosage for adults, as reliable data for children is lacking (cf. Summa, 2020: 92; Steinmann et al., 2016: 19; Heinrichs et al., 2016). Furthermore, it has been argued that with the use of apps and digital infrastructure, risks for children could be better captured and lead to more research data and better access to existing knowledge (see e.g. Rüping 2015). Increased initiatives could even promote "deep medicine," as Eric Topol (2019) suggested. As a concept, deep medicine suggests that AI has the potential to assist physicians in everything they do and to establish a more empathetic and trustful physician-patient-relation that today often suffers because of time-limits. Also, e-health apps, so the argument, could have this assisting quality, which could, in turn, be particularly fruitful in the pediatric context (cf. e.g., Ehrich et al., 2018; Li et al., 2022) and eventually improve individual patient-specific care and research (Morris et al., 2021). All of these points are to a great benefit for the group of children. However, the flip side must also be considered. Although vulnerable groups such as children should not be excluded from research, excessive data collection may violate privacy rights and informational self-determination as has already been pointed out above. In the context of data collection, this primarily relates to the lack of controllability of the flow of information in data-driven medicine and reflects the output orientation of governance and policy, as Patrik Hummel and Matthias Braun (Hummel and Braun, 2020: 1f.) have recently noted. Thus, the concept of group benefit must be applied very mindfully. To gain more clarity, it is useful to list the different stakeholders involved and the potential benefits they might have. There are at least four main groups that need to be distinguished:

(i) researchers and physicians who could benefit from data collection by filling research gaps, finding new associations, enabling even earlier detection and prevention methods, and thus creating better and more individualized treatments;
(ii) (ii.a) individual children and (ii.b) their parents–the data providers–who might not immediately or directly benefit from better treatment options;
(iii) (iii.a) (future) children and (iii.b) their future parents, who are future data providers and could benefit from better treatment options;
(iv) other stakeholders who might profit from the data financially or through power gain, like e.g. insurance companies, corporations, lobbyists, app-providers, etc.

There are at least two further aspects which are to be considered consecutively: (1) the problem of bias that relates to the already addressed issues about data quality, interpretation and classification up above, and (2) the impact of e-health applications on the trust relationship between physician, patient and parents.

(1) In e-health applications for early detection, medical data points would be collected either automatically or manually by users (parents or adolescents themselves). However, recent studies show that the quality and validity of the data sets based on these data points via cell phones or wearable devices such as smart watches are rather poor since they are often unstructured and full of random or systematic errors due to different types of sensors, conditions, or variations in

applicability, which make any interpretation or result based on them likely to be biased (cf. Brault and Saxena 2021: 514f). In fact, bias can enter in various forms and at various stages: (i) in the problem definition according to the developed algorithm, (ii) in the social or technical intervention where certain types of data sets can be incomplete, under- or overrepresented, (iii) when the feature selection is unevenly distributed across different groups, (iv) because of the model's dependency on the data sets, (v) model selection and its accuracy, (vi) design of the user interface and user directory (Brault and Saxena 2021: 515f.). This calls into question the comprehensibility of results as well as of conclusions based on these datasets (McDougall 2019). Furthermore, if cross-sectional data is also collected, conclusions could be even more problematic than only sectional data since it increases the amount of possible errors and incomprehensible conclusions, which not the least raises questions about replicability and reproducibility of the results (Brault and Saxena 2021: 514). There is another aspect to consider with cross-sectional data evaluation: The efficiency of the algorithms of e-health applications relies on "grouping", i.e., on putting individuals into groups according to group-specific characteristics which is, again, a risk factor for bias. The number of characteristics is increased when cross-data connections are included which, in turn, can promote higher intransparency than it would be the case if cross-data connections would not be used. For example, if 10-year-old Betty's social competence, psycho-social development, or language competence is not only tracked by manually entered information in a specific medical app, but also automatically by data from her social media usage time, the postings or pictures she likes or comments on, the music she listens to, and the language she uses in the messages she writes, then this would be a case for cross-data connection. Another example would be if an algorithm puts different children in the same group with higher risks to develop a certain disease, say asthma, and generates treatment or help suggestions, only because they are living in a certain area or have a particular social background, which is based on information that comes from multiple app-trackers but is not necessarily comprehensible since the information of the conclusion cannot be deduced and followed back to the particular apps. A third example–a risk if data is used for early detection or prediction of possible diseases or increased health risks–is that a child could be categorized as part of a certain group before a disease has actually manifested. Not only should this knowledge be sufficiently protected from access by others, but it should also be treated as confidential and with care since the mere knowledge about a certain disposition to develop a disease can be harmful and may lead to self-stigmatization. In fact, knowledge about a potential increased risk for a disease or a probability-based prediction for a future medical condition can already decrease the person's well-being (Bächle 2019: 51f.).

There is controversy about how respective protection measures are or can be implemented in app-based AI-applications. A further critical point of grouping in general is that these groups might not be stable because individuals can move from one group to another quickly depending on new data points. This importantly differs from other forms of grouping supervised by researchers as for instance is the case in medical studies. Ad-hoc groups put together by cross-data connections can be thus more biased and inclusion can be more unfair to individuals than usual data evaluation methods due to automatic or manual inputs (by the user) that are insensitive to the sample size (cf. Brault and Saxena 2021: 514). Note that it can be difficult to notice unfair or harmful grouping (cf. Mittelstadt 2017: 481).

Then again, not only the linkage or reconnection of data, but also the decoupling of data can lead to problems: algorithms for data evaluation can also decouple the presence of traditional disease symptoms from medical diagnosis and then be a hindrance for appropriate recommendations and measurements. One common consequence of these issues is that under- or overtreatment is likely to occur based on e-health applications since their conclusions are likely jeopardized by bias issues.

All this shows that there are many ways in which cross data connections gathered by e-health applications can lead to "informational harm" (Richter and Buyx 2016: 316). Informational harm refers to the occurrence and dependence of highly questionable results based on biased algorithms, which may result in over-, under- or other forms of mistreatment. Informational harm can also include the risk of information loss and discrimination, which is especially problematic for people who already belong to vulnerable groups, as is the case of children. Therefore, data collection and recommendations for preventive measures based on data sets may create an increased risk for mistreatment and incorrect decisions, especially for members of groups considered most vulnerable (Braun et al., 2021: 3). Only if the data is evaluated by trained physicians in collaboration with medical informatics and data scientists, it seems reasonable to expect an improvement of medical preventive care (cf. Daniel et al., 2019; Durán and Jongsma 2021).

(2) e-health applications can only be successfully implemented if pediatricians as well as parents and children have confidence in and can rely on their safety and efficiency (cf. Bates et al., 2018: 1975–6). What is more, the reliance on safety and efficiency is also likely to have an effect on the doctor-patient relationship, where trust is a central element. In pediatrics, the relationship and decision-making processes are more complex because three parties are involved: the minor patient, the physician, and the parents. Usually, children trust their parents in making the right decision for them, while trust towards the physician has to be built up. An essential factor for this is the parents' trust in the doctor. Another factor that can contribute to the child's trust in the doctor in the long term is habituation during visits, in particular during the regular check-ups described above. However, the relationship of trust can be disturbed, especially if children or their parents have the impression that the child's interest is not paramount. E-health applications could fuel such an impression if data collection and use are not transparent. The providers of e-health applications, as for instance, the commercial companies developing the apps, the data storage and integration centers, but also the app interface itself, can have a mediating role in the traditional relationship between patients and physicians. This can have positive effects, like the support for the physician via accessibility of data, or recommendations based on data, or the immediate support and help for the app user. However, it can

also compromise this relationship because trust between the parties is waning. Considering the data collection practice via apps and the use of interpretation of whole data sets based on collected data points, a lack of transparency in one of the factors could undermine trust in the physician: On the side of the patients, trust in the physicians is partly based on their know-how and understanding of the recommended applications, but also on the usage, sharing, accessibility and confidentiality of the output-information which is not provided directly by the familiar physician but rather by an accessible medical platform. As has been discussed above, new forms of data connection and sharing can easily threaten data safety, but also software viruses and hacker attacks can breach this safety. When sensitive health data is leaked because it was not protected by multiple security levels, this increases the risks for re-identification and possible discriminatory practices.

This indicates that the extensive use of individual data to advance medical knowledge for the benefit of patients may result in today's patients and their parents having less trust in physicians. First and foremost, this applies to the trust between the engaged adults, i.e. the physician and the parents. The potential data safety might not be a fundamental concern in the relationship between the children and the physician since theirs is more based on the perceived goodwill of the physician towards the children. This might, however, change when children get more awareness and start to use the provided e-health application for self-tracking at some point. Here, understanding and informed consent have to be considered more thoroughly since the awareness about potential conflicts and effects of e-health applications has an influence on the level of understanding, which is necessary to consent to that praxis in a meaningful and informed way. If parents and/or children do not understand the praxis, consent is not informed. In this case, however, the decision on whether the use of e-health applications is in the best interest of the child has to be reconsidered. This is even more the case, if parents or older children would *only* rely on e-health applications without a physician to evaluate the output–which is an increased danger if e-health applications are incorporated in every-day use, and not in relation to the clinical context. Thus, extensive use of individual data to advance medical knowledge for the benefits of patients now and in future could result in today's (child) patients and their parents, having less trust in physicians and thinking that individual well-being is not the primary concern. This could have an overall negative effect and harm both current and future children, which is important to consider when weighing group benefits. In this light, it seems to make only limited sense to justify the app-based collection of medical data by referring to group benefit.

Note that although there already are data protection concepts developed by data integration centers (cf. Prasser et al., 2018: e57–e65; Mansmann et al., 2020: 30), it is often still unclear how the volume and heterogeneity of the data is to be evaluated and used specifically since the necessary theoretical knowledge and standards for meaningful validation, analyzing, and interpretation of the data is still lacking to create useful infrastructures in the medical context (cf. Krüger-Brand 2015: A1026f.). Health-related data generated by and saved in more secure and regulated environments using laboratory information systems (LIS) differs from other (commercial) self-tracking apps or devices; however the ethical challenges seem to be related to similar issues with only more or less severity: informed consent, privacy, control over personal data and the interpreted output based on the given data (Bächle 2019: 49).

Making sense of data is a complex process in which multiple stakeholders are involved (Neff et al., 2017): we need more comprehensive critical data studies, take technical critiques as a way to actively discuss and contribute to the betterment of these app, and thus improve outcomes by tackling the challenges in an interdisciplinary way, bringing together, scientific and practical knowledge, but also taking into account social dimensions and ethical expertise.

In summary, then, group benefit is a relevant ethical concept that is partially suitable for justifying measures that commit individuals to the benefit of a group or society as a whole. Especially in pediatrics, this concept can be used to justify interventions. It is, however, important to ensure that recourse to group benefit in the discussions does not disrupt trust between physicians, patients, and parents. In addition, the true group benefit must be carefully examined and biases, to which data obtained through app-based applications is particularly susceptible, must be minimized. Otherwise, group benefit could easily turn into group harm.

5 OUTLOOK

Lindsey Knake (2020: 2) recently raised the question "Are we ready for AI in pediatrics?" and answered it herself with "not completely". In this paper, we have specifically discussed ethical implications of e-health applications in early preventive healthcare. We agree with Knake that there are still many challenges that have to be overcome. We further agree with the assessment that existing e-health applications are not readily transferable to the pediatric setting (Kelly et al., 2021: 1). According to our analysis, these challenges revolve around familiar questions about tradeoffs between public benefit on the one hand and individual autonomy, privacy, and freedom of choice on the other. In pediatrics, however, these trade-offs are even more problematic because children belong to a particularly vulnerable group that must be treated with special care and attention. In addition, the physician-patient relationship is more complex in the case of children because parents are involved as another party. It is also important to bear in mind that this is a dynamic relationship in which children mature more and more into self-determined individuals whose ideas become more essential and significant for any decisions in their best interest as they grow older.

Since one of the main reasons for decreased trust seem to be transparency and data safety issues, we need to ensure increased insight in and education about e-health, access rights, and social and legal regulations for patients, parents, care holders, physicians and partaking other stakeholders like app-providers, data platforms etc. Moreover, a comprehensive data ethics needs to provide a framework so that data usage is based on

the principles of autonomy, beneficence, non-maleficence, and equal and just access (Summa, 2020: 86 and Xafis et al., 2019: 235, 245). Furthermore, the quality of data and algorithms must be assessed thoroughly. Brault and Saxena (2021: 516f.) have recently highlighted the need of a catalogue of bias, the development of methodological standards for the use of big data and AI in the medical context encompassing the principle of explicability and an ethical sense of accountability, but also the development of a critical appraisal of AI and big data in medicine. This call can only be stressed with regard to pediatrics. In addition, Clémence Pinel et al. (2020) have addressed the contextual social embedment and relational features of data (that are "never raw") and shed light on how to do "data work" more carefully to contribute in the knowledge production.

These are all important ways to make data usage–and in the long run potentially also e-health applications–more meaningful, valuable and ready for use in the pediatric context and to take advantages of the benefits addressed above. This means, the given challenges should not distract from the fact that the collection of data via e-health applications is potentially beneficial in pediatrics. Established preventive health screenings, as for instance the German U-examinations, could indeed be improved through such applications, to the benefit of both individual patients and pediatrics as a whole. One drawback of this point, however, is that in practice not all potential benefits may materialize immediately, but only in the medium and long term, since sufficient data must first be collected and carefully evaluated. Yet, if implemented successfully, such initiatives can be extended not only to the national level, but also to the European or international level. The biggest risk discussed here was that poor data quality and excessive euphoria about technology will lead to exactly the opposite case, namely that e-health applications could hinder or worsen the established health screenings. In the worst case, the use of e-health applications could generate biased data on the basis of which poor decisions are made, and at the same time damage the trust between physicians, child patients, and parents. There have been few analyses of these groups' priorities, and comprehensive data on possible ethical, legal, and social facilitators and barriers for the implementation of these new technological means for pediatricians, parents, and children so far remains scarce: for instance, there are some general remarks for parent's perceptions about mobile technology use of preschool aged children (Genc 2014), and some for the perceptions of young children's, parents' and industry stakeholders' criteria for selecting apps (Dias and Brito 2021). Therefore, preferences, wishes and ideas in the context of e-health applications should be investigated and evaluated in future research. The best interest of the child must remain the overriding ethical principle guiding trade-offs in individual cases. Comprehensive information for parents and children must ensure that the right to self-determination is respected at all times. If this is the case, then e-health applications can be jointly developed and implemented and improve health care in pediatrics in the long term.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

All publishing fees are covered by Forschungszentrum Jülich GmbH with a 15% discount rate as stated in the pre-payment agreement between Frontiers and Helmholtz Gemeinschaft.

REFERENCES

Akbar, G. L. (2019). Thinking Critically about Self-Determination: A Literature Review. *J. Soc. Work Values Ethics* 16/2, 9–17.

Alqarni, A., Alfaifi, H., Aseeri, N., Gadah, T., and Togoo, R. (2018). Efficacy of a Self-Designed Mobile Application to Improve Child Dental Health Knowledge Among Parents. *J. Int. Soc. Prev. Communit Dent.* 8 (5), 424–430. doi:10.4103/jispcd.jispcd_195_18

Archard, D. (2014). *Children: Rights and Childhood*. London: Routledge.

Bächle, T. C. (2019). "On the Ethical Challenges of Innovation in eHealth," in The Futures of eHealth. Social, Legal and Ethical Challenges, *47–55*. Editors T. C. Bächle and A. Wernick (Berlin: Humboldt Institute for Internet and Society).

Bates, D. W., Landman, A., and Levine, D. M. (2018). Health Apps and Health Policy. *JAMA* 32019, 1975–1976. doi:10.1001/jama.2018.14378

Beauchamp, T. L., and Childress, J. F. (2019). *Principles of Biomedical Ethics*. 8th ed. Oxford: Oxford University Press.

Binik, A. (2018). Does Benefit Justify Research with Children? *Bioethics* 32 (1), 27–35. doi:10.1111/bioe.12385

BMG (2021). Gesundheitsuntersuchungen für Kinder und Jugendliche. Bundesministerium für Gesundheit. Available at: https://www.bundesgesundheitsministerium.de/themen/praevention/kindergesundheit/frueherkennungsuntersuchung-bei-kindern.html [status (accessed: 03.02.2022).

Bol, N., Helberger, N., and Weert, J. C. M. (2018). Differences in Mobile Health App Use: A Source of New Digital Inequalities? *Inf. Soc.* 34 (3), 183–193. doi:10.1080/01972243.2018.1438550

Brault, N., and Saxena, M. (2021). For a Critical Appraisal of Artificial Intelligence in Healthcare: The Problem of Bias in mHealth. *J. Eval. Clin. Pract.* 27, 513–519. doi:10.1111/jep.13528

Braun, M., and Dabrock, P. (2016). Ethische Herausforderungen einer sogenannten Big-Data basierten Medizin. *Z. für Med. Ethik* 624, 313–329.

Braun, M., Hummel, P., Beck, S., and Dabrock, P. (2021). Primer on an Ethics of AI-Based Decision Support Systems in the Clinic. *J. Med. Ethics* 47, e3. doi:10.1136/medethics-2019-105860

BZgA (Bundeszentrale für gesundheitliche Aufklärung) (2022). Datenquellen zu Gesundheit und Lebenswelt von Kindern. Available at: https://www.kindergesundheit-info.de/fuer-fachkraefte/grundlagen/daten-und-fakten/datenquellen/ [accessed: 02.03.2022].

Carroll, J. K., Moorhead, A., Bond, R., LeBlanc, W. G., Petrella, R. J., and Fiscella, K. (2017). Who Uses Mobile Phone Health Apps and Does Use Matter? A Secondary Data Analytics Approach. *J. Med. Internet Res.* 19, 4e125. doi:10.2196/jmir.5604

Chatzakis, C., Floros, D., Papagianni, M., Tsiroukidou, K., Kosta, K., Vamvakis, A., et al. (2019). The Beneficial Effect of the Mobile Application Euglyca in Children and Adolescents with Type 1 Diabetes Mellitus: A Randomized Controlled Trial. *Diabetes Technol. Ther.* 21 (11), 627–634. doi:10.1089/dia.2019.0170

Daniel, G., Silcox, C., Sharma, I., and Wright, M. B. (2019). Current State and Near-Term Priorities for AI-Enabled Diagnostic Support Software in Health Care. Duke-Margolis Center for Health Policy. Available at: https://healthpolicy.duke.edu/sites/default/files/2019-11/dukemargolisaienableddxss.pdf.

Dias, P., and Brito, R. (2021). Criteria for Selecting Apps: Debating the Perceptions of Young Children, Parents and Industry Stakeholders. *Comput. Educ.* 165, 104134. doi:10.1016/j.compedu.2021.104134

Durán, J. M., and Jongsma, K. R. (2021). Who Is Afraid of Black Box Algorithms? on the Epistemological and Ethical Basis of Trust in Medical AI. *J. Med. Ethics* 47 (5), 329–335. doi:10.1136/medethics-2021-107531

Ehrich, J., Gerber-Grote, A., Marg, W., Werner, A., and Levy, C. (2018). Internetbasierte Erfassung und Bearbeitung von Forschungsdaten zur Kindergesundheit. *Monatsschr Kinderheilkd* 166, 487–497. doi:10.1007/s00112-018-0453-y

European Commission (2012). *Communication from the Commission to the Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Health Action Plan 2012–2020 – Innovative Healthcare for the 21st Century (Brussels, 6.12.2012. COM/2012/0736 Final)*. Brussels: European Union.

European Convention on Human Rights (2021/1950). European Convention on Human Rights. Available at: https://www.echr.coe.int/Documents/Convention_ENG.pdf.

Genc, Z. (2014). Parents' Perceptions about the Mobile Technology Use of Preschool Aged Children. *Procedia - Soc. Behav. Sci.* 146, 55–60. doi:10.1016/j.sbspro.2014.08.086

German Ethics Council (2017). *German Version "Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung Stellungnahme."* Big Data and Health – Data Sovereignty as the Shaping of Informational Freedom. Opinion. Executive Summary & Recommendations. Available at: https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-big-data-und-gesundheit.pdf. [accessed: 03.02.2022].

Grote, T. (2021). Trustworthy Medical AI Systems Need to Know when They Don't Know. *J. Med. Ethics* 47 (5), 337–338. doi:10.1136/medethics-2021-107463

Heinrichs, B., Pinsdorf, C., and Staab, T. (2016). Ethische Aspekte des Off-Label-Use. *Jahrb. für Wiss. Ethik* 21, 47–67. doi:10.1515/jwiet-2017-0105

Hsia, B. C., Singh, A. K., Njeze, O., Cosar, E., Mowrey, W. B., Feldman, J., et al. (2020). Developing and Evaluating ASTHMAXcel Adventures: A Novel Gamified Mobile Application for Pediatric Patients with Asthma. *Ann. Allergy, Asthma & Immunol.* 125 (5), 581–588. doi:10.1016/j.anai.2020.07.018

Hummel, P., and Braun, M. (2020). Just Data? Solidarity and Justice in Data-Driven Medicine. *Life Sci. Soc. Policy* 16, 8. doi:10.1186/s40504-020-00101-7

Kantin, H. (2020). Giving Children a Say without Giving Them a Choice: Obtaining Affirmation of a Child's Non-dissent to Participation in Nonbeneficial Research. *Camb Q. Healthc. Ethics* 29 (1), 80–97. doi:10.1017/s0963180119000811

KBGG (2021). Kinderbetreuungsgeldgesetz. Fassung BGBl. I Nr. 221/2021: § 3. Available at: https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20001474 [accessed 02 03, 2022].

Kelly, C. J., Brown, A. P. Y., and Taylor, J. A. (2021). Artificial Intelligence in Paediatrics, in *Artificial Intelligence in Medicine*, ed. by N. Lidströmer and H. Ashrafian (Cham: Springer), Available at: . https://link.springer.com/referenceworkentry/10.1007/978-3-030-58080-3_316-1 [accessed: 11.02.2022].

KIGGS (20182018). KIGGS Wave 2 – First results from cross-sectional and cohort analyses. Studie zur Gesundheit von Kindern und Jugendlichen in Deutschland, Federal Health Reporting, joint service by RKI and Destasis. *J. Health Monit.* 3, 1. Available at: https://www.rki.de/EN/Content/Health_Monitoring/Health_Reporting/GBEDownloadsJ/Journal-of-Health-Monitoring_01_2018_KiGGS-Wave2_first_results.pdf?__blob=publicationFile [status (accessed: 15.07.2021).

Kinder-Richtlinie (2022). Aktuelle Richtlinie: https://www.g-ba.de/richtlinien/15/-- Document: Richtlinie des Gemeinsamen Bundesausschusses über die Früherkennung von Krankheiten bei Kindern (Kinder-Richtlinie). Version of 18. Juni 2015, published in Bundesanzeiger AT 18.08.2016 B1; last update 16. September 2021 published in Bundesanzeiger AT 03.11.2021 B4, effective since 1. Januar 2022. Available at: https://www.g-ba.de/downloads/62-492-2675/Kinder-RL_2021-09-16_iK-2022-01-01.pdf [accessed: 02.03.2022].

Knake, L. A. (2022). Artificial Intelligence in Pediatrics: The Future Is Now. *Pediatr. Res. Insights*, 1–2. doi:10.1038/s41390-022-01972-6

Kokol, P., Završnik, J., and Blažun Vošner, H. (2017). Artificial Intelligence and Pediatrics: A Synthetic Mini Review. *Pediatr. Dimens.* 2 (4), 1–5. doi:10.15761/PD.1000155

Krebs, P., and Duncan, D. T. (2015). Health App Use Among US Mobile Phone Owners: A National Survey. *JMIR mHealth uHealth* 3 (4), e101. doi:10.2196/mhealth.4924

Krüger-Brand, H. (2015). Big Data und Gesundheit. Viele Hoffnungen, viele Ängste. *Dtsch. Ärzteblatt* 11223, A1026–A1027.

Krutzinna, J. I. (2017). Beyond an Open Future. *Camb Q. Healthc. Ethics* 26 (2), 313–325. doi:10.1017/S096318011600089X

Krutzinna, J. (2022). Who Is "The Child"? Best Interests and Individuality of Children in Discretionary Decision-Making. *Int. J. Child. Rights* 30, 120–145. doi:10.1163/15718182-30010005

Li, Y. W., Liu, F., Zhang, T. N., Xu, F., Gao, Y. C., and Wu, T. (2022). Artificial Intelligence in Pediatrics. *Chin. Med. J. Engl.* 133 (3), 358–360. doi:10.1097/CM9.0000000000000563

Liebel, M. (2018). Welfare or Agency? Children's Interests as Foundation of Children's Rights. *Int. J. Child. Rights* 26 (4), 597–625. doi:10.1163/15718182-02604012

Löschke, J., and Heinrichs, B. (2015). Research Involving Minors – a Duty of Solidarity? Ethics in Biology, Engineering and Medicine. *Int. J.* 6, 67–80. doi:10.1615/ETHICSBIOLOGYENGMED.2015013491

Lupton, D. (2014). Apps as Artefacts: Towards a Critical Perspective on Mobile Health and Medical Apps. *Societies* 4 (4), 606–622. doi:10.3390/soc4040606

Mansmann, U. (2020). "Big Data in der Medizin: Konzeptionelle, organisatorische und technische Aspekte", in L. Summa, U. Mansmann, B. Buchner, and M. Schnebbe: *Big Data in der Medizin. Ethik in den Biowissenschaften*. Sachstandsberichte of the DRZE, vol. 22, 13-48. Editors: D. Sturma and D. Lanzerath (Freiburg im Breisgau: Verlag Karl Alber).

Martínez García, E., Catalán Escudero, P., Mateos Arroyo, J., Ramos Luengo, A., Sánchez Alonso, F., and Reinoso Barbero, F. (2020). PainAPPle. Validation and Evaluation of an Electronic Application for the Management of Acute Pain in Pediatric Patients. *Rev. Española Anestesiol. Reanim. (English Ed.* 67 (3), 139–146. doi:10.1016/j.redare.2019.09.007

Mayer-Schönberger, V., and Ingelsson, E. (2017). Big Data and Medicine: A Big Deal? *J. Intern Med.* 283 (5), 418–429. doi:10.1111/joim.12721

McDougall, R. J. (2019). Computer Knows Best? the Need for Value-Flexibility in Medical AI. *J. Med. Ethics* 45 (3), 156–160. doi:10.1136/medethics-2018-105118

Mittelstadt, B. D., and Floridi, L. (Editors) (2016). "The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts," *The Ethics of Biomedical Big Data* (Cham: Springer International Publishing), 445–480.

Mittelstadt, B. (2017). From Individual to Group Privacy in Big Data Analytics. *Philos. Technol.* 30, 475–494. doi:10.1007/s13347-017-0253-7

Montgomery, K. C., Chester, J., and Milosevic, T. (2017). Children's Privacy in the Big Data Era: Research Opportunities. *PEDIATRICS* 140 (2), S117–S121. doi:10.1542/peds.2016-1758O

Morris, A. H., Stagg, B., Lanspa, M., Orme, J., Clemmer, T. P., Weaver, L. K., et al. (2021). Enabling a Learning Healthcare System with Automated Computer Protocols that Produce Replicable and Personalized Clinician Actions. *J. Am. Med. Inf. Assoc.* 28 (6), 1330–1344. doi:10.1093/jamia/ocaa294

MuKiPass (2002). Verordnung des Bundesministers für soziale Sicherheit und Generationen über die Festlegung eines Mutter-Kind-Pass-Untersuchungsprogrammes, die Voraussetzungen zur Weitergewährung des Kinderbetreuungsgeldes in voller Höhe sowie über den Mutter-Kind-Pass (Mutter-Kind-Pass-Verordnung 2002 - MuKiPassV). Österreichische Gesetzgebung. Fassung: BGBl. II Nr. 420/2013,BGBl. II Nr. 470/2001. Available at: https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20001694 [accessed: 02.03.2022].

Müller, H., and Samerski, S. (2016). Big Data: Eine Datenethik Ist Unabdingbar. *Komment. Dtsch. Ärzteblatt* 11340, A1749.

Neff, G., Tanweer, A., Fiore-Gartland, B., and Osburn, L. (2017). Critique and Contribute: A Practice-Based Framework for Improving Critical Data Studies and Data Science. *Big Data* 5 (2), 85–97. doi:10.1089/big.2016.0050

NL Ministry of Health (2022). *Preventie in de Wet publieke gezondheid*. Rijksinstituut voor Volksgezondheid en Milieu Ministerie van Volksgezondheid, Welzijn en Sport) Available at: https://www.loketgezondleven.nl/zorgstelsel/preventie-vijf-stelsel-wetten/preventie-wpg (accessed 02 04, 2022).

O'Neill, O. (Editor) (2016). "Public Health or Clinical Ethics: Thinking beyond Borders," *Justice across Boundaries* (Cambridge: Cambridge University Press), 211–224.

Paycha, F. (1968). Diagnosis with the Aid of Artificial Intelligence: Demonstration of the 1st Diagnostic Machine. *Presse Therm. Clim.* 1051, 22–25.

Pinel, C., Prainsack, B., and McKevitt, C. (2020). Caring for Data: Value Creation in a Data-Intensive Research Laboratory. *Soc. Stud. Sci.* 50 (2), 175–197. doi:10.1177/0306312720906567

Prasser, F., Kohlbacher, O., Mansmann, U., Bauer, B., and Kuhn, K. A. (2018). Data Integration for Future Medicine (DIFUTURE). *Methods Inf. Med.* 57:S 01, e57-e65.doi:10.3414/ME17-02-0022

Radenbach, K. E. (2006). Gruppennützige Forschung an Kindern und Jugendlichen. Ihre ethische und rechtliche Zulässigkeit unter besonderer Berücksichtigung der Bewertung von Vorsitzenden deutscher Ethikkommissionen. Göttingen. Available at: https://ediss.uni-goettingen.de/bitstream/handle/11858/00-1735-0000-0006-B33D-A/radenbach.pdf?sequence=1.

Radovic, A., Vona, P. L., Santostefano, A. M., Ciaravino, S., Miller, E., and Stein, B. D. (2016). Smartphone Applications for Mental Health. *Cyberpsychology, Behav. Soc. Netw.* 197, 465–470. doi:10.1089/cyber.2015.0619

Richardson, B., Dol, J., Rutledge, K., Monaghan, J., Orovec, A., Howie, K., et al. (2019). Evaluation of Mobile Apps Targeted to Parents of Infants in the Neonatal Intensive Care Unit: Systematic App Review. *JMIR Mhealth Uhealth* 7 (4), e11620. doi:10.2196/11620

Richter, G., and Buyx, A. (2016). Breite Einwilligung (broad consent) zur Biobank-Forschung - die ethische Debatte. *Ethik Med.* 28 (4), 311–325. doi:10.1007/s00481-016-0398-4

Rüping, S. (2015). Big Data in Medizin und Gesundheitswesen. *Bundesgesundheitsbl.* 58, 794–798. doi:10.1007/s00103-015-2181-y

Sadowski, J. (2019). When Data Is Capital: Datafication, Accumulation, and Extraction. *Big Data & Soc.*, 1–12. doi:10.1177/2053951718820549

Sadowski, J. (2020). *Too Smart. How Digital Capitalism Is Extracting Data, Controlling Our Lives, and Taking over the World*. Cambridge; London: The MIT Press.

Seidman, L. C., Martin, S. R., Trant, M. W., Payne, L. A., Zeltzer, L. K., Cousineau, T. M., et al. (2019). Feasibility and Acceptance Testing of a Mobile Application Providing Psychosocial Support for Parents of Children and Adolescents with Chronic Pain: Results of a Nonrandomized Trial. *J. Pediatr. Psychol.* 44 (6), 645–655. doi:10.1093/jpep/jsz007

Sharon, T. (2018). When Digital Health Meets Digital Capitalism, How Common Goods Are at Stake? *Big Data & Soc.* 5 (2), 1–12. doi:10.1177/2053951718819032

Stewart, C. (2021). mHealth – Statistics & Facts. Available at: https://www.statista.com/topics/2263/mhealth/#dossierKeyfigures [accessed 16.03.2022].

Sugiyama, K., and Hasegawa, Y. (1984). Computer Assisted Medical Aiagnosis System for Inborn Errors of Metabolism. *Jpn. J. Med. Electron. Biol. Eng.* 22, 942–943.

Summa, L. (2020). "Big Data in der Medizin: Ethische Aspekte", in: L. Summa, U. Mansmann, B. Buchner, and M. Schnebbe: *Big Data in der Medizin. Ethik in den Biowissenschaften*. Sachstandsberichte of the DRZE, vol. 22, 74-121. Editors: D. Sturma and D. Lanzerath (Freiburg im Breisgau: Verlag Karl Alber).

Summa, L., Mansmann, U., Buchner, B., and Schnebbe, M. (2020). Big Data in der Medizin. Ethik in den Biowissenschaften. Sachstandsberichte of the DRZE, vol. 22, ed. by D. Sturma and D. Lanzerath (Freiburg im Breisgau: Verlag Karl Alber).

Taylor, H. J. (2016). What Are 'Best Interests'? a Critical Evaluation of 'Best Interests' Decision-Making in Clinical Practice. *Med. Law Rev.* 24 (2), 176–205. doi:10.1093/medlaw/fww007

Topol, E. (2019). *Deep Medicine. How Artificial Intelligence Can Make Healthcare Human Again*. New York: Basic Books.

Tragomalou, A., Moschonis, G., Manios, Y., Kassari, P., Ioakimidis, I., Diou, C., et al. (2020). Novel E-Health Applications for the Management of Cardiometabolic Risk Factors in Children and Adolescents in Greece. *Nutrients* 125, 1380. doi:10.3390/nu12051380

Tsakiliotes, K. (2021). Challenges of mHealth. Available at: https://www.internetjustsociety.org/challenges-of-mhealth [accessed: 12.05.2022].

UN Convention on the Rights of the Child (1990). UN Convention on the Rights of the Child. Available at: https://www.ohchr.org/en/professionalinterest/pages/crc.aspx.

Universal Declaration of Human Rights (1948). Universal Declaration of Human Rights. Available at: https://www.ohchr.org/Documents/Publications/ABCannexesen.pdf.

UTeilnahmeDatVO (2008). SGV. NRW.: Verordnung zur Datenmeldung der Teilnahme an Kinderfrüherkennungsuntersuchungen/U-Untersuchungen (U-Untersuchung-TeilnahmedatenVO – UTeilnahmeDatVO). Available at: https://recht.nrw.de/lmi/owa/br_text_anzeigen?v_id=10000000000000000719 [accessed: 02.03.2022].

Virani, A., Duffett-Leger, L., and Letourneau, N. (2020). Parents' Perspectives of Parenting App Use. *J. Inf. Nurs.* 5 (1), 8–18.

WHO (World Health Organization) (2011). *mHealth: New Horizons for Health through Mobile Technologies: Second Global Survey on eHealth. Global Observatory for eHealth Series*, 3. Geneva, Switzerland: WHO Press. Available at: https://www.who.int/goe/publications/goe_mhealth_web.pdf [status [accessed: 28.07.2021].

Winkler, H. A. (2017). 2. Die Nationen Überwinden Oder Überwölben? *Frankf. Forum Diskurse* 16 (Vortrag 3), 22–31. doi:10.17104/9783406711749-22

Xafis, V., Schaefer, G. O., Labude, M. K., Brassington, I., Ballantyne, A., Lim, H. Y., et al. (2019). An Ethics Framework for Big Data in Health and Research. *Abr* 11 (3), 227–254. doi:10.1007/s41649-019-00099-x

Yang, Y. C., Islam, S. U., Noor, A., Khan, S., Afsar, W., and Nazir, S. (2021). Influential Usage of Big Data and Artificial Intelligence in Healthcare. *Comput. Math. Methods Med.*, 5812499. doi:10.1155/2021/5812499

Zuboff, S. (2019). *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontiers of Power*. New York: Public Affairs.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Stake and Heinrichs. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 15 August 2022
DOI 10.3389/fgene.2022.902542

Check for updates

OPEN ACCESS

EDITED BY
Tamar Sharon,
Radboud University Nijmegen,
Netherlands

REVIEWED BY
Stephen Hilgartner,
Cornell University, United States
Vasiliki Rahimzadeh,
Stanford University, United States

*CORRESPONDENCE
Giovanni Rubeis,
giovanni.rubeis@kl.ac.at

SPECIALTY SECTION
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Genetics

RECEIVED 23 March 2022
ACCEPTED 19 July 2022
PUBLISHED 15 August 2022

CITATION
Rubeis G, Dubbala K and Metzler I (2022), "Democratizing" artificial intelligence in medicine and healthcare: Mapping the uses of an elusive term. *Front. Genet.* 13:902542. doi: 10.3389/fgene.2022.902542

COPYRIGHT
© 2022 Rubeis, Dubbala and Metzler. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

"Democratizing" artificial intelligence in medicine and healthcare: Mapping the uses of an elusive term

Giovanni Rubeis*, Keerthi Dubbala and Ingrid Metzler

Karl Landsteiner University of Health Sciences, Krems an der Donau, Austria

Introduction: "Democratizing" artificial intelligence (AI) in medicine and healthcare is a vague term that encompasses various meanings, issues, and visions. This article maps the ways this term is used in discourses on AI in medicine and healthcare and uses this map for a normative reflection on how to direct AI in medicine and healthcare towards desirable futures.

Methods: We searched peer-reviewed articles from Scopus, Google Scholar, and PubMed along with grey literature using search terms "democrat*", "artificial intelligence" and "machine learning". We approached both as documents and analyzed them qualitatively, asking: What is the object of democratization? What should be democratized, and why? Who is the demos who is said to benefit from democratization? And what kind of theories of democracy are (tacitly) tied to specific uses of the term?

Results: We identified four clusters of visions of democratizing AI in healthcare and medicine: 1) democratizing medicine and healthcare through AI, 2) multiplying the producers and users of AI, 3) enabling access to and oversight of data, and 4) making AI an object of democratic governance.

Discussion: The envisioned democratization in most visions mainly focuses on patients as consumers and relies on or limits itself to free market-solutions. Democratization in this context requires defining and envisioning a set of social goods, and deliberative processes and modes of participation to ensure that those affected by AI in healthcare have a say on its development and use.

KEYWORDS

artificial intelligence, big data, ethics, digital technologies, democratization

1 Introduction

In his seminal work What Tech calls Thinking: An Inquiry into the Intellectual Bedrock of Silicon Valley, Adrian Daub (2020) describes the way Big Tech uses and reframes concepts like "disruption" and "communication" to shape our understanding of the goals and purposes of the industry. Daub argues that by reframing these concepts, narratives are implanted into the collective consciousness that explain and legitimize the way digital companies aim to change the world. In his view, digital technologies, especially

Big Data applications and machine learning software, often referred to as Artificial Intelligence (AI), are not mere tools for improving communication and data exchange, optimizing work processes, or enabling commodification of social goods. Rather, these tools are framed as enablers of a new, and of course better, way of life. Big Tech is more than just selling products or services; it is about making the world a better place (Daub, 2020).

We are witnessing a similar tendency in discourses on digital technologies and AI in medicine and healthcare. Digital technologies are developed and used for a wide variety of diagnostic and therapeutic practices such as health data management, image recognition, decision support systems and assistive technologies (Briganti and Le Moine, 2020; Mishra, 2022). Value-laden terms like "disruption" (Rubeis, 2020) and "revolution" (Topol, 2012) are prominent in the discourse on these technologies. Recently, the term "democratization" has been added to this list. According to Eric Topol, one of the most prominent voices in the discourse on AI in medicine and healthcare, these technologies will transform medical practices and structures of healthcare systems and thus democratize medicine (Topol, 2012; Topol, 2015; Steinhubl and Topol, 2018; Topol, 2019). In this view, AI is more than just a new tool for improving isolated medical practices. Following Topol, the ultimate goal is "deep empathy" (Topol, 2019). Deep empathy refers to the optimization of data use and work processes, which will free physicians from time-consuming and mechanical tasks, thus leaving them more room to focus on their relationship with patients (Topol, 2019). Topol describes deep empathy as the culmination of a process that combines digitalization and democratization. Another crucial aspect in this discussion is the ownership of, control over, and access to personal health data by patients. Topol links the "suppressive force of doctors to retain control of patient data" to paternalism (Topol, 2019, p.270) and claims that "medical paternalism would fade as consumers didn't simply generate their information but owned it" (Topol, 2019, p.24). In this view, the patient-as-consumer and data-owner is empowered and can face healthcare professionals on an equal level. This "deep medicine", as termed by Topol, is enabled by AI-technologies and the use of big data. When the optimization of workflow of healthcare professionals and the empowerment of patients converge, we will get rid of paternalism for good and thus democratize medicine.

However, this particular use of the term democratization is not universally shared within the ongoing debate on AI in medicine and healthcare. "Democratizing AI" is a vague term that encompasses various meanings, issues, and visions. Its use extends in nuances within two poles, each reflecting competing understandings of the power of biomedical technologies and their agency in innovation processes (Timmermans and Berg, 2003; Metzler and Åm, 2022). One pole consists of the framing of AI as a transformative agent that can democratize medicine and healthcare. Medicine and healthcare are the objects that ought to be made more democratic, and data-intensive technologies and AI are the means to achieve this goal. The other pole consists of uses in which AI is the object that ought to be democratized. This vision is articulated in various nuances. Some actors underline a need to democratize access to technical tools that help develop AI. The tools include open access to code libraries, developer tools, and data sets (Garvey, 2018; Bhattacharya et al., 2021),collaborative learning and crowdsourcing (Bond et al., 2019a; Bond et al., 2019b; Lyu et al., 2020), accessible interfaces (Vanhorn and Çobanoğlu, 2021) and end-user machine learning systems (Traub et al., 2019). Access to these tools allows biomedical experts without software development skills to contribute to wider use of AI in healthcare. There are also calls to "democratize" AI algorithms by preventing sampling bias and tackling the under-representation of groups in training data (Mulvenna et al., 2021; Wong, 2019). Last but not the least, some actors also call for making the development and use of AI-based technologies an object of democratic governance (Nemitz, 2018; Himmelreich, 2022).

In this article, we map the different ways "democratizing AI" and the "democratization" of AI has been used in discourse on AI in medicine and healthcare and use the mapping for a normative discussion of the term. We begin by describing our methods. We then present and discuss our results. In the discussion section, we contextualize the different uses of the term democratizing AI with current approaches in medical and AI ethics. Since we address the topic of democratization in the context of medical AI from the perspective of normative ethics, our discussion will be a normative one. In the concluding section, we summarize the outcomes of our analysis.

2 Methods

This article is based on an empirical engagement with uses of the terms "democratization of AI" and "democratizing AI" in the discourse on medicine and healthcare.

2.1 Materials

In terms of materials, we used peer reviewed articles and grey literature. We searched for peer-reviewed articles on artificial intelligence within and outside healthcare using search terms "democrat*" and "artificial intelligence" or "machine learning". We searched Scopus and PubMed databases through Ebsco search engine along with Nature, Science and Lancet journal databases and Google Scholar. We limited the search to English language but did not specify date range. The search resulted in 2071 articles. After deduplicating, we selected the articles that discussed AI and democratization in detail, as opposed to those that just mentioned the terms. We included all articles that refer to AI-based technologies without defining the term ourselves. This ensured an openness towards different interpretations of AI within the medical context. We complemented this material with

documents from professional societies and international organizations that used some variation of the term "democratizing AI". Finally, we included 35articles for analysis.

2.2 Conceptual lenses

We approached the literature as documents, in which the authors articulated understandings of the meaning and significance of "democratizing AI" in medicine and healthcare, often drawing on tacit understandings of the agency and power of AI and implicit theories on democracy. The documents we analyzed were of recent date (i.e., almost all of them were published after 2016 and several of them were published within the last 2 years) and diverse. They ranged from Editorials, over research articles, to guiding documents of professional societies. A cross-cutting theme was the literature's future-oriented nature. Many articles discussed emerging trends, expected future developments, or called for actions to be taken. In light of the future-oriented nature of the discourse, we approached uses of the term "democratization" and "democratizing" AI as articulations of (sociotechnical) "visions" (Hilgartner, 2015; Jasanoff, 2015), i.e., understandings of the nature of desirable futures attainable through AI, and of the ways in which these futures can, or ought to, be achieved.

We analyzed the documents with "agnostic" lenses (Laurent, 2017). We did not select one definition of democratization as a normative baseline to critically assess our material, but strived to induce various definitions of "democratization" from the authors' writings. This approach was informed by the conceptual understanding that, paraphrasing Tamar Sharon's work on "common goods", "a plurality of conceptualizations" (Sharon, 2018) of democratization and democracy are at work in the discourse on AI. Indeed, democracy can be understood as an "essentially contested concept" (Gallie, 1955). Most people agree on the value and importance of democracy, while they disagree on what democracy is, or ought to be. They agree that democracies are a desirable good, but they disagree on which actions ought to be taken to achieve this good in practice. In broadest terms, democracy refers to the "rule" (as of -cracy) of the "people" (demos) and denotes expectations on equality. However, there are disagreements on the range of objects that ought to be subjected to the rule of the people, on the desirable practices and institutions to organize that rule, and on the boundaries of the demos or people. Thus, definitions of democracies are also visions of what they ought to be. Similarly, "democratization" is a morally charged term. It problematizes a phenomenon as insufficiently democratic, while simultaneously giving moral power and legitimacy to the agents and means of democratization.

2.3 Data analysis

We analyzed the documents qualitatively to develop a better understanding of the uses of "democratization" in medicine and healthcare. We mapped four clusters of visions of democratizing AI from the materials, using the following questions to code the documents and distill and categorize clusters of visions:

- What is the object of democratization? What is the object that should be democratized, and why?
- Who is the "demos" (Doubleday and Wynne, 2011) who is said to benefit from, or that ought to be involved in, "democratizing AI?"
- What kind of theory of democracy is (often tacitly) tied to specific visions?

We phrased these questions after a first reading of the documents and engagements with scholarly literature on interactions between biomedical technologies and social orders in democratic societies (Timmermans and Berg, 2003; Marres, 2007; Doubleday and Wynne, 2011; Jasanoff, 2013). We then commenced with an intense analysis of a small random sample of documents (within the selected documents), seeking to maximize variations within this sample (Silverman, 2015). We coded them along the three questions, and distilled clusters of visions from this initial analysis.

When clustering visions of democratizing AI, we explored whether a specific vision was sufficiently different to be categorized as a distinct one or whether it could to be subsumed under a vision already deduced, following strategies of qualitative content analysis (Schreier, 2012). We analyzed the documents separately and discussed clusters of visions. Once we had agreed on the set of clusters, we used the remaining documents to validate the clusters.

3 Results

We identified four clusters of visions of democratizing AI within the analyzed material. In the following, we will outline these clusters.

3.1 Artificial intelligence for the people: Democratizing medicine and healthcare through Artificial intelligence

The first vision of "democratization" we identified is democratizing medicine and healthcare through AI. Following this view, democratization is based on two factors: data and the technologies to obtain and process them. Data include individual health data that may range from test results deposited in electronic health records, to behavioral data, or social media entries (Steinhubl and Topol, 2018; Topol, 2019; Weissglass, 2021). Data technologies encompass software like machine learning algorithms, data mining tools, and cloud computing, but also hardware like mobile devices (Mulvenna et al., 2021;

Topol, 2019; Burnside et al., 2020; Weissglass, 2021). Mobile devices like smart phones or tablets allow users to generate and collect data outside of clinical settings or medical expertise. They are seen as the crucial device of democratizing healthcare, enabling clinicians to obtain real world data, e.g., through a digitally enhanced experience sampling method or ecological momentary assessment (Mulvenna et al., 2021). These methods are especially relevant for collecting behavioral and lifestyle data. Thus, data can be used to personalize treatment. Also, the fact that patients generate and collect this data themselves is seen as a democratizing effect (Steinhubl and Topol, 2018). Using mobile devices for data collection may also reduce access barriers to healthcare services (Weissglass, 2021). Digital technologies can help improve health surveillance by generating more and potentially better data, especially in settings with low healthcare coverage. In low-and middle-income countries (LMICs), this could contribute to better access to healthcare services and hence more democratic healthcare systems (Weissglass, 2021).

The connection between AI-technologies and health equity in terms of access is considered crucial for healthcare. One example discussed in the body of literature is a machine learning-based point-of-care screening tool for genetic syndromes in children (Porras et al., 2021). This deep phenotyping technology tool uses deep neural networks and facial statistical shape models to assess the risk of a child having one of the genetic syndromes covered by the technology. The tool, the authors argue, can identify the need of patients for referral to a specialist. It may thus assist physicians in their diagnostic practices, especially in areas where access to specialized care and genetic resources is scarce. Although the tool is no substitute to genetic diagnostics, it is referred to as a contribution to democratizing access to the healthcare resources needed (Porras et al., 2021).

Thus, in this first vision of democratizing AI, AI is imagined as a transformative agent that promises to democratize medicine and healthcare. The demos of this vision consists of individual citizens, often referred to as patients or consumers and mostly located in high-income countries (HICs), or a patient population mostly located in LMICs, who could benefit from the transformative power of AI-based technologies in redefining healthcare or providing healthcare through new means. The democratization of medicine and healthcare through AI is thus often also linked to other values, such as empowerment, participation, equity and access to healthcare.

3.2 Artificial intelligence by the people: Democratizing artificial intelligence in medicine and healthcare by multiplying developers, evaluators, and users

The second vision of democratization of AI refers to facilitated access to AI-technologies in terms of design and/or use. Democratizing AI in this respect means making machine learning accessible to non-domain specialists (Dibia et al., 2018; Traub et al., 2019; Gupta, 2020; Kobayashi et al., 2019; Mulvenna et al., 2021; Vanhorn and Çobanoğlu, 2021). The aim is to enable those without technical expertise on AI, such as healthcare professionals as well as biomedical researchers, to handle AI-technologies. Some authors describe this democratization of AI as an already ongoing process, which will contribute to a widespread use of AI in biomedical research and healthcare practices. According to this view, high-performance computer hardware, cloud machine learning tools, accessible software, and affordable online education have already democratized the creation and use of AI (Dibia et al., 2018; Bond et al., 2019b; Mulvenna et al., 2021; Saldívar-González et al., 2022). "Democratizing", in this understanding, is mostly discussed regarding better access to knowledge and tools. Some authors note that health professionals lack the required knowledge and skills for handling AI, which also negatively influences their attitudes towards the technologies (Allen et al., 2019). A basic knowledge of how algorithms work, what their limits are, and how to evaluate them for clinical practice is thus needed.

An important aspect in the context of facilitating better access to AI technologies is the often-lacking infrastructure in hospitals and other healthcare facilities for engaging with AI development (Allen et al., 2019). One approach to overcome this barrier is to provide toolkits or other ready-made solutions for developing and applying AI (Dibia et al., 2018; Sikpa et al., 2019; Vanhorn and Çobanoğlu, 2021). Vanhorn and Çobanoğlu (2021) suggest a virtual reality (VR)-platform as a simplified environment in which users can design, train, and evaluate models. Instead of coding, users handle data sets, in this case images, in an immersive environment where they can grab data sets and shift or sort them. This immersive experience is meant to enable a more intuitive model development without any coding skills. Another approach is the provision of platforms for code-free automated machine learning (AutoML) interfaces, which is explicitly framed as an empowerment of healthcare professionals and biomedical researchers (Nature Machine, 2021).

Thus, the second cluster of visions of democratizing AI shares the technological optimism with the first vision yet problematizes the identity of the visioners of AI. In this vision AI shifts from a transformative agent that renders medicine and healthcare more democratic to an object *in* medicine and healthcare that needs to be rendered more democratic—or indeed, to be democratized. Here, democratizing AI refers to making tools to render AI accessible to biomedical professionals. In this vision, the demos refers to biomedical professionals, who ought to be involved in the development of AI or be able to use models developed with the help of AI. In turn, democratizing AI in this way helps to empower biomedical professionals, augment their expertise, while also disseminating the use of AI in biomedicine. It is important to note that especially Big Tech companies like Amazon, Google, and Microsoft are key players in this vision of "democratizing" AI (Nature Machine, 2021).

3.3 People in artificial intelligence: Democratizing access to and oversight of data

The third cluster of visions shares the second's understanding that the development of AI-based systems needs to be democratized, focusing on access to and oversight of data used to develop AI-based technologies. It addresses the issue that algorithms are often developed, trained, or validated on data from a single institution (Allen et al., 2019; Traub et al., 2019; Gupta, 2020), which is one of several sources of data bias. We identified two related visions on how access to, and oversight of, data can be democratized: strategies for distributive learning and the establishment of 'data commons' or "data trustees".

One vision suggests various technical approaches to "democratize" the data used to develop AI, by making data more representative. Some authors suggest approaches for distributed training of AI models with decentralized data like federated, distributed, or split learning for facilitating data sharing (Allen et al., 2019; Lyu et al., 2020). Another strategy is transfer learning (TL), where instead of training new models at each hospital, pretrained models can be used and adapted, which reduces the number of data sets needed (Gupta, 2020). Other approaches focus on enabling sharing between institutions. For instance, Traub et al. developed a data ecosystem that serves as infrastructure for sharing assets such as data, algorithms, ML models, systems, services, and compute resources (Traub et al., 2019). Used as a marketplace, this ecosystem could facilitate easier access to these assets.

Another vision is centered around strategies for making data more accessible. Democratizing AI in this respect means to facilitate open access to data sets (Bhattacharya et al., 2021). Bhattacharya et al. identify three crucial factors of democratizing health data: Discoverability in data repositories through the provision of meta data, accessibility of data using websites, tools, and interfaces, and interoperability through standardization of data sets (Bhattacharya et al., 2021).

Thus, the third vision of democratizing AI builds on the second vision's understanding that AI in medicine and healthcare needs to be democratized to unleash its transformative potential and problematizes the "means of production" of AI. It does not focus on the hard- and software, or on the skills needed to use them in practice (as the second vision), but addresses the nature of the data needed to develop, train, and evaluate AI. This vision underlines that there is a shortage of high-quality data in biomedicine and healthcare, that can represent the variety of patients, groups, and people that ought to benefit from AI-based technologies in medicine and healthcare. While this vision overlaps with the second vision as being primarily about enabling access to the means of production of AI, here the demos does not only relate to the users of data (i.e., biomedical professionals) but also to the contributors of data or the people whose traces are in the data.

3.4 Making artificial intelligence an object of the rule by the people

A fourth visions of democratizing AI calls to transform AI into an object of the rule by the people, suggesting that "AI should be subject to novel or different forms of democratic governance" (Himmelreich, 2022, p.3). Just as AI has the potential to contribute to human wellbeing, it can also be used, or abused, to achieve undesirable ends. Thus, there is a sense that the development and adoption of AI-based systems cannot be left in the hands of developers, technicians, or big tech alone (Garvey, 2019). To ensure that AI will "serve the public good" (Nemitz, 2018, p.7), AI and the professionals that develop and use it need direction, oversight, and democratic governance by the people, or by authorities acting on behalf of them.

We identified two versions within this cluster of visions of democratic governance for medical AI, which draw upon different approaches towards democratic governance. One vision builds upon traditions of direct democratic governance and calls for the involvement or engagement of people affected by AI in the development and oversight of AI—or public participation for short. This vision builds upon practices of public participation and public engagement, which have become salient in the governance of emerging technologies over the past decades. While they have taken shape in very different ways, participatory practices are sustained by the understanding that people directly or indirectly affected by emerging technologies (as consumers, patients, or citizens), should have a say on the development, use, or oversight of said technologies (Hagendijk and Irwin, 2006; Felt and Fochler, 2008; Gottweis et al., 2008; Doubleday and Wynne, 2011; Braun and Könninger, 2018). In the case of AI, public participation is also referred to as "co-creation" or "co-design (Donia and Shaw, 2021). Involving the public in the development and oversight of AI-based systems is expected to render the latter socially robust and acceptable. Calls to include publics in the design of AI-based systems are tied to the normative understanding that people affected by AI should be involved in their development, or—drawing on the term Harambam and colleagues used for algorithmic news recommenders—people should have "voice" (Harambam et al., 2018) if they are affected by it. Moreover, calls for including publics in the design of AI are also tied to the normative expectation that people's practical knowledge can render AI more intelligent, such as by helping to identify needs (Barclay, 2020), or by learning from publics how they define ethical values (Wong, 2020).

We can also find such calls to engage publics in documents of international and professional societies. For instance, a guidance from the World Health Organization (WHO) on the "Ethics and governance of artificial intelligence for health" suggests that AI technologies should both be "designed by and evaluated with the active participation of those who are required to use the system or will be affected by it, including providers and patients" (World

Health Organisation, 2021, p.29) to ensure "inclusiveness and equity". Similarly, the United Kingdom Academy of Medical Sciences deemed "ongoing engagement with patients, the public and healthcare professionals (. . .) critical to ensuring new AI technologies respond to clinical unmet need, are fit for purpose, and are successfully deployed, adopted, and used." (Academy of Medical Sciences, 2019).

Another version of the vision of democratizing AI by rendering it an object of democratic governance builds upon theories of representative democracy, by entrusting authorities with overseeing the development and use of AI on behalf of the people. While this vision shares a common ground with calls for public participation, calls for democratic governance through enforceable regulations are also responsive to the proliferation of ethical principles for AI over the past few years, which have been developed with the involvement of Big Tech (Mittelstadt, 2019). While these principles are welcomed, voluntary compliance is often deemed insufficient to ensure that AI will help to enable desirable futures. For instance, Paul Nemitz (Nemitz, 2018, p.1) argues that the "key question for AI in democracy" is what can be left to "ethics", or the voluntary self-governance of the industry along a set of ethical principles and practices, "and which challenges of AI need to be addressed by rules which are enforceable and encompass the legitimacy of democratic process, thus laws."

This vision on democratic AI governance allocates responsibilities to state actors and regulatory authorities to make AI a "subject to the rules set by democracy in law" (Nemitz, 2018, p.10). Over the past few years, several state authorities and international and supranational organizations have begun to address how to govern AI through law and law-like measures. For instance, the European Commission (EC) has begun to elaborate on an approach that "places people at the centre of the development of AI (human centric AI) and encourages the use of this powerful technology to help solve the world's biggest challenges" (Ulnicane, 2022, p.261). In a White paper on AI published in February 2020, EC outlined a risk-based common European regulatory framework to ensure "that new technologies are at the service of all Europeans—improving their lives while respecting their rights." (European Commission, 2020, p.1). A central aim of the EC is to oversee AI in such a way as to allow "people [to] to be able to trust it." (European Commission, 2020, p.1) page 1.

Thus, a fourth narrative on democratizing AI builds upon the first three meanings, while also extending the scope of democratizing AI. While it capitalizes on the narrative of AI as a powerful transformative technology, it draws on practices and theories of participatory and representative democracy to underline that AI needs to be made an object of the rule by the people to ensure that it will be developed and used for the people and accepted and trusted by them. In this vision, the demos refers to the people directly or indirectly affected by AI and thus to the public, who should either be engaged in the development of AI or be represented by authorities who govern AI on behalf of them.

4 Discussion

We have mapped different ways the term "democratizing AI" is used in the discourse on AI in medicine and healthcare, describing four clusters of visions that we have distilled from the material. In the following, we discuss our findings along the object of democratization, the demos, and the type of democracy tied to each of the four cluster of visions of democratization. We contextualize the results with research on medical ethics and ethics of digital health technologies.

4.1 The object of democratization: The problem to be solved

The object of democratization, i.e., the practices or structures to be democratized, varies between the four visions we identified. The first vision discusses the democratization of healthcare through AI. The other three visions focus on the democratization of AI in medicine and healthcare, articulating different visions of which dimensions of AI ought to be democratized and of who ought to be engaged. The first vision defines medicine and healthcare as the object of democratization. According to this vision, paternalistic and dysfunctional healthcare systems should be transformed into consumer-oriented non-hierarchical health markets. A more personalized and autonomy-empowering healthcare system, especially in HICs, and better access to healthcare, especially in LMICs, could be the main benefits of the democratization of medicine and healthcare through AI. The second and third cluster of visions identify AI-based technologies as the object of democratization. Enabling better access to health data, models, and algorithms is the main objective in this context. The free exchange of data, knowledge, and technologies is key to improving the quality of AI-based technologies and enabling healthcare professionals to contribute to development and usage of these technologies. In the fourth vision on democratization, the practices of people who envision, develop, and use AI are deemed to be in need of being made an object of democratic governance.

A tendency towards technical solutionism cuts across all four visions, i.e., the notion that genuinely social or political problems can be fixed through technological innovations or the applications of technology (Morozov, 2013; Howard, 2021), even if this solutionism is stronger in the first vision and weaker in the fourth vision. An example is Topol's concept of deep medicine, where AI offers "a technological solution to the profound human disconnection that exists today in healthcare" (Topol, 2019, p.272). In some visions, the solutionist approach is rather vague, simply framing democratization of healthcare as a good that medical AI can deliver (Weissglass, 2021). In others, explicit problems and their AI-based solutions are identified. For example, AI may fix a crucial problem of the supposedly

dysfunctional healthcare system, missing care, because technology "doesn't complain, can work, doesn't get tired" (Topol, 2019).

4.2 The demos

We identified the demos tied to each of the four visions of democratization we derived from our material. In the first vision, democratization of healthcare through AI, the demos relates to individuals and groups empowered by AI. In the second vision of democratizing access to tools to develop AI, the demos consists primarily of biomedical professionals (including researchers and clinicians) who lack technical expertise on AI. In the third vision of democratizing data, the demos is more ambiguous, referring to biomedical professionals who ought to have access to data and to groups of patients who ought to be represented in data. The demos of the fourth vision is closest to the demos as we (and political theory) know it. It refers to the people who would be directly and indirectly affected by AI or to the "public" (Dewey, 1927). They are deemed to be entitled to have a say on how, by whom, and for which purposes AI-based technologies are developed and used—either by having a direct say on them in participatory settings, or by being represented by authorities who govern AI on behalf of publics. Thus, the way in which the demos is tacitly defined and normatively framed in the four visions differ. The difference is particularly visible when the demos of the first vision is compared with the demos of the fourth vision.

The prevalent framing of individuals in the first vision of democratization is the patient-as-consumer, that is, typical for digital health (Lupton, 2013). Two enabling factors of empowerment are especially relevant in this regard: data-ownership and mobile health (mHealth) technologies. For Topol (Topol, 2019), the empowered patient is a consumer, defined by ownership of their own health data. He gives a list of reasons, of which the first two are "It is your body. You paid for it." (Topol, 2019, p.264). Ownership of their own health data empowers patients to act as consumers in the medical encounter. The underlying assumption is that as data-owners and consumers, patients are in a stronger position vis-à-vis medical professionals. Democratization is framed as the antagonist to medical paternalism. The empowered patient is a consumer of health services and an owner of data that emancipates themselves from a paternalistic system.

However, the basic narrative of empowerment through engagement with one's own health data and self-management practices is highly problematic. The idea that AI-based technologies and especially mHealth solutions enable empowerment of patients and lead to more autonomy has been widely criticized (Lupton, 2013; Sharon, 2016; Rubeis et al., 2018; Morley and Floridi, 2020; Rubeis, 2020). Morley and Floridi state that there is simply no clear evidence for the claim that mHealth technologies strengthen patient autonomy (Morley and Floridi, 2020). Furthermore, it has been shown that the supposed empowerment in digital technologies is often a fig leaf for hidden agendas (Rubeis et al., 2018). The rhetoric of autonomy and empowerment is often used to sugarcoat commodification and work optimization using AI-technologies within the healthcare system (Dillard-Wright, 2019). Following Lupton (Lupton, 2013), the emphasis on patient engagement and the patient-as-consumer approach is the outcome of a "neoliberal" agenda that promotes the shift of responsibilities from the collective to the individual. In a similar vein, Sharon (Sharon, 2016) interprets activation of patients as a means of cost-reduction in the health sector, e.g., by reducing contact with health professionals.

In the fourth vision, individuals are not framed as consumers, but as members of a public consisting of citizens with rights (European Commission, 2020; World Health Organisation, 2021). This vision focusses on public engagement and co-design in AI-development as well as regulation and governance. The inherent tension between this vision and the economic and political reality is obvious (Wilson, 2022). AI is almost exclusively shaped by the private sector, which also influences the development of standards and regulations, and public citizen participation in development and decision-making process concerning AI is virtually non-existent.

4.3 The democracy: A libertarian utopia

Building on the dominant patient-as-consumer approach regarding the demos, it is not hard to make the connection to the corresponding type of democracy that underpins the first vision of democratization. By taking the ideas of self-ownership and individual responsibilities as a given, the first vision of democratization of healthcare through AI, channels Lockean individualism typical for libertarian thinking (Olsthoorn, 2019). It reduces democracy to the libertarian idea of being free from external interference by authorities, in our case-medical paternalism. The connection to libertarian theories is sometimes implicit, as in the case of Topol's approach, sometimes explicit, e.g., by referring to libertarian authors (Montes and Goertzel, 2019). Two ideas are crucial for this libertarian approach: engagement and self-management on behalf of users and decentralized access to data.

Engagement and self-management manifest themselves in the aforementioned strong focus on AI-based mHealth technologies. We have seen that these technologies are promoted as enablers of patient empowerment. The use of mHealth for data collection is also sometimes framed as a means to circumvent economic, legal, or political restraints to improve healthcare services (Weissglass, 2021). When faced with supposedly dysfunctional structures and infrastructures, patients should take their health into their own hands. Healthcare is thus framed as an individual responsibility.

Calls for free-flowing data and universal access to technologies also fit well with typical libertarian ideas. Decentralized data repositories and block chain solutions are promoted as pillars of a more democratic healthcare (Montes and Goertzel, 2019; Traub et al., 2019; Bhattacharya et al., 2021). Following this notion, absence of restrictions and access to data and technologies have a democratizing effect.

However, the involvement of Big Tech in medicine and the immense power, that is, given to a rather small number of companies by letting them handle large amounts of data could also be seen as a direct threat to democracy instead of enabling a more democratic healthcare. The idea that a data-based free-market utopia will make healthcare more democratic suffers from a specific libertarian blind spot, i.e., the awareness for structural inequalities and asymmetric power and property relations. These issues manifest themselves in the so-called big data divide between those who provide data and those who possess the means to process data (Mittelstadt and Floridi, 2016). Granting broader access and enabling restriction free data-sharing alone will not resolve this issue, since the intellectual (knowledge) and material resources for processing data remain unequally distributed. The focus on free-flowing data and access ignores the existing digital power concentration that shapes the infrastructures and required markets (Nemitz, 2018).

This begs the question of democratic control, which is a crucial aspect when it comes to big data in healthcare (Gould, 2019; Sangiovanni, 2019). In this regard, democratic control has three main objectives: ensuring that all that are affected by decisions or actions can participate in the decision-making process (all-affected principle), focusing on the common good, and enabling individuals to make use of their freedom. Democratic control thus requires participation, deliberation, and representation that aim at compensating for or preventing unequal power and property relations (Gould, 2019). In order to democratize AI-based healthcare technologies, the infrastructure for developing and distributing them would have to be an object of democratic control instead of trusting in an invisible hand.

4.4 Strengths and limitations

To the best of our knowledge, this is the first study that explored in depth, the different versions of terms "democratization" and "democratizing" that are used in the context of AI in medicine and healthcare. Qualitative methods allowed us to explore and map the emerging topic in detail.

Our mapping of the visions also has limitations, which need to be considered to qualify our results. First, the use of peer-reviewed articles and documents in English limited our attention to comparatively privileged voices in the discourse on AI in medicine and healthcare. A more thorough mapping of visions of democratization, which could capture alternative visions of democratizing AI, would need to extend the materials to other languages and materials. Second, in light of the emerging nature of discourses on democratizing AI in healthcare and medicine, we have neither quantified our results, nor explored the relationship between visions articulated in documents to practices, institutions, and materialities outside the documents. We also did not analyze the political, social, and economic forces that shape the clusters we identified in detail. Further research is needed to address this topic. We see our paper as the first step in that direction.

5 Conclusion

In this paper, we mapped the different ways the terms "democratization" and "democratizing" are used in the discourse on AI in medicine and healthcare and performed a normative analysis of the findings, embedding them in normative engagements with data-intensive technologies. We derived four clusters of visions of democratizing AI from our qualitative analysis of peer-reviewed articles and grey literature: A first cluster of visions focusses on AI for the people and aims at democratizing medicine and healthcare through the further implementation of AI. These visions frame AI as a technological fix to problems in healthcare systems, such as inequity and access barriers, and lead to a personalization of medical practice. A second set of visions shifts to AI by the people and encompasses visions of democratizing AI in medicine and healthcare by facilitating better access to AI technologies for healthcare professionals without a background in data science or informatics. According to this vision, the provision of knowledge and tools, e.g., ready-made toolkits and code-free interfaces, may facilitate better access to AI technologies and enable medical practitioners to contribute to developing AI-based systems and better integrate them in their practice. Access also plays a crucial role in the third set of visions that we described as people in AI. They focus on democratizing access to and oversight of data. The aim is to make the data, that are used to develop, train, and evaluate AI, more representative as well as accessible by applying various strategies for decentralizing data generation and broader dissemination of data. A fourth cluster of visions seeks to make AI an object of the rule by the people and thus a matter of democratic governance. Democratic governance may imply participation of publics in the design and development processes of AI-based systems or the regulation of AI by democratically legitimized authorities.

Our normative analysis shows that democratization in the context of medical AI can be seen as an example of the kind of rhetoric Daub described that aims at shaping our view of how we could or should live. Weak and strong versions of technical solutionism cut across all visions. The supposed potential of AI technologies to fix primarily social and political problems not only raises false hopes but may also obscure the need for alternative solutions. We also highlight that the envisioned democratization in most visions mainly focuses on patients as consumers and relies on or limits itself to free market-solutions. This rather libertarian understanding of democracy ignores the

need for formulating rights that ensure fair distribution and democratic control of AI and the services it provides. This is especially an issue since the development and implementation of AI is largely driven by a small number of companies usually referred to as Big Tech. Protecting the interests of those affected by AI and facilitating a real democratization *of* AI in medicine and healthcare, or even democratization of medicine and healthcare through AI thus requires a rights-based approach instead of technological solutionism or reliance on market forces.

However, the different ways in which the term democratization is used also suggests that this vision is not universally shared. While the imagination that AI-based technologies will help us fix problems in medicine and healthcare underpins all these visions, different understandings of the identity and the place of the demos also show that—who should have the power to envision these futures and who ought to be involved in striving for them, is contested.

This does not mean that democratization is a false term in this context. Our mapping of the ways in which this term is used has helped us to show that it helps to raise important questions about the development and use of AI in medicine and healthcare, about the kind of futures that we strive to attain through AI-based technologies, and who we think ought to be involved and have a say when working towards that future. Specifically, it directs attention to implicit definitions of the demos that ought to be engaged in the development, use, and oversight of AI, and different practices of engagement. In the current uses of the term, these questions are often answered tacitly. The nature of the demos and its appropriate place are presumed. Developing "democratizing AI" into a more robust concept could help us think more systematically through the implicit normativity in the development of AI-based systems. It could be used as what Herbert Blumer (Blumer, 1954) named a "sensitizing concept"—i.e., a concept that makes us attentive to questions to be asked and issues to be taken care of.

For democratization and democracy to be more than misnomers here, a much more substantial theoretical foundation is needed. Democratization in the context of AI in healthcare requires defining and envisioning a set of social goods. It also needs deliberative processes and modes of participation to ensure that those affected by AI in healthcare have a say on its development and use.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

GR made substantial contributions to the conception or design of the work; or the acquisition, analysis or interpretation of data for the work, drafting the work or revising it critically for important intellectual content, provided approval for publication of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. KD was involved in the acquisition, analysis or interpretation of data for the work, drafting the work or revising it critically for important intellectual content and provided approval for publication of the content. IM made substantial contributions to the conception or design of the work; or the acquisition, analysis or interpretation of data for the work, drafting the work or revising it critically for important intellectual content, and provided approval for publication of the content.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Academy of Medical Sciences (2019). *Artificial intelligence and health. Summary report of a roundtable held on 16 January 2019*. [Online] [Accessed] Available: https://acmedsci.ac.uk/file-download/77652269.

Allen, B., Agarwal, S., Kalpathy-Cramer, J., and Dreyer, K. (2019). Democratizing AI. *J. Am. Coll. Radiol.* 16, 961–963. doi:10.1016/j.jacr.2019.04.023

Barclay, L. (2020). *Patient engagement: Why the public should be part of the conversation around AI in healthcare [Online]*. [Accessed] Available: https://www.aidence.com/articles/patient-engagement-health-ai/.

Bhattacharya, S., Hu, Z., and Butte, A. J. (2021). Opportunities and challenges in democratizing immunology datasets. *Front. Immunol.* 12, 647536. doi:10.3389/fimmu.2021.647536

Blumer, H. (1954). What is wrong with social theory? *Am. Sociol. Rev.* 19, 3–10. doi:10.2307/2088165

Bond, R. R., Koene, A. R., Dix, A. J., Boger, J., Mulvenna, M. D., Galushka, M., et al. (2019a). *Democratisation of usable machine learning in computer vision*. ArXiv, abs/1902.06804.

Bond, R. R., Mulvenna, M. D., Wan, H., Finlay, D. D., Wong, A., Koene, A., et al. (2019b). *"Human centered artificial intelligence: Weaving UX into algorithmic decision making." in RoCHI 2019: International Conference on Human-Computer Interaction, Bucharest, Romania.* Available at:http://rochi.utcluj.ro/articole/7/RoCHI2019-Bond.pdf

Braun, K., and Konninger, S. (2018). From experiments to ecosystems? Reviewing public participation, scientific governance and the systemic turn. *Public Underst. Sci.* 27, 674–689. doi:10.1177/0963662517717375

Briganti, G., and Le Moine, O. (2020). Artificial intelligence in medicine: Today and tomorrow. *Front. Med.* 7, 27. doi:10.3389/fmed.2020.00027

Burnside, M., Crocket, H., Mayo, M., Pickering, J., Tappe, A., and De Bock, M. (2020). Do-it-yourself automated insulin delivery: A leading example of the democratization of medicine. *J. Diabetes Sci. Technol.* 14, 878–882. doi:10.1177/1932296819890623

Daub, A. (2020). *What tech calls thinking: An inquiry into the intellectual bedrock of silicon valley*. New York, NY: FSG Originals, MacMillan US.

Dewey, J. (1927). *The public and its problems*. Athens, OH: Swallow Press, 219. 1954.

Dibia, V., Cox, A., and Weisz, J. (2018). *Designing for democratization: Introducing novices to artificial intelligence via maker kits. arXiv preprint arXiv:1805.10723.*

Dillard-Wright, J. (2019). Electronic health record as a panopticon: A disciplinary apparatus in nursing practice. *Nurs. Philos.* 20, e12239. doi:10.1111/nup.12239

Donia, J., and Shaw, J. A. (2021). Co-design and ethical artificial intelligence for health: An agenda for critical research and practice. *Big Data & Soc.* 8, 205395172110652. doi:10.1177/20539517211065248

Doubleday, R., and Wynne, B. (2011). *Despotism and democracy in the United Kingdom: Experiments in reframing citizenship*. Reframing rights. Cambridge, MA: MIT Press Scholarship One.

European Commission (2020). *On artificial intelligence—a European approach to excellence and trust*. European Union Brussels. Available at: https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

Felt, U., and Fochler, M. (2008). The bottom-up meanings of the concept of public participation in science and technology. *Sci. Pub. Pol.* 35, 489–499. doi:10.3152/030234208x329086

Gallie, W. B. (1955). Essentially contested concepts. *Proc. Aristot. Soc.* 56, 167–198. JSTOR. doi:10.1093/aristotelian/56.1.167

Garvey, C. (2018). A framework for evaluating barriers to the democratization of artificial intelligence. *Thirty-Second AAAI Conf. Artif. Intell.* 32, 12194. doi:10.1609/aaai.v32i1.12194

Garvey, C. (2019). Hypothesis: Is "terminator syndrome" a barrier to democratizing artificial intelligence and public engagement in digital health? *Omics a J. Integr. Biol.* 23, 362–363. doi:10.1089/omi.2019.0070

Gottweis, H., Braun, K., Hajer, M., Loeber, A., Metzler, I., Reynolds, L., et al. (2008). Participation and the new governance of life. *BioSocieties* 3, 265–286. doi:10.1017/s1745855208006194

Gould, C. C. (2019). How democracy can inform consent: Cases of the internet and bioethics. *J. Appl. Philos.* 36, 173–191. doi:10.1111/japp.12360

Gupta, I. (2020). *Decentralization of artificial intelligence: Analyzing developments in decentralized learning and distributed AI networks.* arXiv preprint arXiv:1603.04467.

Hagendijk, R., and Irwin, A. (2006). Public deliberation and governance: Engaging with science and technology in contemporary europe. *Minerva* 44, 167–184. doi:10.1007/s11024-006-0012-x

Harambam, J., Helberger, N., and Van Hoboken, J. (2018). Democratizing algorithmic news recommenders: How to materialize voice in a technologically saturated media ecosystem. *Philos. Trans. A Math. Phys. Eng. Sci.* 376, 20180088. doi:10.1098/rsta.2018.0088

Hilgartner, S. (2015). "Science and democracy: Making knowledge and making power in the biosciences and beyond," in *Capturing the imaginary : Vanguards, visions and the synthetic biology revolution Stephen Hilgartner*. Editors S.Hilgartner, C.MIiller, and R. Hagendijk (London: Routledge).

Himmelreich, J. (2022). "Against "democratizing AI"," in *AI & society: Journal of knowledge, culture and communication*, 1–14.

Howard, M. (2021). Wearables, the marketplace and efficiency in healthcare: How will I know that you're thinking of me? *Philos. Technol.* 34, 1545–1568. doi:10.1007/s13347-021-00473-4

Jasanoff, S. (2015). "Future imperfect: Science, technology, and the imaginations of modernity," in *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power*. Editors S. Jasanoff and S.-H. Kim (Chicago: University of Chicago Press).

Jasanoff, S. (2013). *Science and public reason*. London: Routledge.

Kobayashi, Y., Ishibashi, M., and Kobayashi, H. (2019). How will "democratization of artificial intelligence" change the future of radiologists? *Jpn. J. Radiol.* 37, 9–14. doi:10.1007/s11604-018-0793-5

Laurent, B. (2017). Democratic experiments: Problematizing Nanotechnology and Democracy in Europe and the United States *[online]*. Cambridge and London: MIT Press. [Accessed] Available: https://mitpress.mit.edu/books/democratic-experiments.

Lupton, D. (2013). The digitally engaged patient: Self-monitoring and self-care in the digital health era. *Soc. Theory Health* 11, 256–270. doi:10.1057/sth.2013.10

Lyu, L., Li, Y., Nandakumar, K., Yu, J., and Ma, X. (2020). How to democratise and protect AI: Fair and differentially private decentralised deep learning. *IEEE Trans. Dependable Secure Comput.* 19, 1003. doi:10.1109/tdsc.2020.3006287

Marres, N. (2007). The issues deserve more credit: Pragmatist contributions to the study of public involvement in controversy. *Soc. Stud. Sci.* 37, 759–780. doi:10.1177/0306312706077367

Metzler, I., and Åm, H. (2022). How the governance of and through digital contact tracing technologies shapes geographies of power. *policy Polit.* 50, 181–198. doi:10.1332/030557321x16420096592965

Mishra, S. (2022). Artificial intelligence: A review of progress and prospects in medicine and healthcare. *J. electron. Electromed. Eng. Med. Inf.* 4 (1), 1–23. doi:10.35882/jeeemi.v4i1.1

Mittelstadt, B. D., and Floridi, L. (2016). The ethics of big data: Current and foreseeable issues in biomedical contexts. *Sci. Eng. Ethics* 22, 303–341. doi:10.1007/s11948-015-9652-2

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nat. Mach. Intell.* 1, 501–507. doi:10.1038/s42256-019-0114-4

Montes, G. A., and Goertzel, B. (2019). Distributed, decentralized, and democratized artificial intelligence. *Technol. Forecast. Soc. Change* 141, 354–358. doi:10.1016/j.techfore.2018.11.010

Morley, J., and Floridi, L. (2020). The limits of empowerment: How to reframe the role of mHealth tools in the healthcare ecosystem. *Sci. Eng. Ethics* 26, 1159–1183. doi:10.1007/s11948-019-00115-1

Morozov, E. (2013). *To save everything, click here: The folly of technological solutionism*. New York, NY: Public Affairs.

Mulvenna, M. D., Bond, R., Delaney, J., Dawoodbhoy, F. M., Boger, J., Potts, C., et al. (2021). Ethical issues in democratizing digital phenotypes and machine learning in the next generation of digital health technologies. *Philosophy Technol.* 34, 1–16. doi:10.1007/s13347-021-00445-8

Nature Machine, I. (2021). People have the AI power. *Nat. Mach. Intell.* 3, 275.

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philos. Trans. A Math. Phys. Eng. Sci.* 376, 20180089. doi:10.1098/rsta.2018.0089

Olsthoorn, J. (2019). SELF-OWNERSHIP and despotism: Locke on property in the person, divine dominium of human life, and rights-forfeiture. *Soc. Phil. Pol.* 36, 242–263. doi:10.1017/s0265052519000438

Porras, A. R., Rosenbaum, K., Tor-Diez, C., Summar, M., and Linguraru, M. G. (2021). Development and evaluation of a machine learning-based point-of-care screening tool for genetic syndromes in children: A multinational retrospective study. *Lancet. Digit. Health* 3, e635–e643. doi:10.1016/S2589-7500(21)00137-0

Rubeis, G., Schochow, M., and Steger, F. (2018). Patient Autonomy and quality of care in telehealthcare. *Sci. Eng. Ethics* 24, 93–107. doi:10.1007/s11948-017-9885-3

Rubeis, G. (2020). The disruptive power of Artificial Intelligence. Ethical aspects of gerontechnology in elderly care. *Arch. Gerontol. Geriatr.* 91, 104186. doi:10.1016/j.archger.2020.104186

Saldivar-Gonzalez, F. I., Aldas-Bulos, V. D., Medina-Franco, J. L., and Plisson, F. (2022). Natural product drug discovery in the artificial intelligence era. *Chem. Sci.* 13, 1526–1546. doi:10.1039/d1sc04471k

Sangiovanni, A. (2019). Democratic control of information in the age of surveillance capitalism. *J. Appl. Philos.* 36, 212–216. doi:10.1111/japp.12363

Schreier, M. (2012). *Qualitative content analysis in practice*. London: Sage publications.

Sharon, T. (2016). Self-tracking for health and the quantified self: Re-articulating autonomy, solidarity, and authenticity in an age of personalized healthcare. *Philos. Technol.* 30, 93–121. doi:10.1007/s13347-016-0215-5

Sharon, T. (2018). When digital health meets digital capitalism, how many common goods are at stake? *Big Data & Soc.* 5, 205395171881903. doi:10.1177/2053951718819032

Sikpa, D., Fouquet, J. P., Lebel, R., Diamandis, P., Richer, M., and Lepage, M. (2019). Automated detection and quantification of breast cancer brain metastases in an animal model using democratized machine learning tools. *Sci. Rep.* 9, 17333. doi:10.1038/s41598-019-53911-x

Silverman, D. (2015). *Interpreting qualitative data*. Los Angeles, London: Sage.

Steinhubl, S. R., and Topol, E. J. (2018). Digital medicine, on its way to being just plain medicine. *NPJ Digit. Med.* 1, 20175–20181. doi:10.1038/s41746-017-0005-1

Timmermans, S., and Berg, M. (2003). The practice of medical technology. *Sociol. Health Illn.* 25, 97–114. doi:10.1111/1467-9566.00342

Topol, E. (2012). *The creative destruction of medicine: How the digital revolution will create better health care*. New York, NY: Basic Books.

Topol, E. (2015). *The patient will see you now: The future of medicine is in your hands*. New York: Basic Books.

Topol, E. (2019). *Deep medicine: How artificial intelligence can make healthcare human again*. New York: Basic Books.

Traub, J., Quiané-Ruiz, J.-A., Kaoudi, Z., and Markl, V. (2019). *Agora: Towards an open ecosystem for democratizing data science & artificial intelligence*. ArXiv, abs/1909.03026.

Ulnicane, I. (2022). Artificial intelligence in the European union: Policy, ethics and regulation, in *The routledge handbook of European integrations*. Taylor & Francis.

Vanhorn, K., and Çobanoglu, M. C. (2021). Democratizing AI in biomedical image classification using virtual reality. *Virtual Real.* 26, 159–171. doi:10.1007/s10055-021-00550-1

Weissglass, D. E. (2021). Contextual bias, the democratization of healthcare, and medical artificial intelligence in low- and middle-income countries. *Bioethics* 36, 201–209. doi:10.1111/bioe.12927

Wilson, C. (2022). Public engagement and AI: A values analysis of national strategies. *Gov. Inf. Q.* 39, 101652. doi:10.1016/j.giq.2021.101652

Wong, P.-H. (2019). Democratizing algorithmic fairness. *Philos. Technol.* 33, 225–244. doi:10.1007/s13347-019-00355-w

Wong, P.-H. (2020). Democratizing algorithmic fairness. *Philos. Technol.* 33, 225–244. doi:10.1007/s13347-019-00355-w

World Health Organisation (2021). *Ethics and governance of artificial intelligence for health:* WHO guidance *[online]*. [Accessed] Available: https://www.who.int/publications/i/item/9789240029200.

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 22 August 2022
DOI 10.3389/fgene.2022.902960

Check for updates

OPEN ACCESS

EDITED BY
Gabriele Werner-Felmayer,
Innsbruck Medical University, Austria

REVIEWED BY
J. Benjamin Hurlbut,
Arizona State University, United States
Prasanta Panigrahi,
Indian Institute of Science Education and Research Kolkata, India

*CORRESPONDENCE
Tabea Ott,
tabea.ott@fau.de

SPECIALTY SECTION
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Genetics

RECEIVED 23 March 2022
ACCEPTED 15 July 2022
PUBLISHED 22 August 2022

CITATION
Ott T and Dabrock P (2022), Transparent human – (non-) transparent technology? The Janus-faced call for transparency in AI-based health care technologies.
Front. Genet. 13:902960.
doi: 10.3389/fgene.2022.902960

COPYRIGHT
© 2022 Ott and Dabrock. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Transparent human – (non-) transparent technology? The Janus-faced call for transparency in AI-based health care technologies

Tabea Ott* and Peter Dabrock

Chair of Systematic Theology II (Ethics), Faculty of Humanities, Social Sciences, and Theology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

The use of Artificial Intelligence and Big Data in health care opens up new opportunities for the measurement of the human. Their application aims not only at gathering more and better data points but also at doing it less invasive. With this change in health care towards its extension to almost all areas of life and its increasing invisibility and opacity, new questions of transparency arise. While the complex human-machine interactions involved in deploying and using AI tend to become non-transparent, the use of these technologies makes the patient seemingly transparent. Papers on the ethical implementation of AI plead for transparency but neglect the factor of the "transparent patient" as intertwined with AI. Transparency in this regard appears to be Janus-faced: The precondition for receiving help - e.g., treatment advice regarding the own health - is to become transparent for the digitized health care system. That is, for instance, to donate data and become visible to the AI and its operators. The paper reflects on this entanglement of transparent patients and (non-) transparent technology. It argues that transparency regarding both AI and humans is not an ethical principle per se but an infraethical concept. Further, it is no sufficient basis for avoiding harm and human dignity violations. Rather, transparency must be enriched by intelligibility following Judith Butler's use of the term. Intelligibility is understood as an epistemological presupposition for recognition and the ensuing humane treatment. Finally, the paper highlights ways to testify intelligibility in dealing with AI in health care ex ante, ex post, and continuously.

KEYWORDS

Transparency, AI, Learning Systems, Intelligibility, Health Care, Ethics, Infraethics, Data

Introduction

Artificial Intelligence (AI) is an umbrella term for different technologies such as Machine Learning (ML) and Deep Learning (DL) (Iqbal et al., 2021, 11–13). According to the UNESCO, AI systems are "information-processing technologies that integrate models and algorithms that produce a capacity to learn [. . .] leading to outcomes such as prediction and decision-making" (UNESCO, 2022, 10). While they are associated with great hopes for improving the quality of life, they also pose several ethical challenges and require good governance. This is especially important when it comes to health care. AI is expected to be used in nearly all areas of medicine: for improvement in image evaluation and diagnosis finding of different malignancies (Mentzel 2021, 694–704; Aubreville et al., 2019, 67–85; Kashif et al., 2021, 74) up to the detection of stress (Hwang et al., 2018; Oskooei et al., 2021), depression (Uddin et al., 2022, n. p.), and other mental diseases (Lee et al., 2021, 856–864). For the AI to actually improve human diagnosis and treatment, it must be trained with a large amount of non-messy data. These data are categorized as highly sensitive by the GDPR Art. 9. Data relevant for AI-based health care includes not only bodily data but also data collected from daily life. Transactional data from grocery stores, socioeconomic status, education, neighborhood, and physical environment, for example, can become relevant for public health policy (Lu et al., 2020; Artiga and Hinton, 2018, n. p.). These examples show how the measurement of the human and their transparency is extended. At the same time, methods of DL are deployed. This confronts stakeholders with self-learning systems based on a deep neural network with multiple hidden layers (Goswami, 2020, 8–10; Maschewski and Nosthoff, 2021, n. p.). On the one hand, these multiple hidden layers increase the accuracy of a system. On the other hand, they turn the system into a "black box" whose mapping between input and output is no longer comprehensible to the relevant stakeholders (Zerilli et al., 2021, 28–29). Although there are technical approaches to open the black box, questions of modality, execution, and consequences are still open (Lima et al., 2022, 1–18; Arik and Pfister 2019, n. p; Lundberg and Lee, 2017). However, the opaqueness of the AI system is not solely based on the technical complexity of the system. Transparency issues also arise from human-machine interaction within the greater context of a social web of norms, values, and preconceptions that precede and follow the application (Latour 2000). The context of data acquisition, classification (Bowker and Star, 2000, 10–12) as well as the further handling of the output poses challenges for transparency as well. With this change in health care towards its increasing opacity, new questions of transparency arise. Moreover, almost all recent recommendations for governing AI applications cover this topic. Transparency appears as a decisive feature AI should have. This observation provides the starting point of the analysis, which studies the concept of transparency and the assumptions on which the concept is based. As a first step, it should be noted that transparent AI is closely related to the transparency of the people interwoven with it, especially the patients. While the complex human-machine interactions, as well as the AI system itself, tend to become non-transparent, the patient instead becomes seemingly "transparent" by the use of these technologies. Papers on the ethical implementation of AI plead for transparent AI but neglect the factor of the seemingly more and more transparent patient as intertwined with AI. The aim of the paper is to give depth to the concept of transparency and raise awareness for a certain ambiguity. Transparency is "Janus-faced" and can, under certain circumstances, harm human beings and their entitlement to human dignity. Giving more data does not necessarily lead to desired outcomes - e.g., better treatment. The risks and benefits of becoming transparent are not distributed equally among people (Seyyed-Kalantari et al., 2021; Mann and Matzner, 2019; Braun and Hummel, 2022, 4). Obermeyer et al., for example, showed that an AI algorithm perpetuated the systematic inequalities for People of Color. The algorithm identified People of Color as a group with poorer access to care. But instead of changing the situation for the better, the use of the algorithm resulted in less health care spending on Black patients to equally sick White patients (Obermeyer et al., 2019; Röösli et al., 2021, 191). Another example of harmful transparency is the handling of health data of Indigenous people (not only) during the COVID-19 pandemic (Carroll et al., 2019; Carroll et al., 2021). The data collected about Indigenous people is rarely by or for Indigenous people's purposes (Carroll et al., 2019, 3; Walter, 2018, n. p.). Finally, harmful transparency may result from the connection between the health care system and other economically oriented institutions. In Germany, it is nearly impossible to become a civil servant or to get insured against occupational disability if diagnosed with certain conditions. In a second part, the paper offers a suggestion for coping responsibly with this ambivalence. Transparency will then be presented as an "infraethical" (Floridi, 2017, 391–394) prerequisite that needs to be complemented by the actual ethical notion of intelligibility. Here, intelligibility, following Judith Butler, is vital for the humane treatment of a person. For this reason, transparency in the context of AI should be enriched by the concept of intelligibility. Thereby, the vulnerability of an increasingly transparent patient in the digitized treatment situation can be tackled. Finally, building on the concept of intelligibility, participatory strategies for practice are proposed.

The claim for transparent AI in current governance recommendations

One of the key principles for governing AI in health care and beyond appears to be transparency. It is one of the most elaborated terms in current governance guidelines (Fjeld et al., 2020, 41; Jobin et al., 2019, 391; UNESCO, 2022; High-Level Expert Group on Artificial Intelligence, 2019). Often, it is mentioned together with explainability or interpretability. This

paper follows John Zerilli by distinguishing between transparency as an umbrella term and explainability as one of its subcategories (Zerilli et al., 2021, 25). Explainability and the discourse around explainable AI (XAI), according to Zerilli, is very much concerned with technical transparency - especially the transparency of the algorithm (view also: Lima et al., 2022, 3; ACM US Public Policy Council, 2017; Floridi, 2017, 391–394; Arrietta et al., 2020, 85, 88–90). However, transparency covers more than the understandability of the algorithmic decision-making. It encompasses the social dimension regarding responsibility, accessibility, or justifiability, the role of the patient or physician, and last but not least reflections on social attributions or bias as well. In this paper, the focus lies on the broader and fuzzier concept of transparency. When facing the implementation of transparent AI, several difficulties arise.

First, transparency is an ill-defined term, that is used differently in various contexts. This can be illustrated by the following simple questions, which, despite their straightforward nature, hardly ever receive a clear answer: what is transparency? What is to be made transparent? To whom? To what end? And how is it finally implemented? While the last question concerns practical effects, the first three questions introduce a deeper level of transparency, which is often disregarded in current governance papers. Many of those view transparency as an ethical principle (Fjeld et al., 2020, 41–45; High-Level Expert Group on Artificial Intelligence, 2019, 13, 18; WHO, 2021, 26–28) which, adapted in modules (e.g., open-source data), can be implemented in practice. The questions already show that transparency is about making information available, while leaving open what information, for whom, and for what purpose. However, it is quite clear that making transparent requires different action depending on the addressee. Patients have different know-how and emotional involvement than developers, physicians, or deployers. Accordingly, individual addressees of transparency (transparent to whom?) often go hand in hand with different objectives (transparent to what end?). For instance, making the AI system transparent to a patient is usually associated with the aim of effecting trust (Felzmann et al., 2019, 5; Adams, 2018, 17; Lupton, 2015, 576). In contrast, making the AI system transparent to a developer focuses on efficiency or interoperability (Arrietta et al., 2020, 84; Zerilli et al., 2021, 24; Prabhakaran and Martin, 2020, 72). Finally, in societal or legal contexts transparency aims to sustain accountability (Diakopoulos, 2020, 197) or liability.

Outlining this basic definition problem of transparency leads to a first critical observation: there is no timeless or contextless agenda when making AI transparent. Transparency does not follow an all or nothing logic (Ananny and Crawford, 2018, 979; Zerilli et al., 2022, 7). It always (consciously or unconsciously) excludes crucial information and is highly dependent on its sociotechnical contexts (Hasselbalch, 2021, 10–11; Bowker and Star, 2000, 32). Thereby, transparency is treading a fine line between revealing too much information or too (use)less information. Both ways, revealing too much information and risking an information overflow as well as revealing too less or negligible information, would in the end lead to greater opacity. However, even if the balance succeeds, a remaining opacity stays. This is especially true for the complex sociotechnical process in which an AI is embedded. Not only the interplay between data sets and code yields opaqueness (Burrell, 2016, 5): the interaction of different actants (AI, data, humans) is the decisive factor that favors opacity. Transparency must reflect on these blind spots. It must be marked as a limited process, which is neither free of opacity nor reveals "truth" in any form. As Chesterman puts it: "illusory transparency can be worse than opacity" (Chesterman, 2021, 166).

Another important limitation of transparency is its ethical indifference. Transparency does not necessarily draw consequences from what is disclosed.

On the one hand, transparency does not entail ethical judgement. It does not yet constitute a framework with which to evaluate what has been disclosed. Even if a system is classified as transparent - and it has been shown that "making transparent" is very context-dependent and still contains opaque elements - it is not clear that discriminatory structures will be detected (Bowker and Star, 2000, 44–45). Although there is always bias or discrimination (in the sense of differentiation) attached to AI, some forms are considered harmful while others are not. Moreover, "bias is not simply a feature of data that can be eliminated; it is defined and shaped by much deeper social and organizational forces" (Cho, 2021, 2080). The German General Equal Treatment Act (Allgemeines Gleichbehandlungsgesetz, AGG), for example, provides a classification scheme for detecting harmful bias. It states: "The Act protects people who are discriminated against on the grounds of race or ethnic background, gender, religion or belief, disability, age, or sexual orientation" (Federal Anti-Discrimination Agency, 2019). However, discrimination is not easily detectable. First, bias can have different causes: Real world patterns of health inequality and discrimination, data bias resulting from discriminatory datasets, algorithmic bias due to deployment practices, or application injustice that occurs in the context of use (Leslie et al., 2021, 2). Second, AI can discriminate by proxy. This form of bias is even harder to detect (Calderon et al., 2019, 17). Proxy discrimination means that although protected attributes (e.g., gender or ethnicity) are not mapped in the data set, other characteristics (e.g., membership in a specific Facebook group etc.) can indicate them (Zerilli et al., 2021, 59). These other characteristics, so-called proxies, lead again to disadvantages and stigmatization for the affected individuals (cf. the works of Obermeyer et al., 2019; Prince and Schwarcz, 2020). Third, it gets even more problematic when the AI discriminates against new groups (e.g., pet owners or others), some of which are not at all comprehensible to humans and which are not protected by

the AGG or anti-discrimination law (Wachter, 2022). In case two (proxy discrimination) and three (new groups discriminated against) transparency is not sufficient. In these cases, the non-neutral classification system underlying transparency (e.g., the AGG or more subtle forms) does not necessarily protect the people discriminated against (cf. also Bowker and Star, 2000, 319–322; Mann and Matzner, 2019, 5).

On the other hand, transparency is not necessarily associated with power (Ananny and Crawford, 2018, 978). Transparency which pursues the goal of effecting trust does not primarily intend a self-critical analysis of the AI - especially an analysis that is open to revision and aims to bring about change. Thus, if there is no power or will to deal with an AI that has been unmasked as unfair, the concept of transparency loses all its merit as somewhat ethical principle or ideal. In fact, it is ethically indifferent. Often it is economic interests (e.g., insurances) or (historical) power ambivalences that hinder an appropriate response to transparency. One big issue, for example, is the data collection of marginalized groups. Without including them, transparency is likely to become a stigma (cf. Carroll et al., 2019; Wachter and Mittelstadt, 2019). In conclusion, it is misleading to view transparency as an ethical principle, as proclaimed by the current governance guidelines. It is not good per se, like justice, fairness, or non-maleficence, but Janus-faced. Therefore, transparency cannot be set up alongside ethical principles without acknowledging its ambivalence, which arises from its contextualization. This applies particularly to dealing with the permanent remainder of opacity and the handling of "uncovered" injustice.

Skepticism towards the "transparent patient"

Deeply intertwined with transparent AI is the transparent patient whose health data is the lifeblood of the machine. When it comes to transparency of AI in health care, sociotechnical human-machine interactions are involved. Therefore, to define and specify transparency regarding AI, it is essential to consider the transparency of the humans involved. Primarily, these are the data subjects, i.e., patients. Regarding AI, transparency is seen as a desirable goal, while transparency regarding the patient is rather treated with skepticism (Strotbaum and Reiß, 2017, 367–369; Maschewski and Nosthoff, 2021, n. p.; Prainsack, 2017, 50–51; Pasquale, 2015, 3–4). Here, too, the questions "transparent for whom?" and "transparent to what end?" show the multifaceted nature of transparency. Initially, it is hoped that by collecting large and diverse amounts of an individual's data, more accurate diagnoses and treatment decisions can be made. Even social or lifestyle data (e.g., a person's residence, shopping behavior etc.) become relevant (Hague, 2019, 222; Prainsack, 2017, 5–7). Together the various data types form a network of "biomedical big data" (WHO, 2021, 35). The aim is to make a person transparent to enable better diagnosis and treatment.

However, as before, the notion of transparency must be considered as essentially characterized by moments of opacity. The process of making humans transparent in health care is always fragmented. Here, too, classification systems have a significant influence. However, denying the fragmentarity and persistent opacity can lead to serious harm. Transparency is often associated with telling or revealing "the truth" (Ananny and Crawford, 2018, 974). The assumption that "truth is correspondent to, or with, a fact" (David, 2015, n. p.) then could lead to the conclusion that the more facts are revealed, the better the human self can be known (Ananny and Crawford, 2018, 974). In digitized health care, the patient appears as "data body" (Gitelman, 2013, 121). There is a danger that this data body becomes absolute with respect to the data subject: "The data body is the body by which you are judged in society, and the body which dictates your status in the world. What we are witnessing [. . .] is the triumph of representation over being" (Gitelman, 2013, 121). This statement makes clear that our digital representation in health care (and beyond) can gain an ontologically antecedent status. Not solely, but also Christian ethics draws attention to the mysteriousness, and not only puzzling nature of the human being (Jüngel, 2010, 534–536). A human is not the sum of their parts. The reality is more complex than an AI system can describe (Bowker and Star, 2000, 103; Stark, 2014, 94). Therefore, it is also important to consider how the person is embedded in the world in which they live. A diagnosis is preceded by very different notions of a good life, of health and illness etc. For the bodily person, who cannot explain herself entirely, there nevertheless must be the possibility of integrating the AI diagnosis into their narrated and responsive self-perception. It must be clear that the data show a certain part of the person but do not completely remove the opacity of the person - which is not necessarily bad, if seen as a mystery.

The second important aspect is again the ethical indifference of transparency. People give sensitive health data, i. a., with the expectation that it will benefit them. However, to be beneficial, the AI must meet various requirements. For instance, the AI must have been trained with sufficient comparative data from other patients of the same gender, age, disease etc. With lack or underrepresentation of training data of persons with, for example, a certain gender or sexual orientation, "Data Gaps" arise (Criado-Perez, 2019, 217–235; Norris et al., 2020, 2; Hatzenbuehler and Pachankis, 2021, 437; Dankwa-Mullan et al., 2021, 223–224). This can lead to poorer or even erroneous diagnoses and treatment decisions. For this reason, it bears greater risk for some people, especially minorities, to become transparent than for others. The problem gets even more intense when we consider the phenomenon of intersectional discrimination. A person can face discrimination not only on one but on the intersection of several characteristics. Kimberlé Crenshaw makes this particularly explicit regarding the

intersection of gender and race. She claims that antidiscrimination measures overlook people standing at the crossroads of discrimination, namely Black women (Crenshaw, 1989, 140, 149). However, intersectional discrimination can involve other factors as well. Which characteristic or which concurrence of different characteristics (obesity, disability, habits etc.) leads to stigmatization is not clear from the outset as these markers not necessarily appear in the analyzed data. Though, what shows up in the data are proxies. At a first glance, they do not appear as stigmata. For example, living in a certain neighborhood can function as a proxy (Prince and Schwarcz, 2020). Therefore, some people are skeptical about becoming transparent when providing data, for good reason. They are more likely to face increased vulnerability or precarity (Carroll et al., 2019; Butler, 2009, 25). This is due to the fact that there is no response to their transparency - first, on a diagnosis and treatment level, second, on a societal level (e.g., disadvantage in insurance). The data collection on Indigenous people in the United States illustrates this point clearly (Carroll et al., 2019, 3). Although transparency can be damaging to people, it can also bring them into focus and mobilize resources to address their situation (Casper and Moore, 2009, 79). Some may consider this a chicken-or-egg question: without transparency, there will be no better treatment and diagnoses in the future. Vice versa, if there is no prospect of getting good treatment, transparency will be experienced as harmful. Therefore, the paper aims to enrich the actual claim for transparency by a critical societal perspective. Transparency is not an ethical principle per se. A deeper philosophical analysis is needed to portray transparency as Janus faced and, one could say, "infraethical" (Floridi, 2017, 391–394) term.

Transparency as a Janus-faced infraethical concept

It is rightly pointed out that the demand for transparency initially sounds like a desirable ideal. Its status as an "inherent normative good" is often associated with other values such as truth-telling, honesty, or straightforwardness (Viola and Laidler, 2021, 23). Additionally, transparency is often misunderstood as revealing or showing the truth of something. Regarding AI applications, transparency is treated as "a panacea for ethical issues" (Mittelstadt et al., 2016, 6). However, transparency is not enough to address unfairness, discrimination, and opacity (Edwards and Veale, 2017, 21–22). The Janus-faced character of transparency becomes especially evident when considering, first, the remaining opacity and, second, the not necessarily given connection with awareness of injustice and the power to do something about it. As for the first point, the process of making transparent runs the risk of neglecting the veil that is lifted at that very moment (Kilian, 2013, n. p.). If the different filters (Who? What? To Whom? With what aim?), that determine to what extent the veil is lifted, are blanked out, transparency runs the risk of working as an illusion (Adams, 2018, 17). Regarding the second aspect, the only loose connection between transparency and awareness of malpractice or power to change may even threaten human dignity. If the question "Transparent to what end?" is answered with "To build trust" (concerning AI) or "To make visible for the health care system" (concerning humans) is not enriched by a watchful function against instrumentalization, it is misled and again cherishes an illusion.

Finally, this in-depth analysis of transparency as Janus-faced leads to the conclusion that transparency is not an ethical principle per se but an "infraethical" (Floridi, 2017, 391–394) concept. Infraethical means that it is a "not-yet-ethical framework of implicit expectations, attitudes and practices that can facilitate and promote moral decisions and actions" (Floridi, 2017, 192). Thus, regarding the learning system, transparency can build the ground for awareness of malpractice. As for the patient, it is necessary to give as much information as possible to get a chance for better diagnoses and treatment. However, as Floridi puts it: an injustice regime can be transparent, too, without being for this any less evil (Floridi, 2017, 393). To just apply infraethical transparency to foster successful facilitations (e.g., build trust, implement the technique easier, etc.) is not enough protection of human dignity. Rather, what Floridi suggests is that the infraethics must be combined with "morally good values (the right axiology)" (Floridi, 2017, 393) and be shaped by them. In the following, this reminder of Floridi will be taken as a basis. While Floridi primarily refers to transparency in relation to the design of AI, this view will be enriched by the previous investigations on the transparent human. With the focus on the human, a social anthropological perspective challenges the infraethical concept of transparency. It refers to the need for intelligibility, which can be made a critical requirement for transparency claims (cf. Figure 1). In demanding intelligibility as a verification framework for transparent humans in digitized health care, the identified obstacles of transparency will be tackled: That is first, non-reflected opacity, and second, ethical indifference from not recognizing harm and/or lack of agency.

How to avoid increased vulnerability caused by transparency? Using intelligibility as an ethical request

The previous section has shown that transparency is a Janus-faced concept. Its positive or negative impact on an individual is highly contextual and is often driven by a socio-historical or political agenda. Behind this is the idea that "making transparent", firstly, is itself a highly difficult and elusive process of negotiation between humans and the system. It always contains elements of opacity. Secondly, transparency does not yet produce an appropriate response to the exposure. Rather, it is ethically indifferent and can lead to increased vulnerability (cf. Figure 2). Having discussed the ambivalence of transparency, the final section of this paper addresses ways in

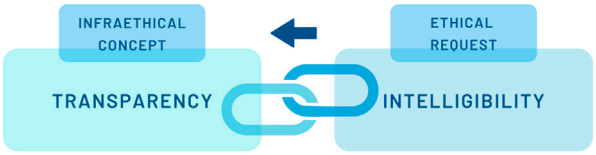


FIGURE 1
Intelligibility as ethical request.

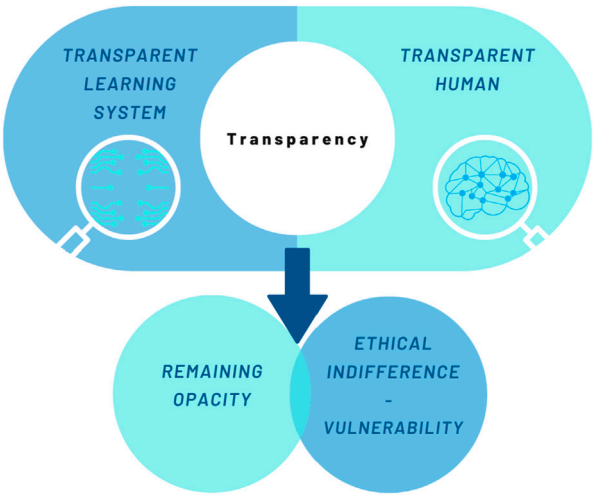


FIGURE 2
Characteristics of transparency.

which transparency can be reframed. The section moves on to describe how to avoid the possible negative effects of human transparency (increased vulnerability, stigma, or harm). Further, it offers a way to address unfairness, discrimination, and opacity in the context of transparent AI. For this purpose, the paper suggests enriching transparency with intelligibility. The term intelligibility is used here in accordance with Judith Butler. Butler uses it when she discusses what precedes personhood. She asks for a "new bodily ontology" in order to rethink "precariousness, vulnerability, injurability, interdependency, exposure, bodily persistence, desire, work and the claims of language and social belonging" (Butler, 2009, 2). Following Hegel, she assumes that humans are necessarily dependent on structures of recognition (Butler, 2009, 2–3). However, these structures of recognition are shaped by norms and classifications. Butler refers to norms as something that operates "to produce certain subjects as 'recognizable' persons and to make others decidedly more difficult to recognize" (Butler, 2009, 6). Consequently, the norms applied have an impact on individual vulnerability or precarity (Butler, 2009, 25). A deeper understanding is provided by Butler's distinction between apprehension and intelligibility. In *Frames of War* Butler defines apprehension as the "knowing that is not yet recognition" (Butler, 2009, 6). Intelligibility, on the other hand, is described as a "general historical schema or schemas that establish domains of the knowable" (Butler, 2009, 6). Butler exemplifies this with the category of gender, which is shaped by the schema of heteronormativity (Butler, 2007, 23–24). Further, Butler notes that intelligibility builds the ground for norms of recognizability. These norms of recognizability in turn prepare the way for recognition (Butler, 2009, 6). In summary, intelligibility is the foundation of the discourse of humans speaking as humans and not "as-if-humans" (Butler, 2004, p 30). Therein, it differs from transparency (and apprehension). Intelligibility is about something preceding (and at the same time following) the visible. In order to follow this ontological description, a distinction between the terms "to perceive" and "to recognize" may be helpful. While perceiving, on the one hand, only grasps the cognitive identification, recognizing, on the other hand, is part of an evaluative acknowledgment (Honneth, 2003, 26–29). The latter reaches to the very roots of being human: to recognize someone means to acknowledge someone as human and therefore as an addressee of human dignity. The concept of intelligibility, according to Butler, offers an explanation for how identities are constructed within normative practices (Halsema, 2005, 216). This way, human dignity violations can be detected. The presupposition of being recognized as a human is to be intelligible as a human. Intelligibility, understood this way, is circumscribed in existing norms. Norms can relate to sex, gender, desire, and race, for example. This observation is of great importance when it comes to AI. In a particular way, the classification and pattern recognition that constitutes AI shows that the technology is embedded in social norms and values (Jasanoff, 2016, 266).

Now, what does this mean for transparency?

First, transparency without the request for intelligibility can lead to the invisibility of a person. This phenomenon is covered in Alex Honneth's essay collection *Unsichtbarkeit. Stationen einer Theorie der Intersubjektivität* (*Invisibility. Stations of a Theory of Intersubjectivity*) where he describes invisibility as "looking through" a person (Honneth, 2003, 11). This form of disregard can be observed when significant characteristics of a person are not well represented in the training data of an AI, but the AI is still applied to that person. It is exceedingly likely that poorer or no diagnosis or treatment outcomes will be achieved. However, one can argue that transparency tackles exactly this problem: it reveals training data to prevent bias. This is certainly true. But the process of making transparent is also subject to norms and classifications - such as anti-discrimination law. As soon as bias by proxy, intersectional discrimination, wrongful classification (Brindaalakshmi, 2021, n. p.), or completely new - sometimes for humans not even understandable - groups (Wachter, 2022) are affected, transparency does not necessarily benefit the persons affected. All four of these

forms of discrimination cannot be identified through the application of existing norms. It needs the question of intelligibility to address these shortcomings of transparency.

Second, transparency that neglects intelligibility can lead to exposure of the human behind the data. If transparency leads to visibility, but visibility leads to social disadvantages, transparency can increase vulnerability. The data collection of Indigenous people (Carroll et al., 2019; Carroll et al., 2021) or Non-Binary people (Brindaalakshmi, 2021) illustrate this point clearly. Without receiving (medical) help or recognition, the exposure is stigmatization. It is perception without recognizing. Therefore, it can be argued that remaining non-transparent can be an advantage since transparency could involve experiencing violence. Becoming transparent can mean being subjected to a norm that is experienced as coercive: this applies especially to those people who do not fit in gender, body, or other group schemata - for people that defy classification.

Although Butler does not use the term intelligibility in an ethical sense, it nevertheless can build the starting point for ethical considerations. Beginning ethical consideration in the perspective of intelligibility questions the fundamentals of the human. It shows the necessity of keeping the notion of the human open to future articulation: "The nonviolent response lives with its unknowingness about the Other in the face of the Other" (Butler, 2004, 35–36). The subject itself is the starting point of the critical evaluation. Their life calls into question the frames which constitute the ontological field (Butler, 2009, 7). Butler considers the deviation from the norm as a potential disruption of the regulatory process that the norm constitutes (Butler, 2004, 52). This norm can be societal (e.g., gender), technological (e.g., due to non-representative data training), or sociotechnical (a combination of both). Some lives exist between, outside, or across the norm. They make a demand on the existing framework, revealing the shifting character of the grids of intelligibility. To detect the disruptive potential of those lives and to make use of it for improving AI is a future challenge. In this regard making transparent is like scratching the surface of the black box to make just a small detail visible. This visibility then has to put up with the critical inquiry of intelligibility. Transparency itself is not a changing force, but it gives hope that sensitivity for intelligibility can make transparency "better", e.g., through iterative transparency with, first, simultaneous knowledge of the opacity due to human-machine interaction and, second, the epistemological power of intelligibility. The challenge to be met is to establish intelligibility as a critical corrective for transparency. It focuses on the human, who is reliant on recognition to uphold human dignity. These considerations will be specified in the following with respect to the transparent human and, finally, derived from this, also for transparent systems.

Now, what is gained by introducing and supplementing the concept of transparency with intelligibility? The paper suggests to make the ethical test criterion for transparent AI the intelligible, i.e., recognizable/acknowledgeable human or patient. Where people are transparent but non-intelligible, as illustrated before with the examples of bias, intersectional discrimination, bias by proxy, discrimination of new and non-protected groups (Wachter, 2022), or data collection of marginalized groups, the existing frameworks become questionable. Intelligibility helps to uncover the "historical *a priori*" (Foucault, 1972, 126–128) in which the AI is embedded. In this regard, critical social analysis can provide starting points for the evaluation of AI and their outcomes. While transparency often follows an all or nothing logic, the term intelligibility opens the opportunity to uncover the essential elements of an AI system: does the system provide an adequate basis for rendering people intelligible? And does it do so not only ex ante during data collection and algorithm design but also continuously during implementation and adaptation, and finally ex post after the actual use case? Further asked: is a person's condition not only disclosed, but is it responded to appropriately in a medical decision-making situation? The response is the pivotal element intelligibility aims at. Paradoxically, it demands a question as an answer. "Who are you?" is the non-violent response to a human made transparent by AI systems. This question acknowledges the "clipping"-character of personhood. It allows the transparent patient to enter an exchange with the transparent AI, which cannot maintain its objectivity claim. Whether a person is intelligible is not possible to tell only from the outside. Thus, AI must be considered in a personal context of life. This contextualization is relevant for all types of AI. It leads, if necessary, to an extension of "grids of intelligibility" (Stark, 2014, 94). Thus, AI systems are tied back to social conditions and vulnerabilities. "The necessity of keeping our notion of the human open to a future articulation is essential to the project of international human rights discourse and politics" (Butler, 2004, 36). Intelligibility draws attention to the frames and norms transparent AI constructs. It challenges the process of making transparent to reveal the conditions of the foundations of being a person. Hence, the claim of intelligibility incorporates sensitivity to socio-historical and political power structures into measures of transparency (Mann and Matzner, 2019, 7).

Conclusion: A space for testifying intelligibility

Finally, it must be asked what transparency looks like that takes the vulnerability of the people involved seriously. Or even more specific: how to generate attention for frames of intelligibility in digitized health care environments? Further, how can this attention then lead to actual changes regarding non-harmful transparency of humans and AI? Typically, two lines of perspectives prevail in the governance of AI regarding the transparent patient (cf. Figure 3): the first shall be referred to here as the *data reduction* or *data parsimony perspectives*. They focus on the right to refuse provision of data. More precisely: a person

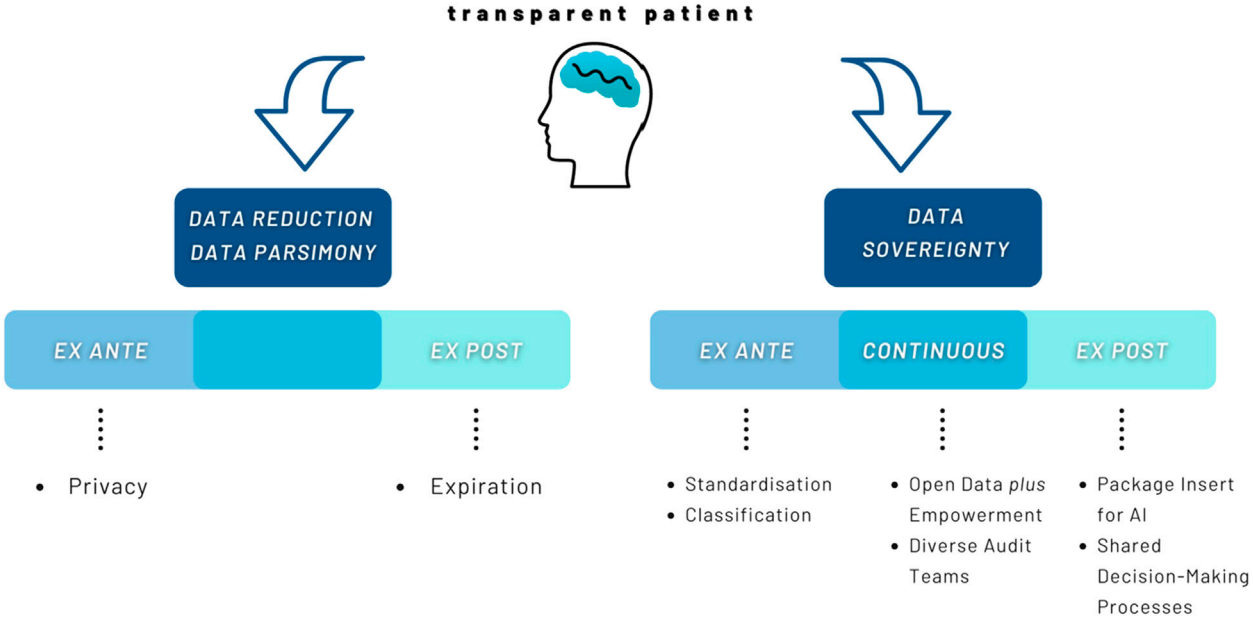


FIGURE 3
Data reduction vs. Data sovereignty.

needs to be sovereign in terms of the information she wants to give right at the beginning - i.e., ex ante. These perspectives often view organizations as surveillance organisms that misuse data or use humans as laboratory animals (Véliz, 2020, 39, 65). Their result is to give no or hardly any data at all or erase it as soon as possible (Mayer-Schönberger, 2009, 171–173) – i.e., ex post. This would, in a sense, lead to conscious and intentional hazarding of the consequences of a person's "non-intelligibility". Considering an increasingly digitized health care system and the benefits that AI offers in terms of diagnosis and treatment, not giving data would lead to health care disadvantages and inequality. Thus, non-intelligibility will not be tackled by giving no data. It rather will exacerbate inequalities and further increase societal problems.

The second line are to be referred to here as *data sovereignty perspectives*. They focus on the development process of AI as well as the outcomes of its use, i.e., ex ante, continuous, or ex post. Behind this is the conviction that not giving data is not an adequate solution to solve problems of (non-)intelligibility and thus violations of human dignity. Instead, data sovereignty perspectives try to deal with the data and suggest solutions on different levels (Hummel et al., 2021b, 22; Hummel et al., 2021a, 9–10; Wachter and Mittelstadt, 2019, 4–5, 13–14). While for data sovereignty perspectives non-intelligibility is not acceptable, the process of making intelligible must likewise meet certain standards in order to not be experienced as violent. Making intelligible goes beyond making transparent. It is sensible to the mysteriousness of the person and their right to be involved in meaning making processes around herself. Further, attention towards frames of intelligibility absorbs the digital exposure and endows it with recognition of harm and agency to address it. The awareness of the need for considering intelligibility as an ethical request for transparency leads to the persons affected first. The humans themselves are the stumbling blocks when it comes to detecting discrimination or stigmatization. Their life in relation to the frames of intelligibility brings forward questions and demands for AI. The patient must be given space for a "discourse of self-reporting and self-understanding" (Butler, 2004, 67).

This comes with several implications regarding the data collection and training process: first, if one fears to experience harm during the process of making intelligible, these fears must be taken seriously. In order to address this concern, spaces must be created in which non-intelligibility or transparency is brought up for discussion. Moreover, non-intelligibility must be the critical trigger point to change the system, in which it is better for people to take on health risks than to become transparent but non-intelligible.

Second, the data that are actually collected have to be standardized. Being aware of the issue of intersectional discrimination could mean involving patients to "capture their characteristics in a way that facilitates readability and interoperability" (Norori et al., 2021, 4). In the case of the Indigenous data collection with no purpose for the people concerned it could mean investing in community controlled data infrastructures (Carroll et al., 2021, 4). On the one hand,

this could ease the verification of the algorithm in the individual treatment situation. On the other hand, it contributes to data sovereignty at a very early stage. However, some thinkers conclude that protected attributes, like gender or ethnicity, should not be collected or classified at any rate (Zerilli et al., 2021, 59). An intelligibility-based approach to AI must reject this anti-classification approach. Rather, it pleads for a use case sensitive procedure that later discloses its modus operandi. This is due to the fact that in health care it is nearly impossible to exclude sensitive information. Often, these attributes appear by proxy and their discriminatory potential is much more difficult to detect afterwards. Also, it is impossible to perceive causal relations between discrepant factors if these are not collected (Ruf and Detyniecki, 2021, 19). Yet, the hope is to gain error-free results independent of a person's group affiliation.

Third, many papers mention the need for Open Data. Open Data and Open Science approaches focus on opening up the development process for people to interfere (Huston et al., 2019, 254). The idea behind this is that "if everything is disclosed, everyone has maximum control". However, several Open Data projects realize that "transparency [alone, authors] is insufficient - a data dump on a portal is not meaningful without sufficient awareness, education, and participation. The same principle applies to algorithms" (Turek, 2020, n. p.). It is not sufficient to only open up the data to the public. The opening process must be supplemented at the same time with opportunities for actual interaction and participation. A study by Schütz et al. shows that people are willing to interact and shape the technologies of the future (Schütz et al., 2019, 137). This goes far beyond transparency and simply being informed (Schütz et al., 2019, 137). The aim must be to enable a diverse set of people to actually check the data sets and to implement heterogenous audit teams. This empowerment of people (e.g., technical literacy, work environments etc.) must be corresponded to by the learning system. The algorithm must, for instance, enable (fast) frame adaptation processes. This is to meet the shifting "grids of intelligibility" and the need to integrate different voices which have not been recognized before. Nevertheless, as the open "debug" competition of Twitter's cropping Algorithm showed (Meunier et al., 2021, n. p.): datasets will not be free from bias nor is it possible to avoid bias completely at further processing stages. The reason for this is that bias is not necessarily caused by the technological component, the code, or the individual use case. It has a socio-historical dimension of discrimination as well (Meunier et al., 2021, n. p.). Therefore, an ex post security mechanism must be implemented that still allows individuals to request their intelligibility or object to their non-intelligibility in the use case. To identify whether the algorithm actually renders humans intelligible can be accompanied by a kind of "package insert" of a learning system. With a package insert for algorithms, an independent and diverse audit team could provide information about the development process and the nature of the training data. This information must be consciously considered within the shared decision-making process between patient and physician. Thus, the package insert functions as a safety or bias warning to avoid harm. It contributes to drawing attention to frames of intelligibility. By being alerted to which groups of people the algorithm produces worse results for, the medical professional can flexibly adjust her decisions. However, not only the medical professional but also the patient should be informed about this package insert in shared decision-making processes. In summary, transparency regarding AI and humans, enriched by the ethical request of intelligibility, demands to make the individual life courses audible. This is to tackle the persistent opacity of humans as well as of AI. Therefore, participatory approaches become important when practical implementation is concerned. This is implied in Bowker and Star's proposal for "a mixture of formal and folk classifications that are used sensibly in the context of people's lives" (Bowker and Star, 2000, 32). Additionally, the learning system must always be open for interference and revision. The shifting grids of intelligibility in everyday life must be representable in the algorithm. That means: the learning system has never finished learning.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

TO is main author of the article. PD commented on all parts of the article and contributed to the writing of the manuscript.

Funding

This work was supported by the German Research Foundation (DFG, German Research Foundation) under Grant SFB 1483 — Project-ID 442419336 and the Bavarian State Ministry of Health and Care, project grant No. PBN-MGP-2010-0004-DigiOnko. The funders had no influence on the study's design, analysis, and evaluation.

Acknowledgments

We are grateful to friends and colleagues in particular Hannah Bleher, Matthias Braun, Eva Maria Hille, David Samhammer, and Max Tretter for their helpful feedback on earlier versions of this paper. Special thanks to Fiona Bendig

and Serena Bischoff, who proofread the final manuscript, and to Svenja Hahn, who helped design the graphics.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

ACM US Public Policy Council (2017). Statement on algorithmic transparency and accountability. *Commun. ACM*. Available at: https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf Accessed Jully 27, 2022.

Adams, R. (2018). The illusion of transparency: Neoliberalism, depoliticisation and information as commodity. *SSRN J.* doi:10.2139/ssrn.3281074

Ananny, M., and Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Soc.* 20 (3), 973–989. doi:10.1177/1461444816676645

Arik, S. O., and Pfister, T. (2019). TabNet: Attentive interpretable tabular learning. Available at: https://arxiv.org/abs/1908.07442 (Accessed May 27, 2022).

Arrietta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58, 82–115. doi:10.1016/j.inffus.2019.12.012

Artiga, S., and Hinton, E. (2018). Beyond health care: The role of social determinants in promoting health and health equity. Available at: https://www.kff.org/racial-equity-and-health-policy/issue-brief/beyond-health-care-the-role-of-social-determinants-in-promoting-health-and-health-equity/ (Accessed May 30, 2022).

Aubreville, M., Goncalves, M., Knipfer, C., Oetter, N., Würfl, T., Neumann, H., et al. (2019). in *Transferability of deep learning algorithms for malignancy detection in confocal laser endomicroscopy images from different anatomical locations of the upper gastrointestinal tract" in biomedical engineering systems and technologies*. Editors A. Cliquet, S. Wiebe, P. Anderson, G. Saggio, R. Zwiggelaar, H. Gamboa, et al. (Cham: Springer International Publishing), 67–85.

Bowker, G. C., and Star, S. L. (2000). *Sorting things out. Classification and its consequences*. Massachusetts: MIT Press.

Braun, M., and Hummel, P. (2022). Data justice and data solidarity. *Patterns* 3 (3), 1–8. doi:10.1016/j.patter.2021.100427

Brindaalakshmi, K. (2021). A New AI Lexicon: Gender. Transgender erasure in AI: Binary gender data redefining 'gender' in data systems. Available at: https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-gender-b36573e87bdc (Accessed June 4, 2022).

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Soc.* 3 (1), 205395171562251. doi:10.1177/2053951715622512

Butler, J. (2009). *Frames of war*. London, New York: Verso.

Butler, J. (2007). *Gender trouble*. New York: Routledge.

Butler, J. (2004). *Undoing gender*. New York: Routledge.

Calderon, A., Taber, D., Qu, H., and Wen, J. (2019). *AI blindspot*. Cambridge: Cambridge University Press.

Carroll, S. R., Akee, R., Chung, P., Cormack, D., Kukutai, T., Lovett, R., et al. (2021). Indigenous peoples' data during COVID-19: From external to internal. *Front. Sociol.* 6, 617895. doi:10.3389/fsoc.2021.617895

Carroll, S. R., Rodriguez-Lonebear, D., and Martinez, A. (2019). Indigenous data governance: Strategies from United States native Nations. *Data Sci. J.* 18 (31), 31–15. doi:10.5334/dsj-2019-031

Casper, M. J., and Moore, L. J. (2009). *Missing bodies: The politics of visibility*. New York: New York University Press.

Chesterman, S. (2021). *We, the robots? Regulating artificial intelligence and the limits of law*. Cambridge: Cambridge University Press.

Cho, M. K. (2021). Rising to the challenge of bias in health care AI. *Nat. Med.* 27 (12), 2079–2081. doi:10.1038/s41591-021-01577-2

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 139–167.

Criado-Perez, C. (2019). *Invisible women. Data bias in a world designed for men*. New York: Abrams Press.

Dankwa-Mullan, I., Zhang, X., Le, P., and Riley, W. T. (2021). "Applications of big data science and analytic techniques for health disparities research," in *The science of health disparities research*. Editors I. Dankwa-Mullan, E. J. Pérez-Stable, K. L. Gardner, X. Zhang, and A. M. Rosario (New York: Wiley), 221–242.

David, M. (2015). The correspondence theory of truth. Available at: https://plato.stanford.edu/entries/truth-correspondence/ (Accessed March 18, 2022).

Diakopoulos, N. (2020). "Transparency," in *The oxford handbook of ethics of AI*. Editors M. D. Dubber, F. Pasquale, S. Das, and N. Diakopoulos (. New York: Oxford University Press), 196–213.

Edwards, L., and Veale, M. (2017). Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for. *Duke Law Technol. Rev.* 16, 18–84. 0.31228/osf.io/97upg

Federal Anti-Discrimination Agency (2019). *Guide to the general equal treatment Act*. Berlin: Explanations and Examples.

Felzmann, H., Villaronga, E. F., Lutz, C., and Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Soc.* 6 (1), 205395171986054. doi:10.1177/2053951719860542

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. C., and Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI*. Cambridge: Berkman Klein Center for Internet and Society.

Floridi, L. (2017). Infraethics-on the conditions of possibility of morality. *Philos. Technol.* 30 (30), 391–394. doi:10.1007/s13347-017-0291-1

Foucault, M. (1972). *The archaeology of knowledge: And the discourse on language*. New York: Pantheon Books.

Gitelman, L. (2013). *Raw data is an oxymoron*. Cambridge: MIT Press.

Goswami, S. (2020). "Deep learning – a state-of-the-art approach to artificial intelligence," in *Deep learning: Research and applications*. Editors S. Bhattacharyya, V. Snasel, A. E. Hassanien, S. Saha, and B. K. Tripathy (Berlin, Boston: Walter de Gruyter), 1–19.

Hague, D. C. (2019). Benefits, Pitfalls, and Potential Bias in Health Care AI. *North Carol. Med. J.* 80, 219–223. doi:10.18043/ncm.80.4.219

Halsema, A. (2005). "Reflexionen über Identität in einer multikulturellen Gesellschaft: Ein Dialog zwischen Ricoeur, Irigaray und Butler," in *Feministische Phänomenologie und Hermeneutik*. Editors S. Stoller, V. Vasterling, and L. Fisher (Würzburg), 208–234.

Hasselbalch, G. (2021). *Data ethics of power. A human approach in the big data and AI era*. Massachusetts: Edward Elgar Publishing Inc.

Hatzenbuehler, M. L., and Pachankis, J. E. (2021). "Sexual and gender minority health disparities: Concepts, methods, and future directions," in *The science of health disparities research*. Editors I. Dankwa-Mullan (New York: Wiley), 429–444.

High-Level Expert Group on Artificial Intelligence (2019). *Ethics guidelines for trustworthy AI*. Brüssel: European Commission..

Honneth, A. (2003). *Unsichtbarkeit: Stationen einer Theorie der Intersubjektivität*. Frankfurt am Main: Suhrkamp.

Hummel, P., Braun, M., Augsberg, S., von Ulmenstein, U., and Dabrock, P. (2021a). *Datensouveränität: Governance-Ansätze für den Gesundheitsbereich*. Wiesbaden: Springer.

Huston, P., Edge, V. L., and Bernier, E. (2019). Reaping the benefits of Open Data in public health. *In: Canada Commun. Dis. Rep.* 45, 252–256. doi:10.14745/ccdr.v45i10a01

Hummel, P., Braun, M., Tretter, M., and Dabrock, P. (2021b). Data sovereignty: A review. *Big Data & Soc.* 8, 205395172098201. doi:10.1177/2053951720982012

Hwang, B., You, J., Vaessen, T., Myin-Germeys, I., Park, C., and Zhang, B.-T. (2018). Deep ECGNet: An optimal deep learning framework for monitoring mental stress using ultra short-term ECG signals. *Telemed. J. E. Health.* 24 (10), 753–772. doi:10.1089/tmj.2017.0250

Iqbal, S., Tariq, M., Ayesha, H., and Ayesha, N. (2021). "AI technologies in healthcare applications," in *Artificial intelligence and internet of things. Applications in smart healthcare*. Editor L. M. Goyal (London: CRC Press), 3–44.

Jasanoff, S. (2016). *The ethics of invention. Technology and the human future*. New York: W. W. Norton & Company.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* 1 (9), 389–399. doi:10.1038/s42256-019-0088-2

Jüngel, E. (2010). *Gott als geheimnis der Welt*. Tübingen: Mohr Siebeck.

Kashif, M., Rehman, A., Sadad, T., and Mehmood, Z. (2021). "Breast cancer detection and diagnostic with convolutional neural networks," in *Artificial intelligence and internet of things. Applications in smart healthcare*. Editor L. M. Goyal (London: CRC Press), 65–84.

Kilian, P. (2013). Unsichtbare Sichtbarkeit. Michel Foucault und die Transparenz. Available at: https://blog.genealogy-critique.net/essays/19/unsichtbare-sichtbarkeit (Accessed March 18, 2022).

Latour, B. (2000). *Die Hoffnung der Pandora. Untersuchungen zur Wirklichkeit der Wissenschaft. Aus dem Englischen von Gustav Roßler*. Frankfurt am Main: Suhrkamp.

Lee, E. E., Torous, J., de Choudhury, M., Depp, C. A., Graham, S. A., Kim, H.-C., et al. (2021). Artificial intelligence for mental health care: Clinical applications, barriers, facilitators, and artificial wisdom. *Biol. Psychiatry. Cogn. Neurosci. Neuroimaging* 6 (9), 856–864. doi:10.1016/j.bpsc.2021.02.001

Leslie, D., Mazumder, A., Peppin, A., Wolters, M. K., and Hagerty, A. (2021). Does "AI" stand for augmenting inequality in the era of coCovid-19ealthcare? *BMJ* 372, n304. doi:10.1136/bmj.n304

Lima, G., Grgić-Hlača, N., Jeong, J. K., and Cha, M. (2022). The Conflict Between Explainable and Accountable Decision-Making Algorithms. Available at: https://arxiv.org/abs/2205.05306 (Accessed June 3, 2022).

Lu, X. H., Mamiya, H., Vybihal, J., Ma, Y., and Buckeridge, D. (2020). "Guiding public health policy by using grocery transaction data to predict demand for unhealthy beverages," in *Explainable AI in healthcare and medicine building a culture of transparency and accountability*. Editor A. Shaban-Nejad (New York: Springer), 169–176.

Lundberg, S., and Lee, S-I. (2017). A unified approach to interpreting model predictions. Available at: https://arxiv.org/abs/1705.07874 (Accessed May 31, 2022).

Lupton, D. (2015). "Donna Haraway: The digital cyborg assemblage and the new digital health technologies," in *The palgrave handbook of social theory in health, illness and medicine*. Editor F. Collyer (New York: Springer), 567–581.

Mann, M., and Matzner, T. (2019). Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination. *Big Data & Soc.* 6, 205395171989580. doi:10.1177/2053951719895805

Maschewski, F., and Nosthoff, A.-V. (2021). Überwachungskapitalistische Biopolitik: Big Tech und die Regierung der Körper. *Z. für Politikwiss.* 32 doi:10.1007/s41358-021-00309-9

Mayer-Schönberger, V. (2009). *Delete: The virtue of forgetting in the digital age*. Oxford: Princeton University Press.

Mentzel, H.-J. (2021). [Artificial intelligence in image evaluation and diagnosis]. *Monatsschr. Kinderheilkd.* 169 (8), 694–704. doi:10.1007/s00112-021-01230-9

Meunier, A., Gray, J., and Ricci, D. (2021). A new AI lexicon: Algorithm trouble. Troublesome encounters with algorithms that go beyond computational processes. Available at: https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-algorithm-trouble-50312d985216 (Accessed June 5, 2022).

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society* 1–21. doi:10.1177/2053951716679679

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., and Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns* 2 (10), 100347–100349. doi:10.1016/j.patter.2021.100347

Norris, C. M., Yip, C. Y. Y., Nerenberg, K. A., Clavel, M.-A., Pacheco, C., Foulds, H. J. A., et al. (2020). State of the science in women's cardiovascular disease: A Canadian perspective on the influence of sex and gender. *J. Am. Heart Assoc.* 9, e015634. doi:10.1161/JAHA.119.015634

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464447–6464453. doi:10.1126/science.aax2342

Oskooei, A., Chau, S. M., Weiss, J., Sridhar, A., Martínez, M. R., and Michel, B. (2021). "DeStress: Deep learning for unsupervised identification of mental stress in firefighters from heart-rate variability (HRV) data," in *Explainable AI in healthcare and medicine building a culture of transparency and accountability*. Editors A. Shaban-Nejad, M. Michalowski, and D. L. Buckeridge (Cham: Springer Nature), 93–105.

Pasquale, F. (2015). *Black box society*. Cambridge: Harvard University Press.

Prabhakaran, V., and Martin, D. (2020). Participatory Machine Learning Using Community-Based System Dynamics. *Health Hum. Rights* 22 (2), 71–74.

Prainsack, B. (2017). *Personalized medicine: Empowered patients in the 21st century?* New York: Wiley.

Prince, A. E. R., and Schwarcz, D. (2020). *Proxy discrimination in the age of artificial intelligence and big data*. Iowa Law Review. Available at: https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/ilr105§ion=35 (Accessed June 13, 2022).

Röösli, E., Rice, B., and Hernandez-Boussard, T. (2021). Bias at warp speed: How AI may contribute to the disparities gap in the time of COVID-19. *J. Am. Med. Inf. Assoc.* 28 (1), 190–192. doi:10.1093/jamia/ocaa210

Ruf, B., and Detyniecki, M. (2021). Towards the right kind of fairness in AI. Available at: http://arxiv.org/pdf/2102.08453v7 (Accessed March 18, 2022).

Schütz, F., Heidingsfelder, M. L., and Schraudner, M. (2019). Co-Shaping the future in quadruple helix innovation systems: Uncovering public preferences toward participatory research and innovation. *She Ji J. Des. Econ. Innovation* 5 (2), 128–146. doi:10.1016/j.sheji.2019.04.002

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., and Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* 27 (12), 2176–2182. doi:10.1038/s41591-021-01595-0

Stark, H. (2014). Judith Butler's post-Hegelian ethics and the problem with recognition. *Fem. Theory* 15 (1), 89–100. doi:10.1177/1464700113512738

Strotbaum, V., and Reiß, B. (2017). ""Apps im Gesundheitswesen – echter medizinischer Nutzen oder der Weg zum gläsernen Patienten," in *E-Health-Ökonomie*. Editors T. Müller-Mielitz and S. Lux (Wiesbaden: Springer Fachmedien Wiesbaden), 359–382.

Turek, H. (2020). *Open algorithms: Experiences from France, the Netherlands and New Zealand (Open Algorithms Blog Series)*. Available at: https://www.opengovpartnership.org/stories/open-algorithms-experiences-from-france-the-netherlands-and-new-zealand/ Accessed Jully 24, 2022

Uddin, M. Z., Dysthe, K. K., Følstad, A., and Brandtzaeg, P. B. (2022). Deep learning for prediction of depressive symptoms in a large textual dataset. *Neural comput. Appl.* 34 (1), 721–744. doi:10.1007/s00521-021-06426-4

UNESCO (2022). *Recommendation on the ethics of artificial intelligence*. Paris.

Viola, L. A., and Laidler, P. (2021). *Trust and transparency in an age of surveillance*. London: Routledge.

Véliz, C. (2020). *Privacy is power: Why and how you should take back control of your data*. London: Penguin Books.

Wachter, S., and Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. *Columbia Bus. Law Rev.* 2, 494–630. doi:10.7916/cblr.v2019i2.3424

Wachter, S. (2022). The theory of artificial immutability: Protecting algorithmic groups under anti-discrimination law. *Tulane Law Review 97*. doi:10.2139/ssrn.4099100 Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4099100 Accessed Jully 24, 2022.

Walter, M. (2018). The voice of Indigenous data. Beyond the markers of disadvantage. Available at: https://griffithreview.com/articles/voice-indigenous-databeyond- (Accessed May 31, 2022).

WHO (2021). Ethics and governance of artificial intelligence for health. Available at: https://www.who.int/publications/i/item/9789240029200 (Accessed March 18, 2022).

Zerilli, J., Bhatt, U., and Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns* 3 (4), 1–10.

Zerilli, J., Dahner, J., Maclaurin, J., Gavaghan, C., Knott, A., Liddicoat, J., et al. (2021). *A citizen's guide to artificial intelligence*. Cambridge: MIT Press.

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 19 September 2022
DOI 10.3389/fgene.2022.903600

Check for updates

OPEN ACCESS

EDITED BY
Silke Schicktanz,
University of Göttingen, Germany

REVIEWED BY
Lorina Buhr,
University of Erfurt, Germany
Joschka Haltaufderheide,
Ruhr University Bochum, Germany

*CORRESPONDENCE
Robin L. Pierce,
r.p.pierce@exeter.ac.uk,
pierce7@post.harvard.edu

SPECIALTY SECTION
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Genetics

RECEIVED 24 March 2022
ACCEPTED 19 August 2022
PUBLISHED 19 September 2022

CITATION
Pierce RL, Van Biesen W, Van Cauwenberge D, Decruyenaere J and Sterckx S (2022), Explainability in medicine in an era of AI-based clinical decision support systems. *Front. Genet.* 13:903600. doi: 10.3389/fgene.2022.903600

COPYRIGHT
© 2022 Pierce, Van Biesen, Van Cauwenberge, Decruyenaere and Sterckx. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Explainability in medicine in an era of AI-based clinical decision support systems

Robin L. Pierce[1]*, Wim Van Biesen[2,3], Daan Van Cauwenberge[4,5], Johan Decruyenaere[3,4] and Sigrid Sterckx[4,5]

[1]The Law School, University of Exeter, Exeter, United Kingdom, [2]Head of Department of Nephrology and Centre for Justifiable Digital Healthcare, Ghent University Hospital, Ghent, Belgium, [3]Centre for Justifiable Digital Healthcare, Ghent University Hospital, Ghent, Belgium, [4]Department of Intensive Care Medicine and Centre for Justifiable Digital Healthcare, Ghent University Hospital, Ghent, Belgium, [5]Department of Philosophy and Moral Sciences, Bioethics Institute Ghent, Ghent University, Ghent, Belgium

The combination of "Big Data" and Artificial Intelligence (AI) is frequently promoted as having the potential to deliver valuable health benefits when applied to medical decision-making. However, the responsible adoption of AI-based clinical decision support systems faces several challenges at both the individual and societal level. One of the features that has given rise to particular concern is the issue of explainability, since, if the way an algorithm arrived at a particular output is not known (or knowable) to a physician, this may lead to multiple challenges, including an inability to evaluate the merits of the output. This "opacity" problem has led to questions about whether physicians are justified in relying on the algorithmic output, with some scholars insisting on the centrality of explainability, while others see no reason to require of AI that which is not required of physicians. We consider that there is merit in both views but find that greater nuance is necessary in order to elucidate the underlying function of explainability in clinical practice and, therefore, its relevance in the context of AI for clinical use. In this paper, we explore explainability by examining what it requires in clinical medicine and draw a distinction between the function of explainability for the current patient versus the future patient. This distinction has implications for what explainability requires in the short and long term. We highlight the role of transparency in explainability, and identify semantic transparency as fundamental to the issue of explainability itself. We argue that, in day-to-day clinical practice, accuracy is sufficient as an "epistemic warrant" for clinical decision-making, and that the most compelling reason for requiring explainability in the sense of scientific or causal explanation is the potential for improving future care by building a more robust model of the world. We identify the goal of clinical decision-making as being to deliver the best possible outcome as often as possible, and find—that accuracy is sufficient justification for intervention for today's patient, as long as efforts to uncover scientific explanations continue to improve healthcare for future patients.

KEYWORDS

transparency, semantic transparency, artificial intelligence in medicine, clinical decision support, causality, explainability

1 Introduction

The combination of "Big Data" and Artificial Intelligence (AI) is frequently promoted as being likely to offer health benefits when applied to medical decision-making (e.g., Fogel and Kvedar 2018; Topol 2019). However, many have rightly observed that AI does not automatically transform data into improved health outcomes (e.g., Beam and Kohane, 2018; Emanuel and Wachter 2019). This technology comes with associated risks, not only at the societal level, but also at the levels of individual patient health and physician responsibility and liability. Moreover, the possibilities for bias, for example, because of a limited appreciation of the clinical context and unintended consequences, for example de-skilling, abound (Cabitza et al., 2017).

One of the features of AI that has garnered considerable attention is the issue of *explainable* AI (London 2019; Lauritsen et al., 2020; Duran and Jongsma 2021; Markus et al., 2021). For many, a basic concern is that if the way an algorithm arrives at a particular output is not known (or knowable) by the physician, this lack of explainability may have an impact on the ability to assess the appropriateness and merits of an output designed to inform treatment or diagnosis. As a consequence, this may also jeopardize the quality of the actual medical decision, as well as the shared decision-making process with the patient. There is however still no consensus on the meaning of "explainability" in the context of AI for clinical decision support systems (CDSS), and even less agreement on what kind of "explainability" is required to adequately address such considerations and for responsible adoption of CDSS (e.g., Adadi and Berrada 2018; Payrovnaziri et al., 2020). In general terms, advocates of the central role of explainability in AI base their view on some version of the argument that "certain actions are morally unjustified given the lack (of) the *epistemic warrants* required for the action to take place," and in the particular context of clinical medicine this implies that "physicians require their beliefs to be epistemically justified before acting," hence "(a) physician is not morally justified in giving a certain treatment to a patient unless the physician has *reliable knowledge* that the treatment is likely to benefit the patient" (Duran and Jongsma 2021: 331-332, emphasis added)[1]. However, the question of what constitutes "reliable knowledge," both conceptually and procedurally, such that it provides epistemic justification, remains elusive. If we understand "reliable knowledge," as used in this context, to refer to a sufficient basis for making an ethically defensible decision in the clinical context, then the term should also point out what should be required of explainability in the use of AI. The relevance of the debate on opacity versus transparency for regulators is also clear from the recent *Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence* (European Commission 2021).

Many commentators (e.g., Rudin, 2021; Van Calster et al., 2019; Shortliffe and Sepulveda 2018) worry that the opaque nature of the decision-making of many AI systems implies that, in the specific context of clinical medicine, physicians and patients cannot and should not rely on the results of such systems. In contrast, some strongly oppose the central role of explainability in AI. These commentators argue that there is no reason to require of AI that which is not required of physicians, and emphasize that a lack of explainability does not necessarily hinder a responsible and effective practice of medicine. For example, philosopher Alex London (2019: 17) claims that, much of the time, physicians cannot explain why they are doing things the way they do, and that their interventions are thus also opaque: "(Medicine's) knowledge of underlying causal systems is in its infancy . . . Medicine is a domain in which the ability to intervene effectively in the world by exploiting particular causal relationships often derives from experience and precedes our ability to understand why interventions work." Veliz et al. (2021) also note that many ill-understood processes have been adopted in medicine. One example that is frequently mentioned is the use of aspirin. Physicians did not exactly know how it works, but they knew that, for certain maladies, it did work and reliably so. However, Veliz et al. (2021) rightly point out that we need to investigate the differences and similarities between opaque algorithms and medical treatments whose workings are opaque: "For starters, the mechanism of aspirin is constant over time, but many black-box algorithms change as they get new information. Furthermore, how aspirin works is a natural fact; how algorithms work depends on us." (Veliz et al., 2021: 340).

Whether it is appropriate to expect "more" explainability from medical AI systems than from physicians is a complicated matter. In London's (2019) view, put simply, it may be unnecessary to expect explainability from medical AI, since accuracy may well be enough in many cases, even if the "why" or "how" cannot be explained or understood. This point is powerfully illustrated by the fact that the consumption of citrus fruits by sailors to prevent scurvy probably saved thousands of lives, as demonstrated in the first ever RCT, despite the fact that it was then unknown how and why it worked[2].

Thus, for some, adequately addressing the feature of "opaqueness" appears to be central to identifying what would

1 There are also other reasons why some have argued we need explainable AI. It is argued, for instance, that explainable AI to avoid widespread discrimination by AI (Gerke, 2021). Although we are aware of these arguments, in this paper we will focus on the need arising from the black box character of CDSS. Nevertheless, we believe the criticism we offer Section 2 of our paper is also relevant to these arguments.

2 See https://www.bbvaopenmind.com/en/science/leading-figures/james-lind-and-scurvy-the-first-clinical-trial-in-history/ [last accessed 8 August 2022].

constitute responsible use of AI in medicine, whereas, for others, it serves little practical purpose.

We see merit in both positions, but in our paper we seek to show that greater nuance is needed in order to get at the underlying function of explainability from the point of view of clinical practice. We aim to contribute to answering the questions which phenomena can be understood as explainability in the context of AI for clinical decision support systems and what kind of "explainability" is required for responsible adoption of such CDSS. While such an analysis provides some insight into *what may be required*, the question begs clarification about the actual value and, hence, *importance* of explainability. To this end, we will explore some key criteria identified in the literature and evaluate whether they are indeed necessary conditions of explainability in a clinical context. First, we will explore whether the concept of transparency furthers explainability. Next, we take up the question of whether accuracy and performance of the AI provide an acceptable form of explanation, and whether this would be sufficient to claim that the AI device is explainable in the sense that it provides a necessary epistemic justification for responsible use in a clinical context. Finally, based on our inquiry we will evaluate whether CDSS currently are able to meet these criteria. Thus, our inquiry into explainability in this paper is focused on the extent to which explainability should be required for AI systems intended for supporting clinical decision-making by physicians and, if so, how this concept of explainability should be understood[3].

Clinical decision support (CDS) can be defined as a process that "provides clinicians, staff, patients, or other individuals with knowledge and person-specific information, intelligently filtered or presented at appropriate times, to enhance health and health care" (Osheroff et al., 2007: 141)[4]. As noted by Musen et al. (2014):

> "Systems that provide CDS do not simply assist with the retrieval of relevant information; they communicate information that takes into consideration the particular clinical context, offering situation-specific information and recommendations. At the same time, such systems do not themselves perform clinical decision making; they provide relevant knowledge and analyses that enable the ultimate decision makers—clinicians, patients, and health care organizations—to develop more informed judgments . . . Systems that provide CDS come in three basic varieties: 1) They may use information about the current clinical context to retrieve highly relevant online documents, as with so-called "infobuttons" . . .; 2) they may provide patient-specific, situation-specific alerts, reminders, physician order sets, or other recommendations for direct action; or 3) they may organize and present information in a way that facilitates problem solving and decision making, as in dashboards, graphical displays, documentation templates, structured reports, and order sets" (Musen et al., 2014: 643-644).

The paper proceeds with Section 2, in which we discuss key conceptual issues necessary to clarify the debate on explainability. We particularly focus on the differences between "explainability" and "transparency," and highlight the crucial importance of *semantic* transparency, as a particular form of transparency that is essential to responsible use of CDSS. Semantic transparency yields a type of explainability that is frequently necessary for accuracy. The clinical case of Acute Kidney Injury (AKI) is provided as an illustration of the importance of this semantic transparency.

In Section 3, we discuss the reasons for why explainability matters in clinical medicine and thus why we need explainability in CDSS. Here, we will build on philosopher of science and technology Duran's (2021) argumentation for the importance of *scientific* explanation for clinical uses of AI and illustrate this with the example of a prediction model for AKI. Based on this analysis, we argue that the need for scientific or causal explainability in *clinical practice* is limited and that a nuanced approach that engages with the function (and relative importance) of explainability is necessary in order to identify what should be required of medical AI. We argue that, in daily clinical practice, it is sufficient most of the time to have an explanation that provides enough justification to (not) do something, but that, in order to improve accuracy in the longer term, increasing understanding of underlying causality is required[5].

Section 4 then focuses on this topic of causal understanding, identifying the key question of whether the Big Data approaches that typically underpin modern CDSS can answer questions pertaining to causality (counterfactual or "why" questions). We provide a brief overview of the intense debate on this question, highlighting philosopher of science and technology

3 We will not be looking at explainability concerning AI in a non-medical context. We hypothesize, however, that several differences exist between explainability of AI in the clinical-medical setting as opposed to other settings. First, the decision of an AI in a medical context cannot be evaluated by merely creating *transparency* regarding the factors that drove the decision, because the relationship between input and output is less clear in a medical context. In other domains there is either a clear and established relation between input and output, or, when such a relation is unknown, it can easily be tested whether or not the use of the algorithm would consistently lead to the desired result even if all other factors were to be varied. Second, medical decisions inevitably reflect values. Even seemingly objective decisions, for example, the thresholds selected to steer medical decisions, in reality reflect one or several values. In this paper, however, we will not focus on these differences.

4 See also Berner and La Lande (2016).

5 In recent years there has been an interesting debate concerning the status and nature of causality within the field of medicine (Kincaid, 2008; Thompson, 2011; Illari & Russo, 2014). In this paper we will focus on the question whether CDSS is able to meet the demands of clinical practice, assuming that the modeling of causal relations plays at least *some* role in clinical practice.

Pietsch's (2021) epistemological analysis of Big Data. Pietsch's (2021) argues that causal understanding is crucial for reliable predictions as well as for effective interventions. We add nuance to his argument on two points: first, that when the accuracy of predictive algorithms operates in the place of explainability, there is no real need for an underlying causal relationship between the data and the outcome; and second, that relying on "variational evidence" allows one to infer a *causal* relationship between a phenomenon and its circumstances. We do not subscribe to the latter—i.e., the claim that causality can be obtained with Big Data approaches relying on machine learning—because, as in many other real-world problems, in a clinical context it is almost never certain that Big Data are complete and representative of all conditions, hence the conditions that would allow for the use of variational induction are simply never present.

In Section 5, we conclude with a summary of our core findings regarding what explainability requires for responsible clinical decision-making.

2 "Explainability" and "transparency": The importance of semantic transparency

The relationship between "explainability" and "transparency" is neither obvious nor clear-cut. Providing one does not necessarily ensure the other. If explainability could be substituted by transparency, then the requirements for explainable AI would be simplified considerably. However, like many commentators, we hold the view that transparency is of limited value as a surrogate for explainability. (Markus et al., 2021: 3). Nevertheless, we will identify a particular type of transparency, semantic transparency, as fundamental to explainability. This, in turn, informs our argument about the nature of the explainability that may ultimately be required.

According to Duran and Jongsma (2021), the concept of "transparency" refers to "algorithmic procedures that make the inner workings of a black box algorithm interpretable to humans" (Duran and Jongsma 2021: 330). In contrast with "transparency," "opacity" refers to the "inherent impossibility of humans to survey an algorithm, both understood as a script as well as a computer process" (ibid.)[6].

Duran and Jongsma (2021) give a clear and helpful explanation of why *transparency*, i.e., providing exogenous algorithms capable of making visible the variables and relations within the black box that are responsible for the outcome, although it can help foster trust in algorithms and their outcomes, but does not answer (all) the problems posed by opacity, as it instead *shifts the question of opacity of the black box algorithm to the question of opacity of those* exogenous algorithms.

According to Duran and Jongsma (2021), those defending the view that "epistemic opacity" is inevitable argue that this is due to the fact that humans are limited cognitive agents and that therefore we should abandon the goal of achieving transparency as a means of cultivating trust in algorithms[7]. Duran and Jongsma (2021), by contrast, argue that "giving up explanation altogether (or reducing explanation to a handful of alleged transparent algorithms) defeats much of the purpose of implementing AI in medical practice" because the predicted improvements in efficiency and accuracy would be nullified by the loss of trustworthiness in the process if explainability were to be given up or reduced to transparency (Duran and Jongsma 2021: 331).

We agree with this point of view yet we would like to make a different contribution to this debate. First, we believe that transparency consists of different "parts" or elements and that a specific part of transparency is fundamental to explainability[8]. More precisely, we argue that *semantic transparency* may address a significant aspect of the problem of opacity. An absence of opacity not only presupposes transparency at the level of *how* symbols and data are *handled* by the AI device, but it necessitates that *exactly what* those symbols and data *represent* be clear and transparent. Therefore, by semantic transparency we refer to the clear and unambiguous usage of terms handled by the CDSS. This forms a crucial element of semantic transparency, the absence of which may serve to undermine any subsequent efforts to provide other forms of transparency, and undermines accuracy, which we argue can provide justification for responsible use of CDSS.

As we explain further in this Section, if the *terminology* used to classify the information that trained an algorithm is unclear, conflated, or insufficiently precise, it will be impossible to obtain

6 Berkeley sociologist and computer scientist Jenna Burrell makes an important distinction between three forms of opacity: opacity as intentional corporate or state secrecy (in order to maintain a competitive advantage); opacity as technical illiteracy (because code writing and reading are specialist skills); and opacity that arises from the characteristics of machine learning algorithms, more specifically from "the mismatch between mathematical procedures of machine learning algorithms and human styles of semantic interpretation" (Burrell 2016: 3). Even though all three forms can be relevant for the context of clinical medicine, we will only be concerned with the third form.

7 According to Duran and Jongsma, this need not worry us too much because the outcomes of medical AI *can* be trustworthy and black box algorithms *can* be reliable, provided that certain epistemic conditions are met, viz. the conditions entailed by the framework of "computational reliabilism" that they propose (Duran and Jongsma 2021: 332; Duran and Formanek 2018). Although they argue that this framework, which does not require transparency, provides "reasonable levels of confidence about the results of opaque algorithms," this claim does not imply that opaque algorithms should be used without any restrictions, as the appropriateness of their use in the context of medicine depends on many factors that are related to ethics rather than epistemology (Duran and Jongsma 2021: 330).

8 By "fundamental" we mean that although semantic transparency is not a sufficient condition of explainability, it is a necessary one.

transparency at any later stage. Given the foundational nature of the classification of training data, semantic opacity arising from imprecise or conflated terminology at this stage would be extremely difficult, if not impossible, to untangle at a later stage for the purposes of transparency. Therefore, transparency is necessary at this semantic level. In terms of clinical implications, a failure to incorporate semantic transparency can affect both: 1) the ability to understand how an output should be translated into action (i.e., what clinical intervention is advisable); and 2) the degree of accuracy (within a generally accurate range) that can be achieved with the output (i.e., how narrow the range of reliable accuracy is). In this way, semantic transparency is an essential element of reliability needed to support the responsible use of CDSS, for otherwise the actual inner workings of the recommendation system will remain largely unknown to the physician regardless of subsequent reductions in opacity. Therefore, we argue that semantic transparency should be a non-negotiable requirement for transparency in the context of using AI for CDSS, because a lack of this foundational transparency could ultimately undermine the principal value of using an AI device at all.

Unfortunately, the importance of semantic transparency with regard to terminology of both input and output parameters and concepts, is often neglected. If the same input or output symbol within the algorithm can represent different items, or different interpretations of an item, it becomes unclear what exactly is being handled by the algorithm, and different users (who explain the working of the decision to themselves) may have different interpretations of what has been done and what the result is. As philosopher of science Wolfgang Pietsch rightly notes, one of the essential conditions for achieving successful prediction based on data is that "the vocabulary is well chosen, meaning that the parameters are stable causal categories" (Pietsch 2015: 910). Transparency at the semantic level means that the definitions and their operationalization in the algorithm should be transparent (i.e., clear and unequivocal at the semantic level). Pietsch's requirement of "stable causal categories" refers to the fact that the definition of these parameters should be stable over time, and thus fixed and unchanging, so that any deviations from this requirement over time can be detected.

However, lack of basic semantic transparency is a widespread problem in decision support systems used in clinical medicine. We can illustrate this with an example from the field of nephrology. Acute Kidney Injury (AKI) is a clinical concept indicating that the kidneys are damaged and will rapidly decline in function. Depending on the definition used, this decline can range from rather benign to a complete loss of function, resulting in the accumulation of water and toxins, potentially leading to the death of the patient. Kidney function can to some extent be replaced by extracorporeal renal replacement therapy (RRT). While RRT can be lifesaving, it is invasive and can have life-threatening side-effects such as bleeding, severe electrolyte disorders or low blood pressure. To date, there is no curative treatment for AKI, so there is a lot of focus on algorithm-based automated prediction and early detection in order to avoid progression to AKI.

The correct evaluation and implementation of such algorithms, however, is hampered by an absence of semantic transparency in the use of *many different definitions* of AKI. A review of algorithm-based prediction models for AKI by Van Acker et al. (2021) found that 44 different definitions were used for AKI. Most of these prediction models claim to predict AKI as defined by the widely accepted Kidney Disease Improving Global Outcomes (KDIGO) initiative (Fliser et al., 2012). However, in reality they use *different interpretations* of this definition, which may even substantially differ from the original KDIGO definition. For example, most interpretations neglect the criterion of urinary output, although this is the most powerful prognosticator in the KDIGO definition (Van Acker et al., 2021). As a consequence, the end user cannot truly know exactly *what* is understood by the algorithm-predicted condition labelled as AKI, and how that label should be coupled to possible interventions. *Transparency of prediction algorithms* for AKI requires that we can know *precisely* which definition of AKI was used, and, as a result, understand the implications for intervention. Transparency on the precise definition of AKI used in the algorithm requires in-depth detail not only regarding the definition itself, but also on the exact *operationalization* of that definition into computer language. Indeed, even when the KDIGO definition is correctly used, differences in operationalization might result in differences in the incidence and prognostic value of the label AKI. For example, "patient weight" could be the real, measured weight of the patient, an estimated weight, or an ideal weight for a person of that age and gender and "during 12 h" can be interpreted as "in every hour for 12 consecutive hours" or "over a 12-h period <6 ml/kg." All these differences in operationalization have a substantial impact on the meaning of the label "AKI" that is provided as an interpretation of the data by the algorithm.

Studies on *interventions* for AKI yield different and contradictory results. This problematic finding is most likely related to the fact that, as mentioned above, different and frequently unspecified definitions are used for AKI, such that, in reality, *different* conditions are being investigated in those intervention studies. Similar instances of conflation, imprecision, and opacity can be found in other fields of medicine, as well (see e.g., Steyaert et al., 2019).

3 The importance of explainability in medicine: Accuracy of the recommendation and scientific explanation

It is necessary to engage with various *normative* issues in order to address the following important questions with regard to

the specific context of clinical medicine: When and why does explainability matter in medicine? What kind of explainability is necessary in order to reach a responsible use of AI?

Some commentators insist that an elaboration of the *need* for explanation is necessary, and that *the reasons identified for demanding explainability determine what is required to achieve it and what is meant by the term* (Markus et al., 2021: 4). According to Markus et al. (2021), "Given that clinical practice presents a range of circumstances that have different needs regarding explainability, different properties of explainability can be traded off depending on the reason explainability is needed" (Markus et al., 2021: 7). They distinguish three reasons why explainability can be useful (Markus et al., 2021: 4):

(1) "to assist in verifying (or improving) other model desiderata" (e.g., fairness, legality, and ethicality);
(2) to manage social interaction ("to create a shared meaning of the decision-making process" and to justify decisions towards colleagues and patients); and
(3) to discover new insights to guide future research.

They argue that clustering explainable AI (XAI) systems on the basis of need guides determinations about explainability given that need informs "the relative importance of the properties of explainability and thus influences the design choice of explainable AI systems" (Markus et al., 2021: 4).

Adadi and Berrada (2018) identify four motivations for explainability:

(1) to justify decisions;
(2) to enable user control;
(3) to improve models; and
(4) to gain new insights.

It is noteworthy that the lists of motivations for explainability offered by Markus and others and by Adadi and Berrada (2018) both include *the need to garner "new insights."* As we explain below, this need may be a compelling motivation for explainability in CDSS.

In Pietsch's (visual) representation of the notion of "data" in his book *Big Data* (Pietsch 2021: 11) (Figure 1), he conveys the epistemological importance of data and summarizes the most important (epistemological) aspects of "data" as follows:

> "Data are marks on a physical medium ... that are meaningfully (i.e., causally or definitionally) related with certain singular facts belonging to a phenomenon of interest. If the data are correctly interpreted in terms of the relationship that they have with those facts, then the data constitute evidence for those facts and thus the phenomenon of interest" (Pietsch 2021: 12).

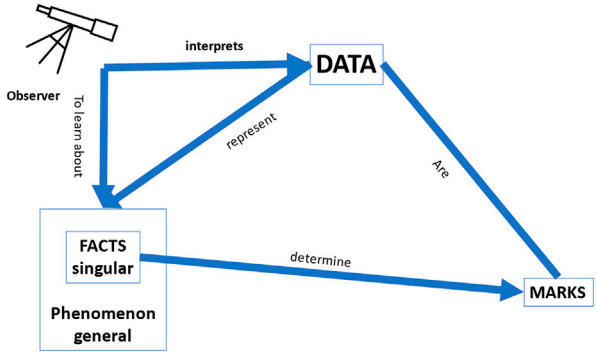


FIGURE 1
Pietsch's representation of data (adapted from Pietsch (2021): 11).

Bearing in mind this general framework of any data that may provide knowledge about the world, we would like to focus on the specific context of *clinical medicine* where data may provide knowledge about health and disease. Imagine the situation of a clinician in a busy Intensive Care Unit, where an (AI or other) observer gives the clinician her interpretation of the available data. The clinician most likely will only be interested in the *accuracy* of how well this interpretation represents the facts and the phenomenon, and that *will suffice as an "explanation"* to justify the acceptance of an advice (see Figure 2 below). As such, *the accuracy of the recommendation provides an explanation in line with one of the needs* identified by Adadi and Berrada (2018), as mentioned earlier in this Section, viz. the need of "justifying decisions": the recommendation of the AI-based CDSS is justifiable because it is deemed to be sufficiently accurate. We should point out that even if the framework in which these recommendations was based would turn out to be "wrong" or misguided, the physician would still be justified in using the CDSS if it would be more accurate than any other tool available to them. It is also important to observe that, as can be seen in the figure, there is no need for a causal relation between the data and the recommendation, as long as the accuracy of the CDSS is better than that of any other tool.

Furthermore, the same would be the case for *a patient who was being informed by her physician/nurse about possible interventions*: the patient would like to understand *how* this physician or nurse has linked her data to other data (in other words, how her data were classified) in order to draw the proposed conclusions. For example, to link the data, the physician might have relied on information from an RCT showing that patients with the same condition have the highest probability of having outcome X if they do Y rather than Z. This justification can thus also assist shared decision making, and so corresponds with other reasons why explainability can be needed, e.g., "to manage social

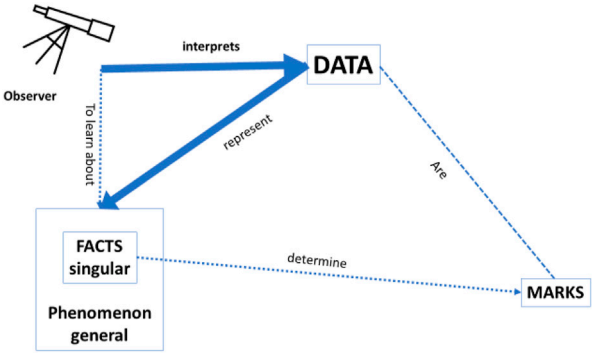


FIGURE 2
Accuracy: how well the data interpretation (i.e., classification/prediction) by the observer represents the facts/phenomenon. In this setting, it is irrelevant whether the data are causally related to the facts/phenomenon and there is no need for a predefined model.

interaction and to create a shared meaning of the decision-making process," "to justify decisions towards colleagues and patients," and "to enable user control" [cf. the lists above of Markus et al. and of Adadi and Berrada (2018)].

Accordingly, we would argue that in clinical decision support systems, knowledge on the *accuracy of the recommendation is an essential part of explainability*, especially for the clinical practitioner and the patient confronted with an acute task of decision making. By accuracy of the recommendation we mean that in daily clinical practice, it is sufficient most of the time to have *an explanation that provides enough justification to (not) do something*. However, *accuracy by itself is insufficient to satisfy all the needs-based criteria*. It provides a justification for using the CDSS in the whole group of patients, but cannot identify possible fallacies of the CDSS in relation to an individual patient, as it cannot provide (causal) insights in the process. This may have serious adverse consequences in individual (exceptional) cases, or in conditions in which the model becomes unstable.

To avoid such adverse consequences, explanation should be about *why* a physician or a CDSS *classified* a patient as belonging to a particular group. Indeed, as noted by Duran, accuracy of prediction or classification does not explain the true relations between the data and the outcome. Duran (2021: 3) argues, convincingly in our view, that explanations must be distinguished from "other epistemic functions, such as predictions, classifications, and descriptions" and that "much of what today is taken to be XAI are, in fact, classifications and predictions" whereas "scientific explanations provide a particular type of valuable information, one that grows our understanding of *why* a given output is the case, rather than organizing our knowledge and possibly forecasting new cases."

Importantly, in some cases, causal models regarding underlying mechanisms might lead the physician to make wrong clinical decisions. For example, dopamine induces diuresis and, on the basis of this physiological property, it used to be administered to patients to prevent AKI; however, we now know from RCTs that the use of dopamine is associated with higher mortality and more AKI, so this practice has been abandoned. While the explanation based on physiology would lead a physician to use dopamine, the "explanation" provided by RCT data would discourage the physician from doing so. This stresses, once again, the importance of taking into account the *aim* of the explanation when defining what can be seen as explanation.

Nevertheless, incorrect predictions are more likely to be avoided (and accuracy is thus more likely to be improved) if one can rely on a *model of the world* rather than on mere associations between input and output. Errors can result, for example, from so-called *tank problems*, where the algorithm bases its recommendation on data that do not have any relevant relation to the facts they represent, but only an—often obscure—association with those facts. (Zech et al., 2018).

What matters most for the *daily practice* of clinicians is "classification." Physicians usually work by classifying a patient into a certain group. In fact, clinical guidelines are generally developed to facilitate this kind of patient classification. In this setting, it is not necessary that the data are causally linked to the outcome, as long as the final classification is accurate. However, a classification is not the same as an explanation in the sense of understanding *why* certain things happen the way they do. It is learning about the world by association, not by making a model. Nonetheless, in order to advance medicine and reduce future errors, the effort to continue seeking to understand the *why*, i.e., the causal mechanisms, is essential (see Figure 3 below). Understanding causal relations in the data might improve accuracy, as this would avoid recommendations based on non-causal correlations, a weakness that is lurking in many deep learning systems. As the medical community has a duty to provide the best care possible, it is justified to use a CDSS with an accuracy higher than that of physicians. The medical community also has a duty, moreover, to improve accuracy by trying to better understand causal relations in the data and thus improving the model, and thereby improving the accuracy of the CDSS in the future.

Scientific explanation (see Duran and Jongsma, 2021) of a prediction model for AKI would require a clarification of *whether* and *why* a given factor has a causal contribution to the development of AKI, and how much of the emergence of AKI *and* the associated mortality is *attributable* to that factor. Such an explanation is even more important since the fact that AKI is *associated* with mortality does not necessarily imply that avoiding AKI would decrease mortality. A scientific explanation would be required for understanding the process as well as for being able to develop strategies to avoid or minimise this factor. Depending on the extent to which the likelihood of

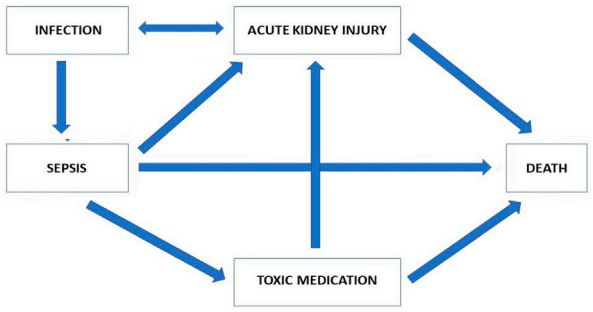


FIGURE 3
Modelling of the most important factors involved in the relationship between AKI and mortality.

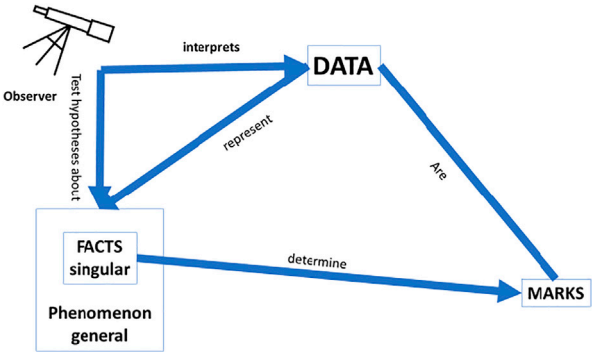


FIGURE 4
Scientific explanation: This form of explanation leads to a better understanding of the world; operationalisation of the data needs to be in line with a causal relation between the facts and the marks; a model that can be tested as a hypothesis is needed to ascertain whether or not the data fit the model; and the facts need to be causally related to the phenomenon they instantiate.

AKI and of mortality are attributable to this factor, that clarification might thus also allow for a reduction of the probability of AKI and thus mortality.

From Figure 3, we can see that AKI in patients with sepsis can be linked to mortality; however, it is equally apparent that AKI is associated with many other factors, which in themselves are associated with mortality. *Explaining* the causal pathway is an essential step to improving the outcome for these patients. Indeed, even if a golden bullet were invented that would totally prevent patients with sepsis from developing AKI, many other pathways leading to death could still exist. As long as the relative impact of the direct association between sepsis and death and the effect of AKI as a consequence of sepsis is not *explained*, it would not be possible to predict the change in mortality of sepsis with AKI by treatment with the golden bullet.

For a scientifically explainable AI (sXAI) for CDSS, one would need an interpretable predictor that helps one *understand* how the phenomenon that is determined by the facts, which are themselves potentially described by the data, causally relates to the *phenomena* (Figure 4). Such understanding could rely on knowledge regarding physiology, counterfactual experiments, etc. Such understanding could enhance and improve the classifications made by physicians in future cases, for it would also allow for generalization of the current data and situations to cases outside the current dataset, precisely because a correct understanding of the *why* would then be available.

In daily life a sufficient explanation to a physician is *an explanation that gives her enough justification to do or not do something*. One simply could not work as a physician if one sought to understand the underlying pathophysiological mechanisms all the time. A physician wants to classify a patient (intervention X will work because the patient belongs to class Y), to achieve a justifiable balance between accuracy and having time to treat all the patients who need her help. However, *in order to improve the accuracy of medical decisions in the longer term*, we need a better understanding of the phenomenon based on causal models. We will now proceed to take a closer look at this.

4 The importance of explainability in medicine: Big data and scientific explanation

Our argumentation so far highlights accuracy as an essential part of explainability of the use of CDSS, but at the same time supports a demand for greater understanding of causality as essential to the advancement of medicine, not least because increased understanding of causal factors is expected to result in increased accuracy. Therefore, we argue that prioritizing accuracy implies that one has to pay attention to causal mechanisms to ensure accuracy in the long term."

Accordingly, if, in order to ensure accuracy in the long-run, we need to be able to make models of causal mechanisms, *a central question remains of whether deep learning and Big Data approaches are helpful at all to answer "why" questions*. If the answer to this question is "no" it seems that it would be impossible for CDSS based on big data or deep learning ever to become completely 100% accurate. In what follows we use the work of Wolfgang Pietsch to try to answer this question.

As shown in Figure 2, in a deep learning approach without a predefined theoretical model, the interpretation of the data by the observer is only assessed by the accuracy of how well the data predict the facts and the phenomenon, but there is no guarantee of a *causal* relation between the data and the facts, and between the facts and the phenomenon. In the approach of Pietsch, a causal relation would be could be uncovered if all relevant factors are included in the dataset, the background

conditions are stable, and all potentially relevant combinations of factors are present. (Pietsch, 2015) In such a setting, a causal relation can be derived between the data and the phenomena by so-called *variational induction*. Pietsch stresses that the variety of evidence is crucial for variational inductivism: "confirmation . . . increases . . . with observing as many different situations in terms of changing circumstances as possible" (Pietsch 2021: 30). However, in clinical practice, one can never be certain that all relevant factors are represented in the dataset in all potentially possible combinations. Therefore, whereas the claim of causal conclusions based on variational induction is correct in theory, in practice it does not hold. This is an important nuance, all too often neglected in Big Data analysis. It explains why, for example, all the reports in the literature of successful applications of deep learning in the context of medicine—with the term "successful" referring to cases where the classification skills of the algorithm were comparable to those of experienced clinicians—concern cases where the number of data used to train the system was very high, and/or with a strongly restricted focus (e.g., to the question "diabetic retinopathy or not" in patients with diabetes, and not "what is the eye disease in this person" in the general population), precisely to ensure that all potential combinations of relevant factors can be present in the data set.

In other words, data can teach us something about the world *via* the association between the data and the facts, but only in the *specific context* of where, how and when the data were generated, and not about what would happen in a *counterfactual* world where some of the parameters are different. The fact that the relations between the data and the facts and the facts and the phenomenon are only associational does not preclude accurate predictions as long as the circumstances in which the predictions are made remain identical. However, as soon as the circumstances change (e.g., if the algorithm is used in a different hospital), the algorithm might become biased as the relation between the facts and the data in the original algorithm was not causal. If, for example, one of the data points that determined the classification by an algorithm was the type of X-ray machine used, this is of course not causally related to the type of lung disease that needs to be diagnosed. The only way to get out of this conundrum is to have a theoretical framework of the world, as this would identify which (combination) of data elements are necessary to accept that "all potentially relevant factors are included in the dataset, the background conditions are stable, and all potentially relevant combinations of factors are present," as requested by Pietsch (2015). Indeed, as noted earlier, in a clinical setting, the theoretically correct concept of variational induction allowing causal conclusion, only holds when a pre-specified model of the world is used to guarantee that all potentially relevant factors are present in the dataset.

In clinical settings big data sets never contain all the relevant data, and, given that inclusion of irrelevant data can lead to "tank problems," it is essential to *build a model* of the condition to allow for generalisation. The only way to achieve such a model is by *exploring causal relations* between the data and the observed phenomena, i.e., by scientific explanation. Therefore, to ensure the accuracy of our interventions in the long term, we have to continuously improve our theoretical models by studying causal mechanisms.

However, in the daily life of physicians, understanding such causal relations is not *per se* sufficient to select a certain intervention. There will always be a need to validate whether in reality the assumed causal relations will lead to improved outcomes, as is exemplified by the dopamine case. Scientific explanation *alone* cannot replace accuracy as a justification for using a certain intervention if the intervention is not tested in clinical trials. Therefore, the view of Pietsch needs further nuance: explanation understood as clarifying causal mechanisms and/or development of a model is necessary to improve accuracy of an existing CDSS, but is on itself insufficient to justify the use of the new CDSS. In order to justify this use, the accuracy of the improved CDSS should be better than that of human physicians, the previous version of the CDSS or other tools for decision making in the context at hand, making accuracy essential part of the explanation of why it is justified to use the CDSS.

5 Concluding remarks

The potential of AI to serve as a valuable aid in medical decision-making is significant but is still some distance away on the horizon. The acceptance and integration of AI-driven systems in everyday clinical practice depends on multiple factors. In this paper, we have focused on what kind of explainability is necessary to use CDSS responsibly in a clinical context. We identified three factors that are crucial to explainability in the context of responsible use of CDSS.

First, we identified semantic transparency, a specific type of transparency, as a critical component of transparency's contribution to explainability, and an essential element of what is required for responsible use of AI systems in the context of CDSS. Second, as some scholars have noted, the importance of explainability varies according to need. We have found that, in daily clinical practice, most of the time, accuracy should and does serve as a necessary and sufficient basis for responsible use of AI in CDSS by physicians. Third, building on Duran's (2021) case for the need for scientific explanation, we have argued that in order to improve accuracy in the longer term, and thus to reduce the incidence of interventions that negatively affect the survival and health of future patients, understanding

underlying causal mechanisms is necessary. When we understand the underlying mechanisms, we can understand why some patients respond to particular treatments and others do not. Scientific explanation is thus necessary to enhance accuracy. However, understanding a causal mechanism of a disease, a diagnostic test, or an intervention does not necessarily lead to improved outcomes when acted upon in the clinical reality. This can only be achieved with clinical trials. Therefore, scientific explanation is *in itself* insufficient to justify clinical actions.

We support the view that transparency is of limited value as a surrogate for explainability (Markus et al., 2021: 3). Nevertheless, we have identified semantic transparency as fundamental to explainability, in that *semantic transparency* may address a significant aspect of the problem of opacity. That is, given the foundational nature of the classification of training data, semantic opacity arising from imprecise or conflated terminology at this stage would be extremely difficult, if not impossible, to untangle at a later stage for the purposes of transparency. However, lack of basic semantic transparency is a widespread problem in decision support systems used in clinical medicine. For this reason, we stress this type of transparency as essential to explainability for two reasons: 1) it provides a specific type of accuracy that is necessary for the responsible use of CDSS; and 2) given that semantic transparency yields precision, it furthers the ability to derive causal explanations, which in turn, leads to increased accuracy. Understanding causal relations in the data can improve accuracy, as this would avoid recommendations based on non-causal correlations, a weakness that is lurking in many deep learning systems. However, in the daily life of physicians, understanding such causal relations is not *per se* sufficient to select a certain intervention. There will always be a need to validate (by means of Randomized Controlled Trials) whether in reality the assumed causal relations will lead to improved outcomes.

Our goal should be to create support systems for clinical decision-making that give the best possible outcome as much of the time as possible; that are as good as they can be until the *why* is understood; that actively "seek" causality; that are compatible with subsequent value-based choices; and that are open to improvement[9]. We fully concur with London's (London 2019: 20) recommendation that "regulatory practices should establish procedures that limit the use of machine learning systems to specific tasks for which their accuracy and reliability have been empirically validated."

London also rightly observes that the pathophysiology of disease is often uncertain and the mechanisms through which interventions work is frequently not known or, if known, not well understood (London 2019: 17). However, we would submit that this is a reason to strive more, rather than less, for understanding and explanation. As Aristotle observed and London has carried forward in his work, medicine is both a science and an art. We take the view that it is indeed both, but that although accuracy may be prioritized with regard to the patient in the clinic today, there are practical and pressing reasons to attend to causal knowledge in order to best serve tomorrow's patient.

9 The latter being something which may be hindered or prevented by proprietary regimes such as patents and trademarks. This is a highly important topic, but we cannot elaborate on it here.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

The lead author (RP) co-led the conceptualization, argumentation, drafting and revision of the manuscript. WvB led the scientific/technical analysis, was actively involved in the development of argumentation, and was a key contributor to the writing and revision. DvC contributed valuable research and was actively involved in the writing and revision. JD developed and critiqued key points and was actively involved in writing and revision. SS co-led the conceptualization, drafting, and revision and was actively involved in writing, revision, and argumentation. All authors were actively involved in the writing and development of this manuscript.

Funding

WvB and SS have received funding from the Research Foundation Flanders (FWO) (project number 3G068619).

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE access* 6, 52138–52160. doi:10.1109/access.2018.2870052

Beam, A., and Kohane, I. (2018). Big data and machine learning in health care. *JAMA* 319 (13), 1317–1318. doi:10.1001/jama.2017.18391

Berner, E., and La Lande, T. J. (2016). "Overview of clinical decision support systems," in *Clinical decision support systems: Theory and practice*. Editor E. Berner (Cham: Springer), 3–22.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Soc.* 3 (1), 205395171562251–12. doi:10.1177/2053951715622512

Cabitza, F., Rasoini, R., and Gensini, G. (2017). Unintended consequences of machine learning in medicine. *JAMA* 318 (6), 517–518. doi:10.1001/jama.2017.7797

Duran, J. (2021). Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare. *Artif. Intell.* 297, 103498. doi:10.1016/j.artint.2021.103498

Duran, J., and Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds Mach. (Dordr).* 28 (4), 645–666. doi:10.1007/s11023-018-9481-6

Duran, J., and Jongsma, K. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *J. Med. Ethics* 47, 329. doi:10.1136/medethics-2020-106820

Emanuel, E., and Wachter, R. (2019). Artificial intelligence in health care: Will the value match the hype? *JAMA* 321 (23), 2281–2282. doi:10.1001/jama.2019.4914

European Commission (2021). *Proposal for a regulation of the European parliament and of the Council laying down harmonised Rules on artificial intelligence*. Brussels: COM. (2021) 206 final, published 21/04/2021.

Fliser, D., Laville, M., Covic, A., Fouque, D., Vanholder, R., Juillard, L., et al. (2012). A European renal best practice (ERBP) position statement on the kidney disease improving global outcomes (KDIGO) clinical practice guidelines on acute kidney injury: Part 1: Definitions, conservative management and contrast-induced nephropathy. *Nephrol. Dial. Transpl.* 27 (12), 4263–72. doi:10.1093/ndt/gfs375

Fogel, A., and Kvedar, J. (2018). Artificial intelligence powers digital medicine. *NPJ Digit. Med.* 1, 5. doi:10.1038/s41746-017-0012-2

Gerke, S. (2021). *Health AI for good rather than evil? The need for a new regulatory framework for AI-based medical devices [SSRN scholarly paper]*. Available at: https://papers.ssrn.com/abstract=4070947.

Illari, P. M., and Russo, F. (2014). *Causality: Philosophical theory meets scientific practice*. First edition. Oxford University Press.

Kincaid, Harold (2008). Do we need theory to study disease?: Lessons from cancer research and their implications for mental illness. *Perspect. Biol. Med.* 51 (3), 367–378. doi:10.1353/pbm.0.0019

Lauritsen, S., Kristensen, M., Olsen, M., Larsen, M., Lauritsen, K., Jørgensen, M., et al. (2020). Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nat. Commun.* 11 (1), 3852–11. doi:10.1038/s41467-020-17431-x

London, A. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Cent. Rep.* 49 (1), 15–21. doi:10.1002/hast.973

Markus, A. F., Kors, J. A., and Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J. Biomed. Inf.* 113 (2021), 103655. doi:10.1016/j.jbi.2020.103655

Musen, M., Middleton, B., and Greenes, R. (2014). in *Clinical decision-support systems. Biomedical informatics*. Editors E. Shortliffe and J. Cimino (London: Springer), 643–674.

Osheroff, J. A., Teich, J. M., Middleton, B., Steen, E. B., Wright, A., and Detmer, D. E. (2007). A roadmap for national action on Clinical Decision Support. *J. Am. Med. Inf. Assoc.* 14, 141–145. doi:10.1197/jamia.M2334

Payrovnaziri, S., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J., et al. (2020). Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review. *J. Am. Med. Inf. Assoc.* 27 (7), 1173–1185. doi:10.1093/jamia/ocaa053

Pietsch, W. (2015). Aspects of theory-ladenness in data-intensive science. *Phil. Sci.* 82, 905–916.

Pietsch, W. (2021). *Big data*. Cambridge: Cambridge University Press.

Rudin, C. (2021). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1 (5), 206–215. doi:10.1038/s42256-019-0048-x

Shortliffe, E., and Sepulveda, M. (2018). Clinical decision support in the era of artificial intelligence. *J. Am. Med. Assoc.* 320 (21), 2199–2200. doi:10.1001/jama.2018.17163

Steyaert, S., Holvoet, E., Nagler, E., Malfait, S., and Van Biesen, W. (2019). Reporting of "dialysis adequacy" as an outcome in randomised trials conducted in adults on haemodialysis. *PloS one* 14 (2), e0207045. doi:10.1371/journal.pone.0207045

Thompson, R. P. (2011). "Causality, theories and medicine," in *Causality in the sciences*. Editors P. McKay Illari, F. Russo, and J. Williamson (Oxford University Press), 25–44. doi:10.1093/acprof:oso/9780199574131.003.0002

Topol, E. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* 25 (1), 44–56. doi:10.1038/s41591-018-0300-7

Van Acker, P., Van Biesen, W., Nagler, E., Koobasi, M., Veys, N., and Van Massenhove, J. (2021). Risk prediction models for acute kidney injury in adults: An overview of systematic reviews. *PLoS One* 16 (4), e0248899. doi:10.1371/journal.pone.0248899

Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, E. W., and Collins, G. S. (2019). Predictive analytics in health care: How can we know it works? *J. Am. Med. Inf. Assoc.* 26 (12), 1651–1654. doi:10.1093/jamia/ocz130

Veliz, C., Prunkl, C., Phillips-Brown, M., and Lechterman, T. M. (2021). We might be afraid of black-box algorithms. *J. Med. Ethics* 47 (5), 339–340. doi:10.1136/medethics-2021-107462

Zech, J., Badgeley, M., Liu, M., Costa, A., Titano, J., and Oermann, E. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.* 15 (11), e1002683. doi:10.1371/journal.pmed.1002683

frontiers | Frontiers in Genetics

TYPE Policy and Practice Reviews
PUBLISHED 04 October 2022
DOI 10.3389/fgene.2022.927721

Check for updates

OPEN ACCESS

EDITED BY
Karen Herrera-Ferrá,
Mexican Association of Neuroethics (AMNE), Mexico

REVIEWED BY
Rabia Saleem,
University of Derby, United Kingdom
Ivan Sammut,
University of Malta, Malta
Garbiñe Saruwatari,
Instituto Nacional de Medicina Genómica (INMEGEN), Mexico

*CORRESPONDENCE
Janos Meszaros,
janos.meszaros@kuleuven.be

SPECIALTY SECTION
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Genetics

RECEIVED 24 April 2022
ACCEPTED 06 July 2022
PUBLISHED 04 October 2022

CITATION
Meszaros J, Minari J and Huys I (2022), The future regulation of artificial intelligence systems in healthcare services and medical research in the European Union.
Front. Genet. 13:927721.
doi: 10.3389/fgene.2022.927721

COPYRIGHT
© 2022 Meszaros, Minari and Huys. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

The future regulation of artificial intelligence systems in healthcare services and medical research in the European Union

Janos Meszaros[1,2]*, Jusaku Minari[3] and Isabelle Huys[1,2]

[1]Division of Clinical Pharmacology and Pharmacotherapy, Department of Pharmaceutical and Pharmacological Sciences, KU Leuven, Leuven, Belgium, [2]Centre for IT and IP Law (CiTiP), KU Leuven, Leuven, Belgium, [3]Uehiro Research Division for iPS Cell Ethics, Center for iPS Cell Research and Application (CiRA), Kyoto University, Kyoto, Japan

Despite its promising future, the application of artificial intelligence (AI) and automated decision-making in healthcare services and medical research faces several legal and ethical hurdles. The European Union (EU) is tackling these issues with the existing legal framework and drafting new regulations, such as the proposed AI Act. The EU General Data Protection Regulation (GDPR) partly regulates AI systems, with rules on processing personal data and protecting data subjects against solely automated decision-making. In healthcare services, (automated) decisions are made more frequently and rapidly. However, medical research focuses on innovation and efficacy, with less direct decisions on individuals. Therefore, the GDPR's restrictions on solely automated decision-making apply mainly to healthcare services, and the rights of patients and research participants may significantly differ. The proposed AI Act introduced a risk-based approach to AI systems based on the principles of ethical AI. We analysed the complex connection between the GDPR and AI Act, highlighting the main issues and finding ways to harmonise the principles of data protection and ethical AI. The proposed AI Act may complement the GDPR in healthcare services and medical research. Although several years may pass before the AI Act comes into force, many of its goals will be realised before that.

KEYWORDS

GDPR—General Data Protection Regulation, artificial intelligence, AI Act, healthcare, medical research, data protection, automated decision-making, European Union

1 Introduction

Information technology (IT) companies invest heavily in and cooperate with healthcare organisations to apply their technology in healthcare services and medical research (Corrales Compagnucci et al., 2022). Google (Shetty, 2019) and Apple (Apple, 2021) are present in a growing number of medical fields, from diagnosing cancer to predicting patient outcomes. IBM has made great efforts to apply its artificial intelligence (AI) technology in healthcare by partnering with hundreds of hospitals, healthcare

organisations and researchers worldwide to translate data into better care (IBM Watson Health in Oncology, 2020).

Despite the promising results, the proliferation of AI applications in healthcare and medical research faces technological, legal and ethical issues. The main technological issues are the lack of interoperability and standardisation among medical IT systems (Brindha, 2012). From the ethical perspective, healthcare decisions often involve complex judgments and grasping the social context, which AI applications still struggle to replicate or simulate (Louwerse et al., 2005). Reliability and transparency are crucial aspects of building trust in care relationships (Wachter, 2010), and the opaque nature of AI applications might undermine these relationships (Cabitza and Zeitoun, 2019). Moreover, algorithms can underperform in novel cases of drug side effects and underrepresented populations, possibly leading to discrimination (Garcia, 2017).

Building and training AI systems require a vast amount of accurate data, which can contain sensitive medical information in healthcare services and medical research. Therefore, data protection is a critical legal matter, especially in the European Union (EU), under the General Data Protection Regulation (GDPR). The GDPR prohibits solely automated decision-making (ADM) and processing of health data, with a few exemptions, such as if it is done with the patient's consent or for the public interest. Hence, using health data with AI systems for ADM can face significant legal restrictions. However, the GDPR encourages innovation and technological developments, especially in scientific research, where there are several broad exemptions. Our paper elucidates how these special rules affect the development and application of AI systems in healthcare and medical research.

Nevertheless, the GDPR only partly covers the regulation of AI systems, with rules on processing personal data and protecting data subjects against ADM. It does not provide comprehensive protection against AI systems. Thus, AI regulation has become a central policy question in the EU (European Commission, 2019a), moving from a soft-law approach, with its non-binding guidelines, to a legislative approach that calls for a new regulatory framework on AI by proposing the AI Act. The proposal aims to establish horizontal rules for the development and application of AI-driven products, services and systems in the EU.[1] With the proposed AI Act, the EU aims to establish a technology-neutral definition of AI systems in EU law and to lay down a classification system for AI systems with different requirements and obligations tailored to a "risk-based approach".

Given that the interaction between the GDPR and the proposed AI Act may result in a complex legal framework in the future, we elucidate herein the emerging regulatory issues on AI systems in healthcare services and medical research in the EU. We first analyse the legal background of ADM and scientific research in the GDPR. We then introduce and clarify the proposed AI Act regarding healthcare services and medical research. Finally, the article concludes with a novel elaboration on the connection between the principles of data protection and ethical AI.

2 Data protection and automated decision-making in healthcare and medical research

Traditionally, health data are collected and processed for specific purposes, such as diagnosis and direct care. Thus, data protection and medical laws worldwide encompass the purpose limitation principle, which means that health data should not be processed for a new purpose, except if certain conditions are met. However, modern healthcare systems and applications, such as AI medical devices, can collect and process a vast amount of health data that can be used for scientific research and policy planning (Vayena and Tasioulas, 2016). In the age of big data and AI, technology provides unprecedented opportunities for the secondary use of health data (Coorevits et al., 2013; see also Corrales Compagnucci, 2019). It would need disproportionate efforts to acquire explicit consent from a large number of data subjects for new processing purposes, which poses complex ethical, legal and technical challenges (Burton et al., 2017). Hence, the purpose limitation principle is increasingly being challenged by researchers and policymakers to provide more efficient care while saving on expenses. Countries must balance citizens' autonomy, the public interest, and safeguards when healthcare data are reused for secondary purposes to address these challenges (Rumbold and Pierscionek, 2017). The onset of the coronavirus disease 2019 (COVID-19) pandemic became another vital reason to harvest health data to protect public health and address the current pandemic and future ones.

Health data are defined broadly in the GDPR as "personal data related to the physical or mental health of a natural person, including the provision of healthcare services, which reveal information about his or her health status".[2] The GDPR generally prohibits processing sensitive data, such as health data.[3] However, it provides several exemptions from this prohibition, including the case of public health emergencies during the COVID-19 pandemic. These exemptions include when "processing is necessary for reasons of public interest in the area of public health, such as protecting against serious cross-border threats to health"[4] or when "necessary for reasons of

1 Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Brussels, 21.4.2021, COM (2021) 206 final.

2 GDPR Article 4 (15).

3 GDPR Article 9.

4 GDPR Article 9(i).

substantial public interest".[5] The most practical legal basis for private companies' processing of data is the data subjects' consent or a legitimate interest.[6] For governments, public interest might be a more appropriate legal basis than the data subjects' consent. The European Data Protection Board has emphasised that consent is not the optimal basis of public authorities' processing of data due to the power imbalance between the citizens and the authorities (European Data Protection Board 2021), which is also true in the context of the COVID-19 outbreak (European Data Protection Board 2020; see also Fedeli et al., 2022).

2.1 Profiling and (solely) automated decision-making

The GDPR's rules on profiling and (solely) ADM have significantly impacted the application of AI systems in healthcare services and medical research. It is crucial to differentiate profiling, ADM, and solely ADM from each other.

The GDPR defines profiling as follows:

> Any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.[7]

The most important elements of profiling are 1) automated processing and 2) evaluating the personal aspects of a natural person. As Article 29 Working Party highlighted, "evaluating" indicates that profiling may involve assessing or judging a person. A simple classification of people does not constitute profiling.[8] For instance, when a healthcare provider sorts patients by age or gender without predictions or further assessment, it is not considered profiling. The Council of Europe's Recommendation[9] identified three stages of profiling: 1) data collection, 2) automated analysis to identify correlations and 3) identifying the characteristics of present or future behaviour. Therefore, when COVID-19 patients' electronic health records with automated analysis systems are combined with their current diagnoses to predict the severity of their diseases, it constitutes profiling.

ADM means an automated decision regarding an individual, with meaningful human involvement, whereas "solely ADM" does not have meaningful human involvement and is a decision made exclusively by an algorithm. By contrast, profiling does not involve a decision and can be only a source of both types of ADM (see the examples in Table 1).[10] The first element of solely ADM is a "decision" (regarding an individual). In this regard, solely ADM affects healthcare services more than medical research because the primary goal of scientific research is producing new knowledge rather than making decisions regarding individuals (Meszaros and Ho 2021). The second element is the "lack of meaningful human involvement". To qualify as meaningful human involvement, "the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture. It should be carried out by someone who has the authority and competence to change the decision".[11] In healthcare services, a medical professional's expected level of oversight to reach "meaningful" involvement is still a debated topic. It needs to be more than routine approval to effectively protect patients against the potential errors of AI systems. The third element is "legal effects or similarly significant consequences", which might significantly affect a person's legal status or rights. A legal effect requires that the decision affects someone's legal rights, such as the freedom to associate with others, vote in an election, or take legal action. A legal effect may also affect a person's legal status or rights under a contract. Entitlement to or denial of a social service also belongs here.[12] Decisions in healthcare services thus fulfil this condition. The GDPR permits profiling and ADM for data controllers based on specific legal grounds, with appropriate safeguards. However, solely ADM is generally prohibited, with specific exceptions, such as explicit consent or Member State law (see Table 2).

Overall, the GDPR's prohibition of solely ADM has a significant effect on the application of AI systems in healthcare services, which might be avoided in several ways, such as with meaningful human involvement.

5 GDPR Article 9(h).

6 GDPR Article 6 and 9.

7 GDPR Article 4 (4).

8 Article 29 (Working Party). Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679 (2018) 7.

9 Council of Europe. The protection of individuals with regard to automatic processing of personal data in the context of profiling. Recommendation CM/Rec (2010)13 and explanatory memorandum.

10 Ibid.

11 Article 29 (Working Party). Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679 (2018) 21.

12 Ibid 21.

TABLE 1 Examples of profiling and (solely) automated decision-making in healthcare services related to COVID-19.

	Examples
Profiling	The patient's COVID-19 diagnosis is combined with her electronic health records (EHR). The AI system creates her health profile to predict the future severity of her disease (e.g., patients with diabetes have an increased chance of severe COVID-19 symptoms)
Solely automated decision-making	An AI system decides alone, **without human involvement**, if the COVID-19 patient can leave the hospital
Automated decision-making	There is a **meaningful human involvement:** the AI system in the hospital only supports the medical professionals who are making the final decisions

2.2 Scientific research in the General Data Protection Regulation

The GDPR has special rules on scientific research, encouraging innovation and technological development in and through such areas.[13] There are several exemptions from the strict rules in GDPR for scientific research. For instance, personal data can be used further without the data subjects' consent for research purposes, and the right to erasure (the right to be forgotten) can be rejected. It is not an uncommon practice in scientific research, especially in medical sciences, to process personal data for a purpose different from the original one (i.e., "secondary use" or "further processing") to pursue new findings (Auffray et al., 2016). The GDPR acknowledges that "it is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection".[14] This recognition is crucial because it became more difficult to obtain consent under the GDPR as the consent must be unambiguous and specific to the processing operation.[15] The GDPR, in principle, forbids data controllers from processing sensitive personal data,[16] and as a general rule, researchers may use sensitive data only with specific legal grounds, such as explicit consent.[17] However, the GDPR also intends to ease the restrictions on processing sensitive data by explicitly permitting processing for research purposes. To obtain this permission, data controllers must apply appropriate safeguards,[18] such as de-identification.

The GDPR defines scientific research as "technological development and demonstration, fundamental research, applied research and privately funded research" conducted by both public and private entities.[19] Furthermore, the GDPR supports technological and scientific developments by citing the Treaty on the Functioning of the European Union to achieve the European Research Area.[20] However, the GDPR defines scientific research in the recital part, which is not legally binding.[21] Therefore, the EU Member States can tailor its scope, resulting in a fragmented legal landscape across the EU, which is against the main goal of GDPR. The European Data Protection Supervisor also highlighted the possible misinterpretation of this exemption. For instance, a company doing research may interpret the pertinent provisions in GDPR as allowing the retention of personal data for indefinite periods and denying data subjects' rights to information (European Data Protection Supervisor, 2020). Due to this broad exemption for research purposes, it is crucial to clarify and harmonise the definition of scientific research and appropriate safeguards at the EU level (Amram, 2020; Ducato, 2020).

2.3 The impact of scientific research on data subjects' rights in the General Data Protection Regulation

The GDPR has a special legal regime for scientific research, which heavily influences the data subjects' rights. When personal data are processed for scientific research purposes, Union or Member State law may provide for derogations from the rights of access (Article 15), rectification (Article 16), erasure (Article 17) and restriction of such processing (Article 18) and from the right to object (Article 21). These derogations are provided if these rights are likely to render impossible or seriously impair the achievement of the research purposes and if such derogations are

13 The GDPR Recital 157 also highlights that "By coupling information from registries, researchers can obtain new knowledge of great value with regard to widespread medical conditions such as cardiovascular disease, cancer and depression".

14 GDPR Recital 33 and 65.

15 GDPR Article 4 (11).

16 GDPR Article 9 (1).

17 GDPR Article 9 (1) (a).

18 GDPR Article 9 (2) (j).

19 GDPR Recital 159.

20 Treaty on the Functioning of the European Union, Article 179 (1). The Union shall have the objective of strengthening its scientific and technological bases by achieving a European research area in which researchers, scientific knowledge and technology circulate freely, and encouraging it to become more competitive, including in its industry, while promoting all the research activities deemed necessary by virtue of the other chapters of the treaties.

21 In the EU law, a recital is part of the text, usually the beginning of the law, which explains the reasons for the provisions, and it is not normative, thus legally not binding. Recitals are usually general statements. The GDPR Recital gives guidelines for understanding the normative text and its purposes.

TABLE 2 The impact of profiling, automated decision-making and scientific research on the data subjects' rights in the General Data Protection Regulation (Meszaros, 2022).

	Profiling	Decision-making with profiling	Solely automated decision-making with profiling	Scientific research (no automated decision-making)
Prohibitions for data controllers	**Allowed** (based on specific legal grounds)		**General prohibition** (with exceptions)	**Allowed** (based on specific legal grounds)
Data subjects' rights	Right to be informed - data collected directly (Art. 13) and indirectly (Art. 14 (3)) Right of access (Art. 15) Right to rectification (Art. 16) Right to erasure (Art. 17) Right to restriction (Art. 18) Right to data portability (Art. 20) Right to object (Art. 21)			Right to information in the case of **directly** collected data (Art. 13) Right to data portability (Art. 20)

necessary for the fulfilment of the research purposes.[22] However, two rights remain for the data subjects in every case: the right to information and data portability (see Table 2).[23]

With the aforementioned special rules on scientific research, GDPR attempts to balance privacy and the "ethical and scientific imperative" to share personal data for scientific research (Meszaros, 2022). These rules provide robust protection for data subjects. However, the application of AI systems requires a more specific, novel regulation, which the EU aims for with the proposed AI Act.

3 The European Union Artificial Intelligence Act proposal

3.1 The regulation of artificial intelligence in the European Union

As the GDPR only partly covers the regulation of AI systems, mainly through processing personal data and protecting of data subjects against ADM, it does not provide comprehensive protection against AI systems. The regulation of these systems requires a more complex legal landscape with strict enforcement, especially in healthcare services and medical research. While the EU does not yet have a specific legal framework for AI, the European Commission (EC) highlighted the necessity of using a regulatory approach to promote this emerging technology and address the associated risks (European Commission, 2020). Due to the economic, legal and social implications of AI, in recent years, AI regulation has become a central policy question in the EU (European Commission, 2019a).

The EU adopted a soft-law approach with its non-binding Ethics Guidelines for Trustworthy AI (European Commission, 2019b) and Policy and Investment Recommendations in 2019 (European Commission, 2019c). However, with the publication of Communication on Fostering a European Approach to Artificial Intelligence (European Commission, 2021) in 2021, the EU shifted towards a legislative approach and called for a new regulatory framework on AI.

The EU unveiled a proposal for the AI Act in April 2021. The legislation would lay down a harmonised legal framework for developing and applying AI products and services. The AI Act aims to ensure that the AI systems made available in the EU market are safe, respect EU law, and provide legal certainty to facilitate investment and innovation in AI. The act seeks to facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation.[24] By comparison, it took GDPR more than 4 years from the proposal stage to be adopted, with a 2-year implementation period before it came into force. Although several years may pass before the proposed AI Act comes into force, similar to what happened with GDPR, many of its goals may be realised before that, in healthcare services and medical research.

22 GDPR Article 89.

23 GDPR Articles 13 and 20.

24 Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM (2021) 206), Explanatory Memorandum and Recitals 1 and 5.

3.2 Definition of artificial intelligence

There is no precise, globally accepted definition of AI. According to the High-Level Expert Group on Artificial Intelligence (AI HLEG),[25] AI is a scientific discipline that includes several approaches and techniques, such as machine learning (ML), reasoning, and robotics.[26] To ensure legal certainty, the EC aims to define AI more clearly in the proposed AI Act as a "software that is developed with [specific] techniques and approaches[27] and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations or decisions influencing the environments they interact with".[28] This broad definition covers AI systems that can be used on a standalone basis and those that can be used as product components.

Annex 1 of the AI Act proposal lists the techniques and approaches used to develop AI. Similar to the UNESCO's Recommendation on the Ethics of Artificial Intelligence, the proposed AI Act defines "AI system" as a range of software-based technologies that encompasses "machine learning", "logic and knowledge-based" systems and "statistical" approaches (UNESCO, 2021). ML is a branch of AI and computer science which focuses on using data and algorithms to imitate how humans learn, gradually improving its accuracy.[29] ML methods are applied in various fields of science, leading to more evidence-based decision-making. Deep learning is a family of ML models based on deep convolutional neural networks (Schmidhuber, 2015). These techniques are gaining popularity because they may achieve human-level performance in various medical fields (LeCun et al., 2015), such as detecting skin cancer (Esteva et al., 2017) and diabetic retinopathy (Ting et al., 2017). The EU plans to update Annex 1 with new approaches and techniques as these emerge, providing flexibility to the proposed AI Act.

3.3 Risk-based approach

The proposed AI Act will introduce a risk-based approach to regulating AI systems. With this solution, the legal intervention is tailored to different risk levels, distinguishing between 1) unacceptable risk, 2) high risk, 3) low or minimal risk.

3.3.1 Prohibited risk

The proposed AI Act explicitly bans harmful AI practices considered threats to people's safety, livelihoods and rights. Accordingly, it prohibits making the following available in the EU market or putting them into service or using them in the EU: 1) AI systems that deploy harmful manipulative "subliminal techniques"; 2) AI systems that exploit specific vulnerable groups (e.g., those with physical or mental disabilities); 3) AI systems used by public authorities or on their behalf for social-scoring purposes and 4) "real-time" remote biometric identification systems in publicly accessible spaces for law enforcement purposes, except in a limited number of cases.

In the context of using health data, "social scoring" may have relevance.[30] In essence, social scoring means using an AI system to evaluate the trustworthiness of individuals based on their behaviours or personal characteristics, leading to the detrimental or unfavourable treatment of an individual or a group of people.

From a medical perspective, an existing medical condition (e.g., mental disorder) may form a base for predictive social scoring. The relationship with healthcare authorities and adherence to public health measures may also be factors for social scoring, such as following quarantine measures or receiving vaccinations. As social scoring is an unacceptable risk, the EU aims to prohibit using AI for such purposes.

Detrimental or unfavourable treatment might be in a different social context and unrelated to the contexts in which the data were originally generated or collected. For instance, a person guilty of tax evasion cannot use public transport or some public health services due to social scoring. This unfavourable treatment would be unjustified or disproportionate.

3.3.2 High-risk artificial intelligence systems

The proposed AI Act lists high-risk AI systems in the eight specific areas below.

(1) Biometric identification and categorisation of natural persons: This may be crucial in healthcare services, such as for identifying and sorting patients in a hospital based on their medical history and appointments.

25 The High-Level Expert Group on Artificial Intelligence was tasked by the European Commission to provide advice on its artificial intelligence strategy.

26 54 High-Level Expert Group on Artificial Intelligence. A definition of AI: Main capabilities and scientific disciplines (2019), p. 8.

27 Listed in Annex 1 of the AI Act.

28 AI Act, Article 3 (1) and Recital 6.

29 https://www.ibm.com/cloud/learn/machine-learning [Accessed June 11, 2022].

30 AI Act proposal, Article 5(c): Social scoring means the "... use of AI systems by public authorities or on their behalf for the evaluation or classification of the trustworthiness of natural persons over a certain period of time based on their social behaviour or known or predicted personal or personality characteristics, with the social score leading to either or both of the following: 1) detrimental or unfavourable treatment of certain natural persons or whole groups thereof in social contexts which are unrelated to the contexts in which the data was originally generated or collected; 2) detrimental or unfavourable treatment of certain natural persons or whole groups thereof that is unjustified or disproportionate to their social behaviour or its gravity".

(2) Management and operation of critical infrastructure: This may include the software for managing public healthcare services and electronic health records.
(3) Education and vocational training: AI systems will also affect the education of medical professionals. Students need to learn about AI products and services and prepare to use them due to their current proliferation in healthcare services and medical research.
(4) Employment, worker management and access to self-employment: The workforce in both public and private health services and research institutes may be affected by this future regulation.
(5) Access to and enjoyment of essential private and public services and benefits: As both public and private health services are mentioned here, the proposed AI Act may have a crucial impact on these fields.
(6) Law enforcement
(7) Migration, asylum and border control management
(8) Administration of justice and democratic processes

The list of high-risk AI systems in the annexe of the proposed AI Act provides flexibility for the EU as it can be modified and expanded in the future.[31] There are several requirements for these high-risk AI systems, such as risk management and data governance.[32] The providers of these systems are required to register their systems in an EU-wide database before making them available in the market or deploying them into service. However, several types of AI products already fall under conformity assessment, such as medical devices. These products remain under their current assessment framework.

3.3.3 Low- and minimal-risk at systems

Low- or minimal-risk AI systems can be developed and used in the EU without conforming to any additional legal obligations. However, the proposed AI Act envisages the voluntary creation of codes of conduct to provide safe and reliable services. Examples of these AI systems are those interacting with humans (e.g., chatbots) and provide emotional recognition. These tools may help interact with patients in healthcare services and participants in medical research.[33]

4 Discussion and actionable recommendations

To realise AI's potential in healthcare and medical research, new laws regulating AI systems are necessary (Humerick, 2018), based on the existing guidelines and harmonised with GDPR. The proposed AI Act is a crucial step herein. However, harmonisation with GDPR is an essential legal issue that needs to be discussed. AI HLEG[34] has laid down the most important principles of ethical AI. We expand these principles into the healthcare context and elaborate on their connection with the GDPR's data protection principles, providing a novel perspective. Our goals are to highlight the critical issues on AI in healthcare and to provide recommendations for applying GDPR and the proposed AI Act in the future.

(1) Technical robustness and safety: To prevent or minimize the probability of unintentional harm, AI applications in healthcare and research need to be secure and resilient. Technical robustness also means ensuring a fallback plan in case something goes wrong and being accurate, reliable and reproducible. The GDPR and the proposed AI Act require technical robustness and safeguards for processing personal data and deploying AI systems.[35] However, both do not detail these safeguards due to the rapidly changing technological environment, providing "future-proof" regulation. The necessary safeguards, such as "pseudonymisation", differ among the EU Member States (Meszaros and Ho, 2018). Therefore, the required safeguards and the review process by authorities need harmonisation, especially in the case of AI systems for healthcare services and medical research (Malgieri, 2019).

The proposed AI Act provides two types of conformity assessments depending on the AI system: self-assessment and assessment by notified bodies. Regarding self-assessment, the developer of an AI system is responsible for compliance with the requirements on quality and safety. When the assessment is conducted by a notified body, an independent third party certifies the AI system's compliance. However, the review process by notified bodies needs to be harmonised in the EU, otherwise, the developers of AI systems will opt for the less strict notified bodies, resulting into forum shopping.

(2) Privacy and data governance: There is a complex connection between the GDPR and the proposed AI Act. They may complement each other and share definitions related to data protection, such as their rules on biometrics and special

31 AI Act Articles 7 and 8.
32 AI Act Articles 8–15.
33 AI Act Title IV.
34 Following the launch of its Artificial Intelligence Strategy in 2018, the European Commission appointed a group of 52 experts to provide advice regarding its implementation. The group members were selected through an open selection process and comprised representatives from academia, civil society and industry. https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence (Accessed October 12, 2020).
35 AI Act, Article 10 (5), GDPR Article 89.

categories of data.[36] The AI Act clarifies that it should not be understood as providing legal grounds for processing personal data, including special categories of personal data.[37] Therefore, in general, the AI Act does not provide a legal basis for the primary or secondary use of personal data, especially those under special categories, such as health data.

However, there are exemptions from the above rule, such as the concept of a "regulatory sandbox". A "regulatory sandbox" is a "safe space in which businesses can test innovative products, services, business models and delivery mechanisms without immediately incurring all the normal regulatory consequences of engaging in the activity in question" (Financial Conduct Authority, 2015). Regulatory sandboxes were first used within the financial technologies (FinTech) sector but have expanded into other sectors, including healthcare (Leckenby et al., 2021; see also Fenwick et al., 2018). The AI Act will provide a legal basis for processing personal data for developing certain AI systems in the public interest within the AI regulatory sandbox, in line with the GDPR.[38]

(3) Human agency and oversight: These are essential, especially in high-risk AI systems. Human oversight has a central role in the proposed AI Act,[39] which states that it "will also facilitate the respect of other fundamental rights by minimising the risk of erroneous or biased AI-assisted decisions in critical areas". As we previously highlighted, the GDPR's restrictions on solely ADM can be avoided with meaningful human involvement. However, to qualify as having meaningful human involvement, "the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture" and "it should be carried out by someone who has the authority and competence to change the decision".[40] Overall, proper oversight is necessary, especially in the case of AI medical devices and applications, for patient and research participant safety.
(4) Transparency: Transparency is one of the data-processing principles in GDPR,[41] which prevails through several rights, such as the right to access and be informed.[42] In the proposed AI Act, transparency is required for specific AI systems, such as high-risk ones. In healthcare services and medical research, decisions need to be transparent and explainable for safety and trust. Furthermore, scientific research aided by AI applications should be transparent for reproducibility and inquiries about bias and safety.
(5) Diversity, non-discrimination and fairness: The data used to train AI systems need to be diverse to avoid bias. This requirement is of utmost importance in the case of AI systems because they might cause harm to populations underrepresented in healthcare. Therefore, one of the aims of the AI Act proposal is to "minimise the risk of algorithmic discrimination, in particular concerning the design and the quality of data sets used for the development of AI systems complemented with obligations for testing, risk management, documentation and human oversight throughout the AI systems' lifecycle".[43]
(6) Accountability, societal and environmental well-being: As highlighted by AI HLEG, mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes.[44] Certain actors, such as the government, IT, or special insurance companies, should be held responsible for the unintended consequences of these services. Finally, when AI is used for healthcare and research, it is crucial to use it transparently to benefit the whole society by respecting democratic values and decisions.

Overall, the black-box nature of AI applications and devices cannot be an excuse for complying with privacy and safety regulations. The proposed AI Act also highlights that it complements the GDPR without prejudice.[45] These two regulations can be the main pillars of safety and innovation in AI systems for healthcare and medical research.

5 Conclusion

The GDPR's prohibition of solely automated decision-making significantly effects the application of AI systems in medical research and healthcare services. While in medical research, the main focus is on innovation and efficacy, in healthcare services (automated) decisions are made frequently, even rapidly. Therefore, the GDPR's restrictions on solely automated decision-making apply mainly to healthcare services. Hence, the rights of patients and research participants may differ significantly.

The proposed AI Act introduced a risk-based approach to AI systems based on the principles of ethical AI. We highlighted the

36 GDPR Article 9.
37 AI Act proposal Recital 41.
38 AI Act Recital 72.
39 AI Act Article 14.
40 Article 29. Working Party on Profiling (2018), p. 21.
41 GDPR Article 5 (1)a.
42 GDPR Article 15.
43 AI Act 1.2. Consistency with existing policy provisions in the policy area.
44 High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI (2018).
45 AI Act 1.2. Consistency with existing policy provisions in the policy area.

complex connection between the GDPR and the proposed AI Act. For instance, they may complement each other and share the same definitions related to data protection. In some cases, the AI Act may provide a legal ground for processing personal data. Human agency and oversight must also be harmonised, especially the expectations of meaningful human involvement, in connection with the GDPR's rules on solely automated decision-making.

The current and future regulation of AI and data protection in the EU need to align well to provide a safe and innovative future. Although several years may pass before the proposed AI Act comes into force, many of its goals may start being realised before that. Harmonising the data protection principles and ethical AI is a complex but desirable goal, especially in healthcare services and medical research.

Author contributions

JMe is the first author. JMi and IH are shared second authors. All authors listed have made a substantial, direct contribution and approved the final version of the manuscript.

Funding

JMi is supported by the SECOM Science and Technology Foundation.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Amram, D. (2020). Building up the "Accountable Ulysses" model. The impact of GDPR and national implementations, ethics, and health-data research: Comparative remarks. *Comput. Law Secur. Rev.* 37, 105413. doi:10.1016/j.clsr.2020.105413

Apple (2021). *The future of healthcare is in your hands*. Available at: https://www.apple.com/healthcare/?fbclid=IwAR1A0cAPQow-4T-2tClTMPQUn4l4nQjURLisCzRdn6N24iw9iErKWSj2UV4 (Accessed May 3, 2021).

Auffray, C., Balling, R., Barroso, I., Bencze, L., Benson, M., Bergeron, J., et al. (2016). Erratum to: Making sense of big data in health research: Towards an EU action plan. *Genome Med.* 8 (1), 118. doi:10.1186/s13073-016-0376-y

Brindha, G. (2012). A new approach for changes in health care. *Middle East J. Sci. Res.* 12 (12), 1657–1662. doi:10.5829/idosi.mejsr.2012.12.12.19

Burton, P., Banner, N., Elliot, M. J., Knoppers, B. M., and Banks, J. (2017). Policies and strategies to facilitate secondary use of research data in the health sciences. *Int. J. Epidemiol.* 46 (6), 1729–1733. doi:10.1093/ije/dyx195

Cabitza, F., and Zeitoun, J-D. (2019). The proof of the pudding: In praise of a culture of real-world validation for medical artificial intelligence. *Ann. Transl. Med.* 7 (8), 161. doi:10.21037/atm.2019.04.07

Coorevits, P., Sundgren, M., Klein, G. O., Bahr, A., Claerhout, B., Daniel, C., et al. (2013). Electronic health records: New opportunities for clinical research. *J. Intern. Med.* 274 (6), 547–560. doi:10.1111/joim.12119

Corrales Compagnucci, M. (2020). "Big data, databases and "ownership" rights in the cloud," in *Perspectives in law, business and innovation* (Singapore: Springer). doi:10.1007/978-981-15-0349-8

Corrales Compagnucci, M., Fenwick, M., Haapio, H., Minssen, T., and Vermeulen, E. P. M. (2022). "Technology-driven disruption of healthcare & "UI layer" privacy-by-design," in *AI in eHealth: Human autonomy, data governance & privacy in healthcare*. Editors M. Corrales Compagnucci, M. L. Wilson, M. Fenwick, N. Forgó, and T. Bärnighausen (Cambridge, Cambridge Bioethics and Law: Cambridge University Press), 19–67.

Ducato, Rossana (2020). Data protection, scientific research, and the role of information. *Comput. Law Secur. Rev.* 37, 105412. doi:10.1016/j.clsr.2020.105412

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118. doi:10.1038/nature21056

European Commission (2019a). *Communication on building trust in human-centric artificial intelligence*. Brussels, 168. COM Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52019DC0168&from=BG (Accessed October 26, 2021).

European Commission (2021). *Communication on fostering a European approach to artificial intelligence*. Brussels. Available at: https://digital-strategy.ec.europa.eu/en/library/communication-fostering-european-approach-artificial-intelligence (Accessed March 11, 2021).

European Commission (2019b). *Ethics guidelines for trustworthy AI*. Brussels. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (Accessed May 26, 2021).

European Commission (2019c). *Policy and investment recommendations for trustworthy artificial intelligence*. Brussels. Available at: https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence (Accessed February 11, 2021).

European Commission (2020). *White paper on artificial intelligence - a European approach to excellence and trust*. Brussels. Available at: https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (Accessed October 26, 2021).

European Data Protection Board (2021). *EDPB document on response to the request from the European Commission for clarifications on the consistent application of GDPR, focusing on health research*. Available at: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_replyec_questionnaireresearch_final.pdf (Accessed February 9, 2022).

European Data Protection Board (2020). *Guidelines 04/2020 on the use of location data and contact tracing tools in the context of the COVID-19 outbreak*. Available at: https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_20200420_contact_tracing_covid_with_annex_en.pdf (Accessed February 16, 2022).

European Data Protection Supervisor (2020). *A preliminary opinion on data protection and scientific research*. Available at: https://edps.europa.eu/data-protection/our-work/publications/opinions/preliminary-opinion-data-protection-and-scientific_en (Accessed October 22, 2021).

Fedeli, P., Scendoni, R., Cingolani, M., Corrales Compagnucci, M., Cirocchi, R., and Cannovo, N. (2022). Informed consent and protection of personal data in genetic research on COVID-19. *Healthcare* 202210, 349. doi:10.3390/healthcare10020349

Fenwick, M., Vermeulen, E. P. M., and Corrales, M. (2018). "Business and regulatory responses to artificial intelligence: Dynamic regulation, innovation ecosystems and the strategic management of disruptive technology," in *Robotics, AI and the future of law*. Editors M. Corrales Compagnucci, M. Fenwick, and N. Forgó. 1 edn (Singapore: Springer), 81–103.

Perspectives in Law, Business and Innovation. doi:10.1007/978-981-13-2874-9_4

Financial Conduct Authority (2015). *Regulatory sandbox*. Available at: https://www.fca.org.uk/publication/research/regulatory-sandbox.pdf (Accessed March 25, 2020).

Garcia, M. (2017). Racist in the machine: The disturbing implications of algorithmic bias. *World Policy J.* 33 (4), 111–117. doi:10.1215/07402775-3813015

Humerick, M. (2018). Taking AI personally: How the EU must learn to balance the interests of personal data privacy & artificial intelligence. *Santa Clara Comput. High Tech L. J.* 34, 415. Available at: https://digitalcommons.law.scu.edu/chtlj/vol34/iss4/3 (Accessed February 10, 2021).

IBM Watson Health in Oncology (2020). Scientific Evidence. Available at: https://www.ibm.com/downloads/cas/NPDPLDEZ.

Leckenby, E., Dawoud, D., Bouvy, P., and Jónsson, P. (2021). The sandbox approach and its potential for use in health technology assessment: A literature review. *Appl. Health Econ. Health Policy* 19, 857–869. doi:10.1007/s40258-021-00665-1

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi:10.1038/nature14539

Louwerse, M. M., Graesser, A. C., Lu, S., and Mitchell, H. H. (2005). Social cues in animated conversational agents. *Appl. Cogn. Psychol.* 19 (6), 693–704. doi:10.1002/acp.1117

Malgieri, G. (2019). Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations. *Comput. Law Secur. Rev.* 35 (Issue 5), 105327. doi:10.1016/j.clsr.2019.05.002

Meszaros, J., and Ho, C-H. (2021). AI research and data protection: Can the same rules apply for commercial and academic research under the GDPR? *Comput. Law Secur. Rev.* 41, 105532. doi:10.1016/j.clsr.2021.105532

Meszaros, J., and Ho, C-H. (2018). Big data and scientific research: The secondary use of personal data under the research exemption in the GDPR. *Hung. J. Leg. Stud.* 59 (No 4), 403–419. doi:10.1556/2052.2018.59.4.5

Meszaros, J. (2022). "The next challenge for data protection law: AI revolution in automated scientific research," in *AI in eHealth: Human autonomy, data governance & privacy in healthcare*. Editors M. C. Compagnucci, M. L. Wilson, M. Fenwick, N. Forgó, and T. Bärnighausen (Cambridge: Cambridge Bioethics and Law, Cambridge University Press), 264.

Rumbold, J. M. M., and Pierscionek, B. K. (2017). A critique of the regulation of data science in healthcare research in the European Union. *BMC Med. Ethics* 18, 27. doi:10.1186/s12910-017-0184-y

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Netw.* 61, 85–117. doi:10.1016/j.neunet.2014.09.003

Shetty, S. (2019). *A promising step forward for predicting lung cancer*. Available at: https://blog.google/technology/health/lung-cancer-prediction/ (Accessed May 15, 2021).

Ting, D. S. W., Cheung, C. Y., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., et al. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 318 (22), 2211–2223. doi:10.1001/jama.2017.18152

UNESCO (2021). *Recommendation on the ethics of artificial intelligence*, 10. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000380455 (Accessed June 9, 2022).

Vayena, E., and Tasioulas, J. (2016). The dynamics of big data and human rights: The case of scientific research. *Phil. Trans. R. Soc. A* 374 (2083), 20160129. doi:10.1098/rsta.2016.0129

Wachter, R. M. (2010). Why diagnostic errors don't get any respect-and what can be done about them. *Health Aff.* 29 (9), 1605–1610. doi:10.1377/hlthaff.2009.0513

frontiers | Frontiers in Genetics

TYPE Perspective
PUBLISHED 26 January 2023
DOI 10.3389/fgene.2023.1098439

Check for updates

OPEN ACCESS

EDITED BY
Aviad Raz,
Ben-Gurion University of the Negev, Israel

REVIEWED BY
Zhongshan Cheng,
St. Jude Children's Research Hospital, United States
Netta Avnoon,
Tel Aviv University, Israel

*CORRESPONDENCE
Marie-Christine Fritzsche,
marie-christine.fritzsche@tum.de

SPECIALTY SECTION
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Genetics

RECEIVED 14 November 2022
ACCEPTED 04 January 2023
PUBLISHED 26 January 2023

CITATION
Fritzsche M-C, Akyüz K, Cano Abadía M, McLennan S, Marttinen P, Mayrhofer MT and Buyx AM (2023), Ethical layering in AI-driven polygenic risk scores—New complexities, new challenges. *Front. Genet.* 14:1098439. doi: 10.3389/fgene.2023.1098439

COPYRIGHT
© 2023 Fritzsche, Akyüz, Cano Abadía, McLennan, Marttinen, Mayrhofer and Buyx. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Ethical layering in AI-driven polygenic risk scores—New complexities, new challenges

Marie-Christine Fritzsche[1,2]*, Kaya Akyüz[3,4], Mónica Cano Abadía[3], Stuart McLennan[1,2], Pekka Marttinen[5], Michaela Th. Mayrhofer[3] and Alena M. Buyx[1,2]

[1]Institute of History and Ethics in Medicine, TUM School of Medicine, Technical University of Munich, Munich, Germany, [2]Department of Science, Technology and Society (STS), School of Social Sciences and Technology, Technical University of Munich, Munich, Germany, [3]Biobanking and Biomolecular Resources Research Infrastructure Consortium - European Research Infrastructure Consortium (BBMRI-ERIC), Graz, Austria, [4]Department of Science and Technology Studies, University of Vienna, Vienna, Austria, [5]Helsinki Institute for Information Technology HIIT, Aalto University, Helsinki, Finland

Researchers aim to develop polygenic risk scores as a tool to prevent and more effectively treat serious diseases, disorders and conditions such as breast cancer, type 2 diabetes mellitus and coronary heart disease. Recently, machine learning techniques, in particular deep neural networks, have been increasingly developed to create polygenic risk scores using electronic health records as well as genomic and other health data. While the use of artificial intelligence for polygenic risk scores may enable greater accuracy, performance and prediction, it also presents a range of increasingly complex ethical challenges. The ethical and social issues of many polygenic risk score applications in medicine have been widely discussed. However, in the literature and in practice, the ethical implications of their confluence with the use of artificial intelligence have not yet been sufficiently considered. Based on a comprehensive review of the existing literature, we argue that this stands in need of urgent consideration for research and subsequent translation into the clinical setting. Considering the many ethical layers involved, we will first give a brief overview of the development of artificial intelligence-driven polygenic risk scores, associated ethical and social implications, challenges in artificial intelligence ethics, and finally, explore potential complexities of polygenic risk scores driven by artificial intelligence. We point out emerging complexity regarding fairness, challenges in building trust, explaining and understanding artificial intelligence and polygenic risk scores as well as regulatory uncertainties and further challenges. We strongly advocate taking a proactive approach to embedding ethics in research and implementation processes for polygenic risk scores driven by artificial intelligence.

KEYWORDS

genomics, polygenic risk score, deep neural network (DNN), machine learning (ML), artificial intelligence—AI, stratification, predictive medicine, ethical

1 Introduction

Machine learning (ML) techniques, in particular deep neural networks (DNNs), are increasingly being developed to generate polygenic risk scores (PRSs) using electronic health records (EHRs) as well as genomic and other health data (Ho et al., 2019; Badré et al., 2021; Elgart et al., 2022). While this may allow greater accuracy, performance and prediction ability of PRSs, it also presents a range of increasingly complex ethical challenges.

PRSs are defined as "a weighted sum of the number of risk alleles an individual carries" (Lewis and Vassos, 2020). In medicine, PRSs estimate an individual's risk of a specific condition or disease based on their genetic makeup. Even though the genomes of individuals are to a large extent similar, there are genetic differences, which are called genetic variants (Broad Institute, 2021). If a genetic variant is more common in individuals who have a specific disease, it may be associated with an increased risk of that disease (Broad Institute, 2021). A PRS takes into account all these risk variants, however minimal their effect, to estimate an individual's risk of developing a disease (Broad Institute, 2021). Recently, PRSs have been developed to offer targeted risk prediction for a rapidly increasing number of conditions, including complex common diseases and conditions, such as breast cancer (Mavaddat et al., 2019), type 2 diabetes mellitus (Läll et al., 2017), coronary heart disease (Khera et al., 2016; Inouye et al., 2018), obesity (Khera et al., 2019), depression (Mitchell et al., 2021) and schizophrenia (Trubetskoy et al., 2022). Researchers aim to develop PRSs as a tool to prevent and more effectively treat serious diseases, disorders and conditions by identifying those at high risk who would benefit from targeted therapies.

The ethical and social implications of many PRS applications in medicine have already been widely discussed (e.g., Adeyemo et al., 2021; Knoppers et al., 2021; Slunecka et al., 2021). However, their confluence with ML has not yet been sufficiently considered in either literature or practice. We argue that the interaction between different and novel layers of ethical and social concerns pertaining to artificial intelligence (AI) and big data, as well as PRSs in research and translation into the clinical setting, stand in need of urgent consideration. This includes ethical aspects of AI as well as ethical and social implications of precision medicine and PRSs. We highlight potentially increasing complexities and the need to explore which new ethical and social issues arise from increased use of AI techniques for different PRS applications. We do so in the hope that those who aim to embed PRSs in healthcare systems take a proactive approach to embedding ethics during the research and implementation process. After giving a brief overview of the background to AI-driven PRSs, we consider the many ethical layers involved, beginning with the ethical and social implications of PRSs, then moving on to the challenges in AI ethics, and finally, exploring potential complexities of AI-driven PRSs.

2 Background to PRSs and AI-driven PRSs

Early studies on PRSs (Purcell et al., 2009; Dudbridge, 2013) applied the so-called *classic PRS method* (Choi et al., 2020), where the risk is calculated as a weighted sum (i.e., a linear regression) of a set of genetic risk alleles for given single nucleotide polymorphisms (SNPs) (see also Figure 1). The relevant subset of SNPs is selected using a genome-wide association study (GWAS), usually conducted in a cohort different from the target cohort, such that SNPs exceeding a certain p-value threshold are included in the calculation of the risk in the target population. Instead of using a subset corresponding to the significant SNPs, it is possible to include a much larger number of SNPs in the weighted sum to calculate the risk. When so many SNPs are included, it is necessary to prevent overfitting by applying shrinkage on the linear regression weights using either classic

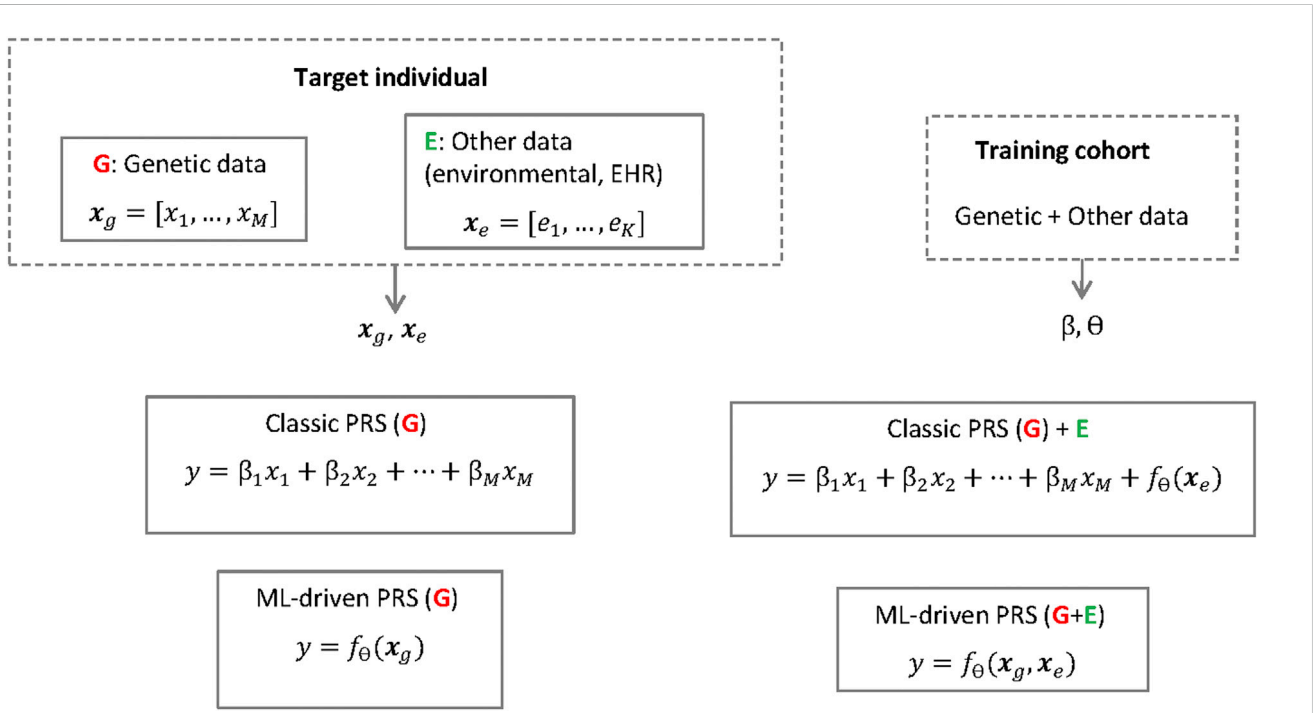


FIGURE 1
Classic PRSs and ML-driven PRSs the polygenic risk score for a target individual and phenotype of interest (y) is based on the individual's genetic data (x_g) but can also include other data types (x_e). The score is calculated using a linear regression (with weights β) or a machine learning model f_θ (e.g. a neural network with parameters θ). The parameters (β, θ) are learned using a separate training cohort. Note, however, that while the linear regession cofficients β are often publicly available or can be derived from published summary statistics, to train the neural network f_θ it is necessary to have access to individual level data in the training cohort.

techniques such as the LASSO or ridge regression (Mak et al., 2017) or Bayesian methods (Ge et al., 2019), the latter having given rise to some of the most popular implementations today (Vilhjalmsson et al., 2015). The SNP weights in a PRS can be derived from effects sizes published for the GWAS cohort, where the effect of each SNP on the risk has been estimated one-SNP-at-a-time, by accounting for linkage disequilibrium (LD) between the SNPs (Choi et al., 2020). Therefore, to apply classic PRS, individual-level data are only needed for the target individuals, but not from the GWAS cohort.

Recent years have witnessed attempts to replace the linear regression in PRS calculations with more sophisticated ML methods, which promise increased accuracy due to less restrictive modeling assumptions (Ho et al., 2019; Elgart et al., 2022). For example, DNNs which belong to the broader class of deep learning (DL), have been tested in PRSs for breast cancer, leading to improved scores compared to other statistical and ML estimation methods (Badré et al., 2021). A DNN processes input SNP data by passing them successively through multiple layers, where each layer takes the features from the previous layer as input, updates them, and passes the updated features forward to the next layer. In this way, features in higher layers can represent arbitrary, non-linear combinations of SNPs instead of the simple linear summation in conventional PRSs, which may better reflect the underlying biology.

Besides applying DL to modeling the genetic component, DL can alternatively be used to extract additional predictive features from EHRs (Miotto et al., 2016), which can be combined with the genetic data as input in PRS calculation (Dias and Torkamani, 2019). For example, using non-genetic risk factors together with genetic data improves the accuracy of breast cancer (Lee et al., 2019) and coronary artery disease (Inouye et al., 2018) risk modeling with the potential to enhance risk-based screening. However, current models typically build on combining genetics and EHR features additively (i.e., a simple summation), leaving room for the development of more complete approaches, for example a DNN that takes as input the different risk factors jointly to learn about the complex interplay between them.

Current research aims to pool and assess genomic data from biobanks, cohorts or registries on an unprecedented scale by combining it with environmental, other -omics data and health data such as EHRs. Considering the increasing heterogeneity of data that is used in the development of PRSs, more complex uses of AI have also been employed, such as making use of deep phenotypic information in medical images and EHRs to support downstream genetics analyses (Dias and Torkamani, 2019). Currently, PRSs typically only involve the genetic component, which is easier to interpret. The challenges in interpretation mainly occur when other data types are included, such as EHRs or gene expression data, the latter being different from SNPs that are currently used as data for PRSs (see also Figure 1). Other such data types are likely to increase in use, so the major challenges regarding the black box nature of the DNN models will probably be more relevant in the (near) future. Although researchers aim to reveal more and more causal relations, to date, analyses with AI for PRSs are mainly limited to correlations and improving predictions, which can result in inconclusive evidence (see Section 4). Barriers to the explainability of AI for PRSs also exist due to the statistical-probabilistic properties and the difficulty of the model to uncover the more complex biological, chemical and physical mechanisms that influenced it. In addition, there is a risk of potentially superfluous or inflated correlations due to the limitations of the method through phenomena such as the recently observed "cross-trait assortative mating" (Border et al., 2022). The risks of misinterpretation of (AI-driven) PRSs by clinicians, patients and other stakeholders involved should not be underestimated, especially as there may be a risk of drawing conclusions about causal relationships too quickly and where knowledge of statistics and causality/correlation claims is too low in many groups involved. Although the difference between causation and correlation is well understood by scientists, authors point to the need for education of the public about such differentiations for PRSs (Slunecka et al., 2021).

3 Ethical and social implications of PRSs

The potential benefits of the clinical use of PRSs may be manifold, both for individuals and/or society: identifying individuals at risk, improving the precision and range of differential diagnoses and treatments, as well as promoting the development of intervention thresholds. Incorporating polygenic risk profiles into population screening is expected to increase efficiency in contrast to screening stratified by age (Chowdhury et al., 2013; Torkamani et al., 2018; Kopp et al., 2020), while use of combined PRSs for various conditions in healthcare systems may contribute to early identification of potential non-genetic interventions and increased life expectancy (Meisner et al., 2020). Thus, PRSs may benefit individuals and represent a dramatic improvement of public health with potential socio-economic impacts. This has led to demands by PRS advocates within the medical community for a radical rethinking of PRSs as clinical instruments that could inform clinical decisions, such as in the prioritisation of psychosocial or pharmaceutical interventions "rather than treat/not treat decisions" (Lewis and Vassos, 2020).

While they come with important benefits, discussions in the literature on the multiple ethical and social implications for the medical use of PRSs range from social and distributive justice questions to debates on scientific validity and clinical utility (Babb de Villiers et al., 2020; Lewis and Vassos, 2020; Knoppers et al., 2021; Lewis and Green, 2021; Slunecka et al., 2021; Widen et al., 2022). In the context of PRS development and clinical implementation, ethical debates reflect those on monogenic genetic findings (Lewis and Green, 2021). Common concerns relate, for example, to genetic determinism as well as the concepts of ancestry/ethnicity, where tools such as AI for risk stratification may not be representative of human diversity and whose development and use may distract attention from the social determinants of health (Lewis and Vassos, 2020; Knoppers et al., 2021; Lewis and Green, 2021). Particular concerns about the risk of genetic discrimination and eugenics are raised with regard to the application of PRSs for embryo screening (Treff et al., 2019; Tellier et al., 2021; Turley et al., 2021); most recently for pre-implantation genetic testing (PGT) (Kozlov, 2022) and premature direct-to-consumer genetic testing/genetic counselling (Docherty et al., 2021), which are also intertwined with marketability and commercialisation. Furthermore, due to underrepresentation of already underserved communities in the research process, some authors note that health disparities could increase through the use of PRSs in the clinical setting (Martin et al., 2019a).

There has been extensive discussion of the clinical and/or personal utility of PRSs (Torkamani et al., 2018; Lambert et al., 2019; Wald and Old, 2019; Lewis and Vassos, 2020; Moorthie et al., 2021; Sud et al.,

2021). Scientific and clinical validity are challenges on multiple levels (Janssens, 2019; Lewis and Vassos, 2020; Knoppers et al., 2021), which touch ethical as well as epistemic concerns. PRSs, for example, do not cover the full risk for certain diseases because of the multiple factors involved. This includes e.g. environmental factors (Slunecka et al., 2021) and complex interactions between environments and PRSs (Domingue et al., 2020). Due to this complexity, interpretation of PRSs poses serious challenges, especially in relation to minors (Palk et al., 2019). From an ethical point of view, the necessity of communicating the limitations of risk prediction with PRSs therefore has to be considered in clinical applications. To this end, "effective and clear risk communication by trained professionals" should "minimize potential psychosocial effects" (Adeyemo et al., 2021). However, in this context, there is a lack of standardised PRS disclosure for individuals (Brockman et al., 2021; Lewis et al., 2022) as well as for kin, such as cascade screening for family members (Reid et al., 2021). Tools for standardisation of PRS disclosure have been developed for certain diseases, such as coronary artery disease (Widen et al., 2022), but the need for additional research on a broader range of populations and better standardisation has been emphasised (Brockman et al., 2021).

Given that PRSs are still an emerging field, there is remarkable heterogeneity around their application and reporting, thus constraining the implementation of PRSs in clinical settings (Slunecka et al., 2021). Publicly accessible catalogues and reporting standards for PRSs have been developed that are responsive to the current research landscape to allow reporting on the design and validation of PRSs within the literature (Lambert et al., 2021; Slunecka et al., 2021; Wand et al., 2021), such as the NHGRI-EBI, an extensive database of summary statistics of GWAS (Buniello et al., 2018). One aim of these efforts is to generate comparable PRSs metrics of performance (Lambert et al., 2021). This should increase the reproducibility and transparency of the PRS development process as well as support studies evaluating the clinical utility of the respective PRSs (Lambert et al., 2021). External and systematic PRS studies with benchmarking should also contribute to these aims (Wand et al., 2021). Another practical ethical issue is that the application of PRSs for medical purposes is presently uncertain under the majority of legal frameworks (Lewis and Vassos, 2020; Adeyemo et al., 2021).

Moreover, some authors also point out the importance of seeing PRSs in the respective context (Chatterjee et al., 2016; Torkamani et al., 2018; Slunecka et al., 2021), considering that the scope and diversity of available data (for instance, ancestry) and the techniques used to produce and use the scores are continuously changing (Trubetskoy et al., 2022). This therefore necessitates consideration, e.g., of the particular PRSs and the disease for which the PRSs were designed and the sophistication of the PRS itself. Consequently, the ethical and social implications need to be explored, taking into account the respective context. For example, specific ethical concerns in PRSs have been increasingly described for psychiatric conditions from informational risks in the use of the PRS in clinical setting, to the research showing links between the condition and social factors such as socioeconomic status or potential use in prenatal testing among others (Agerbo et al., 2015; Loh et al., 2015; Martin et al., 2019b; Palk et al., 2019; Docherty et al., 2021; Murray et al., 2021). This may differ for other conditions, for instance, in terms of actionability or potential for stigmatisation.

4 Challenges in AI ethics

There is much debate on ethical aspects around AI in healthcare (Morley et al., 2020), the role that AI should play (Rigby, 2019), the role and ethical implications of "explainability for AI in healthcare" (Amann et al., 2020), and ethical challenges of ML (Vayena et al., 2018) and of DL in healthcare (Char et al., 2018; Miotto et al., 2018). In particular, the following ethical and social challenges are often discussed in AI ethics (Mittelstadt et al., 2016; Floridi et al., 2021; Tsamados et al., 2021): How to ensure fairness and justice, overcome biases, ensure explainability, transparency, traceability, accountability, privacy, confidentiality, data protection and patient safety–how to design AI for the common good.

In AI ethics, not only are there normative concerns about algorithms such as "unfair outcomes" and "transformative effects", but also epistemic concerns such as "inconclusive evidence", "inscrutable evidence" and "misguided evidence" (Mittelstadt et al., 2016; Tsamados et al., 2021), and often epistemic and normative concerns come together as in the case of traceability. Issues such as the black box problem, accountability and transparency can be subsumed under inscrutable evidence (Mittelstadt et al., 2016). The black box problem in ML hinges on the lack of explainability as to how results are generated. The importance of this is also reflected in European law like the EU General Data Protection Regulation (GDPR) (European Parliament and Council of the European Union, 2016), which entails a general "right to explanation" (Goodman and Flaxman, 2017) for users and a future where explainability could become a legal requirement for ML specifically. The proposed Artificial Intelligence Act of July 2021 explicitly includes the requirement that AI systems be explainable for high-risk sectors (European Commission, 2021). The literature in recent years has repeatedly underlined the need for explainable AI (xAI) in medicine (Ribeiro et al., 2016; Hudec et al., 2018; Holzinger et al., 2019; Azodi et al., 2020), which is seen as (part of) a possible solution to many of the above-mentioned challenges in AI applications in healthcare.

Inconclusive evidence (Mittelstadt et al., 2016) involves ethical issues of causality and correlation, probabilities and predictions. Inconclusive evidence and incorrect causal associations and correlations are a problem for any statistical model, which can be the result, e.g. of biased sampling or hidden contamination. Authors generally point to the need to understand causality of the representations in ML systems (Pearl, 2009; Gershman et al., 2015; Peters et al., 2017; Holzinger et al., 2019). Furthermore, as substructures from genomic and population data are correlated, this can potentially result in false causal associations (Sohail et al., 2019) and misleading information based on bias embedded in genomic data (see Section 5.1). Increasing the robustness of the detected effects across different populations would go some way towards separating true causal effects from spurious associations. In genetics, replicating the findings in multiple cohorts is usually a stipulation, but more work is required to ensure inclusion of more diverse populations (see Section 5.1).

The topic of "misguided evidence leading to bias" (Mittelstadt et al., 2016) and "unfair outcomes leading to discrimination" (Mittelstadt et al., 2016) are key issues in AI ethics. In medical AI, biases (Obermeyer et al., 2019) abound, and the replication of biases and the amplification of real-world injustices by algorithms poses a serious risk.

There are many different proposals for frameworks on how the challenges of applying AI in medicine should be addressed ethically, which principles and values are of particular importance and which guidelines should be followed. Ethical challenges exist in terms of principles, not only regarding which principles should be considered crucial, but also in terms of differences in what the principles mean, e.g. what justice encompasses, as there are many different forms of justice derived from different philosophical theories and different underlying values (Whittlestone et al., 2019). Furthermore, there is the question of what "for the good of society" means—What would AI that is focused on the common good look like? This would need to be discussed and defined in each context (Whittlestone et al., 2019).

Another challenge usually arises when principles conflict with each other, as is often the case with AI in healthcare. Explainability is often not technically possible, and the benefits of AI can vary in significance, so the trade-off would have to be weighed up for each AI system and context. Another major ethical challenge around AI is putting principles into practice. Authors point out that attention needs to be paid to the tensions and conflicts that arise in this process and that these need to be addressed (Whittlestone et al., 2019) so that risks can be avoided and the benefits of AI can be reaped.

5 Bringing ethical and social aspects of PRSs and AI ethics together—New complexities for AI-driven PRSs?

In bringing ethical and social implications of PRSs and of AI ethics together, we would like to point out potential new complexities for AI-driven PRSs. Particularly around the following topic clusters which will be discussed in detail in what follows.

1) More complexity regarding fairness and justice
2) Challenges in building trust, communication and education
3) Privacy and autonomy challenges
4) Regulatory uncertainties and further challenges

5.1 More complexity regarding fairness and justice

Although many researchers point out the opportunities of xAI and interpretable ML (iML), two ethically relevant issues with respect to explanatory methods remain generally difficult to solve: different biases within datasets leading to biased DNN and suspicion of bias in results leading to unfairness (Ras et al., 2018). This could apply also to ML application for PRSs on multiple levels: many biases in PRS development can be linked to biases in the combination of EHRs with genomic and further health data as well as in the substructures of this data.

Firstly, the majority of genetic studies lack diversity (Sirugo et al., 2019). PRSs have mainly been developed with datasets from European populations and predictions of genetic risk are susceptible to unequal outputs (performance levels) across different populations as they are underrepresented in training data, which hinders generalisability (Martin et al., 2019a). Authors observe that research infrastructures like biobanks may suffer from "recruitment bias" as a risk which "infringes on the principle of justice, influences representativity of biobank collections and has implications for the generalizability of research results and ability to reach full statistical power" (Akyüz et al., 2021).

Secondly, further data biases can be linked to many other factors. There is a considerable gap in medical studies on the representation of women (Daitch et al., 2022) as the case of cardiovascular disease also shows (Burgess, 2022). More broadly, gender bias can be found in written documents used for certain ML techniques (Bolukbasi et al., 2016). Gender bias may also occur when heteronormative paradigms are not met, e.g., when data on gender and sex do not match and are therefore automatically excluded for analysis, which is currently a common practice in genomics (American Medical Association, 2018; Ganna et al., 2019). EHRs can contain multiple biases resulting e.g., from physician bias or certain delivery of care (Ching et al., 2018; Gianfrancesco et al., 2018), and even laboratory measurements (which are considered less biased) can show bias resulting from the patient health state and healthcare process (Pivovarov et al., 2014)—although they may be representative regarding population (Kerminen et al., 2019; Adeyemo et al., 2021). Overall, there are substructures in genomics and other health data that can be linked to actual differential causal relationships between health outcomes and putative risk factors. Other substructures can be traced to external factors such as cultural practices, socioeconomic status and other non-causal factors that relate to healthcare provision, access to medicine and clinical trials (Gianfrancesco et al., 2018; Dias and Torkamani, 2019; Sirugo et al., 2019).

Apart from the bias in data, *machine bias* has to be mentioned in the context of AI use for PRSs. This encompasses the biases that are learned by the models (Ching et al., 2018; Dias and Torkamani, 2019). In this context, one criterion for iML for genetic risk prediction could be whether a certain model is adequately interpretable for bias to be detected (Ching et al., 2018; Dias and Torkamani, 2019). Authors call for standards of fairness in order to diminish disparities caused by bias of ML in genetic risk prediction (Dias and Torkamani, 2019; McInnes et al., 2021). Moreover, they point to the necessity for careful application of AI and differentiation between the various forms of bias arising when AI is applied to genetic risk prediction (Dias and Torkamani, 2019). Tools are already being developed to help eliminate machine bias. This is not only intended to eliminate bias of ML, but also to create diagnostic systems that are much freer from human bias than classical diagnostics by physicians allow (Shen et al., 2019). These and further innovative sorts of techniques should also be consistently considered for ML use for PRSs.

In addition to injustice due to biases, *injustice and unfairness regarding data access and sharing data and algorithms* is also an issue for AI-driven PRSs. In this regard, biased processes and results are co-produced, potentially sustaining existing inequalities and unfairnesses. Further, apart from comprehensibility, accessibility can be considered as the second main component of transparency in generating information about how algorithms function (Mittelstadt et al., 2016). While many advances have been made thanks to international initiatives and large interdisciplinary research consortia, authors still highlight the ongoing need to collect, harmonise and share data in genomics and healthcare (Diao et al., 2018; Lambert et al., 2021). The Polygenic Risk Score Task Force of the International Common Disease Alliance has called for the "GWAS research community, global biobank collaborations, and private direct-to-consumer companies" (Adeyemo et al., 2021) to create requirements for public sharing of summary statistics using standardised formats, with the aim of avoiding the exacerbation of

worldwide health inequalities (Adeyemo et al., 2021). However, sharing DL models with the biomedical data and health records of individuals not only faces legal and technical barriers but also poses a major "cultural challenge" (Ching et al., 2018). A culture that rewards discovery rather than the production of data will have a difficult time motivating researchers to share their hard-earned datasets (Ching et al., 2018). However, as is pointed out in recent articles, it is this data that would drive DL (Ching et al., 2018).

Apart from well-known privacy regulations and standards in medical and biological research (Ching et al., 2018), factors such as the costs related to regulations for medical devices may also play an important role in access to PRSs, creating inequalities among populations, subgroups and countries (Adeyemo et al., 2021). Not only does global cooperation contribute to more equity in medical research and healthcare, it also serves an important role for the improvement of clinical validity and utility of PRSs (Adeyemo et al., 2021; Knoppers et al., 2021). Moreover, an open exchange of AI models for genetic risk prediction with the medical and scientific communities is called for to enhance transparency, where the model sharing should include details such as model weights, source codes and meta diagrams (Dias and Torkamani, 2019). Synthetic genetic and phenotypic data (Abadi et al., 2016) is suggested for genomic projects (Moorthie et al., 2021) and is already being tested in PRS development to provide greater diversity in genetic data, avoid biases and privacy issues. Furthermore, protecting data and privacy are very relevant for public-private partnerships (Murdoch, 2021), which play an increasingly important role for the implementation and dissemination of PRSs.

5.2 Challenges in building trust, communication and education

One of the greatest challenges in translating PRSs to the clinical setting is the communication of PRSs. This includes communication to and dialogues with the public(s) and patients as well as educating all other stakeholders involved. The challenge of communicating PRSs in the clinical setting, particularly for doctors (Fiske et al., 2019), is magnified when explaining AI-driven PRSs.

In general, we highlight the need for reflection on epistemological questions around AI use for PRSs and the corresponding normative aspects. It is important to ask what it means to explain, interpret and understand AI-driven PRSs. This should ideally incorporate different perspectives for certain stakeholders and involve further associated questions, e.g., what researchers consider an explanation to be, what kind of explanation users want and need (Slunecka et al., 2021) and what criteria are relevant for explainable PRSs. With the advance of xAI and iML, it is also worth considering how much/what kind of explainability is required for the clinical application of PRSs and how much/what kind of interpretability is clinically meaningful.

With regard to the literature reviewed and the existence of different definitions of explainability, explicability, interpretability and comprehensibility in scientific teams and clinical settings, we argue that awareness of these differences of terms must be raised both in scientific publications and in practice. This would also have the ultimate goal of improving the explainability of the risk scores and the underlying AI mechanisms.

Stakeholders in research and development as well as healthcare areas are constrained to consider the uncertainty of AI-generated PRS predictions and thus need to develop means of dealing with them in a structured, transparent and responsible way. Even if a more explainable ML for PRSs is developed, the question of how to communicate and generally deal with uncertainty due to lack of explainability of ML for PRSs nevertheless requires discussion and translation into appropriate standards. Embedded ethics approaches (McLennan et al., 2022) in both the research and clinical settings could help resolve the challenge of detecting and reflecting on ethical issues as well as communicating them.

Regarding communication of AI-driven PRSs, there is a clear need for engagement with technical, medical and ethical aspects of PRSs and AI for all the different stakeholders involved. We strongly recommend adopting interactive/participatory engagement practices (Horst et al., 2017), especially between clinicians and patients for AI-driven PRSs. This means limiting or avoiding deficit models of communication, i.e., unchallengeable, non-reflexive (Wynne, 1993) communication, which sees audiences (including any actors other than experts) as deficient both in knowledge and capacity to comprehend (Bell et al., 2008). In light of the developments in e-health, citizen-patients are not considered passive recipients of information, but rather self-informing, active individuals (Felt et al., 2009). Furthermore, the respective educational, socioeconomic and cultural background of individual patients and their families has to be considered when, for example, physicians explain PRSs (Slunecka et al., 2021).

In general, one of the biggest challenges of AI-driven PRSs today is trust in AI/AI-driven PRSs and trust in the medical institutions that will use these technologies on a large scale. However, there is a lack of specificity in the literature on issues of trust in the recently developed AI-driven PRSs. This represents a future issue that will need to be addressed with interdisciplinary teams.

Problems of AI explainability add complexity to matters of trust for AI-driven PRSs. Lack of transparency and lack of human understanding of AI black boxes raises the question of how all kinds of end-users create their relationship with AI. Scholars emphasise the importance of explainable AI (Holzinger et al., 2019) and DL models in medicine by arguing for the trust-building effect they have (Ribeiro et al., 2016). They point to the importance of understanding the rationale underlying the predictions of ML modelling when evaluating trust, which is considered crucial for decisions on the use of new models and actions based on predictions (Ribeiro et al., 2016). Interpretability is reflected in the "fidelity-interpretability trade-off" (Ribeiro et al., 2016) and is key to building trust in AI among healthcare professionals. Practitioners are very unlikely to accept a DL system that they do not understand (Miotto et al., 2018). It is noted that the interpretability of the model in genomics is critical to convincing health professionals of the validity of the actions the prediction system recommends, e.g., to explain which phenotypes drive certain predictions (Miotto et al., 2018).

The High-Level Expert Group on AI of the European Commission proposes trust as one of the defining principles for their AI ethics guidelines (High-Level Expert Group on AI, 2019). However, the technical solutions to the issue of trust, as discussed above, are unlikely to become available in definitive form. We therefore suggest that the social and relational considerations are paramount if we are to create a workable framework for establishing trust. This means the question of how trust is built needs to be addressed by adopting a more reflexive and interdisciplinary perspective. This also includes discussion of the *trustworthiness* of AI use for PRSs. Which is to say, discussions about dependable, trustworthy ML use for the PRSs and what requirements and criteria should be placed on the trustworthiness of AI for PRSs

must perforce address contextual questions, such as what trust means in a particular situation or context. The FUTURE-AI initiative has created dynamic best practices for trustworthy AI in healthcare (Future AI, 2022). Empirical and theoretical studies on the ethical and social issues of AI-driven PRSs and trustworthiness are needed so that this knowledge can then be integrated into the development and application of AI-driven PRSs.

Overall, we recognise that education and training of AI-driven PRSs would need to cover tech/AI literacy, risk interpretation/statistical knowledge, genomics/PRS knowledge, communication skills and ethical reflection skills of the stakeholders involved—of course with different granularities depending on the stakeholders: patients, relatives of patients, various public(s), healthcare professionals, medical/nursing students, researchers, technicians, ethics committees, clinical ethics teams, business partners and all the other stakeholders involved in research and development as well as the translation, implementation and application of AI-driven PRSs.

For AI in medicine generally, there is a need to increase education and training for different stakeholders in the healthcare system on applications of technology driven by data (Meskó et al., 2017; Xu et al., 2019). For ML in genomics, authors stress the need to bridge the gaps regarding clinical knowledge and interpreting models (Diao et al., 2018). Others consider the training of clinical staff to be a major challenge for the implementation of PRSs in the clinical setting (Torkamani et al., 2018; Slunecka et al., 2021). The unique nuances of PRSs and GWAS development are mostly unfamiliar to clinicians at this point (Martin et al., 2019a). Concrete suggestions have been made for enhancing education about PRSs for inclusion in the regular curriculum for medical students and in the ongoing education for medical professionals, covering the limitations of PRSs and different forms of risk (Slunecka et al., 2021). In addition, there are different proposals for how experts in genetic risk assessments could be involved in the clinical setting. Furthermore, education of the public(s) is crucial in implementing PRSs for public screening. The website of the National Human Genome Research Institute of the National Institutes of Health (UK), for instance, aims to explain to the public how PRSs work and how to interpret them. Apart from that, sensitivity, reflection and discussion on relationality and power relations of patients, doctors, healthcare and research institutions as well as biotechnology/genomics companies are important issues in the development of AI-driven PRSs. Based on a renewed understanding of how citizens engage with physicians and information technologies in health setting (Felt et al., 2009), empowering citizens and patients is among the key developments for the application of AI-driven PRSs.

5.3 Privacy and autonomy challenges

When large amounts of genomic data and EHRs are used to generate PRSs with AI, privacy is a key issue. A crunch question is whether protection of personal/patient data trumps transparency and right of access to data or *vice versa*. There are also multiple questions revolving around the extent to which anonymisation can be ensured with the large amounts of data used for PRSs, new AI technologies and what informed consent should look like for different uses of PRSs driven by them. For example, the differential privacy method, in which noise is added to data to prevent revealing individual information in case summaries of the data were to be published, does not scale easily to high-dimensional genetic data (Roth and Dwork, 2013). While there are efforts in medicine and PRS development aimed at protecting privacy (Abadi et al., 2016; Simmons et al., 2016; Ching et al., 2018; Beaulieu-Jones et al., 2019; Zhang et al., 2021), it is unclear how these could be implemented or policed on a large scale for AI-driven PRSs. Despite the discourse of exceptionalism of big data research, privacy is still an issue that is tightly entangled with autonomy (Rothstein, 2015). However, in a data-rich environment, genomic data, which is by definition shared in differing amounts with biological relatives, poses further challenges to our understandings and practices of privacy and autonomy, but also anonymisation or risk of genomic identifiability, raising the necessity for a "post-identifiability" lens (Akyüz et al., 2023). Thus, privacy and autonomy are challenges in their own right due to the peculiarity of genomic data.

5.4 Regulatory uncertainties and further challenges

As for healthcare in general, the need for complementary measures to explainability such as regulation (Markus et al., 2021), enhancing the quality of healthcare data for DL (Miotto et al., 2018) and external validation (Markus et al., 2021) have to be considered for AI-driven PRSs. The need for regulatory measures for PRSs in general is highlighted in the literature reviewed (Adeyemo et al., 2021; Knoppers et al., 2021; Slunecka et al., 2021). Standardisation of regulation frameworks for PRSs as medical devices (Adeyemo et al., 2021) is urgently required. With AI-driven PRSs, it is even more important to establish internationally standardised regulation frameworks which are responsive to the dynamic and fast-evolving technical and scientific findings around PRSs. Flexible, on-demand "*ad hoc*" guidance to positively enhance ongoing algorithm improvement (Vayena et al., 2018; Dias and Torkamani, 2019) would support the ethically sound development of AI-driven PRSs. However, regulatory measures can be a burden for people with access to PRS technology (Knoppers et al., 2021). In this sense, the challenge of creating a balance between sufficient regulation and rapid scientific advancement in the application of AI for PRSs must be considered in the development of AI-driven PRSs.

Beyond the ethical concerns mentioned above, further ethical challenges of AI-driven PRSs, such as informed consent procedures for AI-driven PRSs in absentia of explainability could become even more relevant in the future. In addition, the importance of AI for ethics committees has to be emphasised as does the need to involve research ethics committees and clinical ethics committees in the translation and implementation of AI-driven PRSs.

6 Conclusion

Our article has delineated the multiple layers of ethical and social concerns associated with PRSs, AI for PRSs and AI-driven PRSs in medicine. A clear limitation of most ML-based approaches compared with the *classic PRS* method is the requirement for individual level data to train the models, whereas the latter uses publicly available summary statistics about estimated effect sizes. Hence, there is room for development of new ways to leverage published summary statistics in training of more flexible ML-based PRS methods. Another limitation and future challenge common to all PRS methods is the poor generalisability of the scores in populations with different ancestries, which also stems from different allele frequencies, linkage disequilibrium and genetic effect sizes in different populations (Wang et al., 2022). Regarding the use of AI in PRS, there is great potential for improvement by developing models that integrate a

variety of health data types and risk factors into comprehensive predictors of disease risk (Dias and Torkamani, 2019). The clinical utility of PRSs is currently hotly debated; thus, more research is warranted on the best ways to implement PRSs as part of clinical practice, either to improve diagnoses, personalise treatments, or as part of preventive medicine (Torkamani et al., 2018; Choi et al., 2020). In particular, the additional challenges for the clinical implementation posed by the AI based PRS methods remain to be addressed. Furthermore, our discussion of some of the ethical issues that need to be considered in AI-driven PRS is in no way exhaustive. Rather, this article can serve as a basis for further discussions of the ethical challenges that could arise from the future application of AI-driven PRSs.

Where PRSs, ML and big data are part of the picture, we have teased out the more complex ethical challenges emerging from the relation between them, as well as pertaining to them individually. Based on a comprehensive review of the existing literature, we argue that this stands in need of urgent consideration for research and translation into the clinical setting. Different layers of ethical implications could lead to more challenges for explainability of AI-driven PRSs, more complexity of fairness with biases in data (sets) and ML for PRSs and biased outputs, more challenges in building trust, communication and education as well as regulatory uncertainties for and challenges in privacy and autonomy of AI-driven PRSs. Among these, we would especially like to highlight a lack of specificity in the literature on issues of trust in the more recent instantiations of AI-driven PRSs. We maintain that this is a future challenge that will need to be addressed in interdisciplinary, multi-stakeholder teams. The fact that the lack of explainability seems to be an inherent problem of certain ML techniques, which may never be fully solved, should not hinder efforts to make ML for PRSs more explainable and trustworthy for all stakeholders involved in the healthcare system. It has become clear that much of the more explainable PRSs depends not only on more explainable ML techniques, but also on awareness, context- and user-specific communication and engagement, education and training for all stakeholders. In addition, there are limitations to the influence of explainable ML that relate to ethical and social aspects associated with large amounts of data, such as EHRs, genomic and other health data fed into ML models. Apart from more technical research on e.g. techniques of explainable ML for PRSs, more ethical analyses are needed, covering epistemic and normative aspects of AI-driven PRSs including methods of normative and empirical ethics. We have also pointed out that hitherto there are few to no regulatory guidelines, and a lack of commensurate up-to-date research, let alone clear advice on how to communicate the potential implications, costs or benefits of these technological advances to and between the various stakeholders involved. For this, technical and bioethical content as well as discussions on the larger societal implications and public health aspects should also be included in the training for students and healthcare professionals. Although there are efforts to address the ethical and regulatory challenges of AI-driven PRSs, more work is required when AI tools are used with more complex health data such as EHRs and medical images or real world data. This should be an important item on the agenda of citizens, policymakers, scientists and funders of AI-driven PRS development as a co-production. This approach would make an important contribution to the clinical utility of PRSs in terms of transparency, responsibility and finally trustworthiness.

If we fail to address these challenges, the danger is that not only will advances in AI and/or the applications of PRSs outstrip our ability to understand or regulate them, but that the potential for overreliance and indeed misapplication or misuse from an ethical and social standpoint may create further and insurmountable complexities in the future.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

M-CF conceptualised and wrote the manuscript, and KA, MCA, and PM contributed to the writing process. MTM, SM, and AB edited the manuscript. All authors read and approved the final manuscript. M-CF conducted the literature review and theoretical analyses for the manuscript. KA and MCA supported the literature review and analyses. AB supervised the project.

Funding

This manuscript benefited from funding from INTERVENE (INTERnational consortium for integratiVE geNomics prEdiction), a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101016775. PM received funding from the Academy of Finland (Flagship programme: Finnish Center for Artificial Intelligence FCAI, and grants 336033, 352986).

Acknowledgments

The authors thank the TUM Language Center for the language editing.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

Where authors are identified as personnel of the Biobanking and BioMolecular resources Research Infrastructure (BBMRI-ERIC), the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of BBMRI-ERIC.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., et al. (2016). *Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. Vienna, Austria: Association for Computing Machinery, 308–318.

Adeyemo, A., Balaconis, M. K., Darnes, D. R., Fatumo, S., Granados Moreno, P., Hodonsky, C. J., et al. (2021). Responsible use of polygenic risk scores in the clinic: Potential benefits, risks and gaps. *Nat. Med.* 27 (11), 1876–1884. doi:10.1038/s41591-021-01549-6

Agerbo, E., Sullivan, P. F., Vilhjálmsson, B. J., Pedersen, C. B., Mors, O., Børglum, A. D., et al. (2015). Polygenic risk score, parental socioeconomic status, family history of psychiatric disorders, and the risk for schizophrenia: A Danish population-based study and meta-analysis. *JAMA Psychiatry* 72 (7), 635–641. doi:10.1001/jamapsychiatry.2015.0346

Akyüz, K., Chassang, G., Goisauf, M., Kozera, Ł., Mezinska, S., Tzortzatou, O., et al. (2021). Biobanking and risk assessment: A comprehensive typology of risks for an adaptive risk governance. *Life Sci. Soc. Policy* 17 (1), 10. doi:10.1186/s40504-021-00117-7

Akyüz, K., Goisauf, M., Chassang, G., Kozera, Ł., Mežinska, S., Tzortzatou-Nanopoulou, O., et al. (2023). Post-identifiability in changing sociotechnological genomic data environments. *BioSocieties*. (Accepted for publication).

Amann, J., Blasimme, A., Vayena, E., and Frey, D. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inf. Decis. Mak.* 20 (1), 310. doi:10.1186/s12911-020-01332-6

American Medical Association (2018). *AMA adopts new policies at 2018 interim meeting*. National Harbor: Press Release.

Azodi, C. B., Tang, J., and Shiu, S-H. (2020). Opening the black box: Interpretable machine learning for geneticists. *Trends Genet.* 36 (6), 442–455. doi:10.1016/j.tig.2020.03.005

Babb de Villiers, C., Kroese, M., and Moorthie, S. (2020). Understanding polygenic models, their development and the potential application of polygenic scores in healthcare. *J. Med. Genet.* 57 (11), 725–732. doi:10.1136/jmedgenet-2019-106763

Badré, A., Zhang, L., Muchero, W., Reynolds, J. C., and Pan, C. (2021). Deep neural network improves the estimation of polygenic risk scores for breast cancer. *J. Hum. Genet.* 66 (4), 359–369. doi:10.1038/s10038-020-00832-7

Beaulieu-Jones, B. K., Wu, Z. S., Williams, C., Lee, R., Bhavnani, S. P., Byrd, J. B., et al. (2019). Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation Cardiovasc. Qual. Outcomes* 12 (7), e005122. doi:10.1161/CIRCOUTCOMES.118.005122

Bell, A. R., Davies, S. R., and Mellor, F. (2008). *Science and its publics*. Newcastle: Cambridge Scholars.

Bolukbasi, T., Chang, K-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Adv. Neural Inf. Process. Syst.* 29, 4349–4357.

Border, R., Athanasiadis, G., Buil, A., Schork, A., Cai, N., Young, A., et al. (2022). Cross-trait assortative mating is widespread and inflates genetic correlation estimates. *Science* 378, 754–761. doi:10.1126/science.abo2059

Broad Institute (2021). Polygenic scores. Available at: https://polygenicscores.org/explained/ (Accessed December 20, 2022).

Brockman, D. G., Petronio, L., Dron, J. S., Kwon, B. C., Vosburg, T., Nip, L., et al. (2021). Design and user experience testing of a polygenic score report: A qualitative study of prospective users. *BMC Med. Genomics* 14 (1), 238. doi:10.1186/s12920-021-01056-0

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2018). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47 (D1), D1005–D12. doi:10.1093/nar/gky1120

Burgess, S. N. (2022). Understudied, under-recognized, underdiagnosed, andundertreated: Sex-based disparities incardiovascular medicine. *Circ. Cardiovasc. Interv.* 15, e011714. doi:10.1161/CIRCINTERVENTIONS.121.011714

Char, D. S., Shah, N. H., and Magnus, D. (2018). Implementing machine learning in health care — addressing ethical challenges. *N. Engl. J. Med.* 378 (11), 981–983. doi:10.1056/NEJMp1714229

Chatterjee, N., Shi, J., and Garcia-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* 17 (7), 392–406. doi:10.1038/nrg.2016.27

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* 15 (141), 20170387. doi:10.1098/rsif.2017.0387

Choi, S. W., Mak, T. S., and O'Reilly, P. F. (2020). Tutorial: A guide to performing polygenic risk score analyses. *Nat. Protoc.* 15 (9), 2759–2772. doi:10.1038/s41596-020-0353-1

Chowdhury, S., Dent, T., Pashayan, N., Hall, A., Lyratzopoulos, G., Hallowell, N., et al. (2013). Incorporating genomics into breast and prostate cancer screening: Assessing the implications. *Genet. Med.* 15 (6), 423–432. doi:10.1038/gim.2012.167

Cowls, J., and Floridi, L. (2018). Prolegomena to a white paper on an ethical framework for a good AI society. SSRN: https://ssrncom/abstract=3198732 2018.

Cui, T., Mekkaoui, K. E., Havulinna, A., Marttinen, P., and Kaski, S. (2021). "Improving neural networks for genotype-phenotype prediction using published summary statistics,". bioRxiv.

Daitch, V., Turjeman, A., Poran, I., Tau, N., Ayalon-Dangur, I., Nashashibi, J., et al. (2022). Underrepresentation of women in randomized controlled trials: A systematic review and meta-analysis. *Trials* 23 (1), 1038. doi:10.1186/s13063-022-07004-2

Diao, J. A., Kohane, I. S., and Manrai, A. K. (2018). Biomedical informatics and machine learning for clinical genomics. *Hum. Mol. Genet.* 27 (R1), R29–R34. doi:10.1093/hmg/ddy088

Dias, R., and Torkamani, A. (2019). Artificial intelligence in clinical and genomic diagnostics. *Genome Med.* 11 (1), 70. doi:10.1186/s13073-019-0689-8

Docherty, A., Kious, B., Brown, T., Francis, L., Stark, L., Keeshin, B., et al. (2021). Ethical concerns relating to genetic risk scores for suicide. *Am. J. Med. Genet. Part B Neuropsychiatric Genet.* 186 (8), 433–444. doi:10.1002/ajmg.b.32871

Domingue, B. W., Trejo, S., Armstrong-Carter, E., and Tucker-Drob, E. M. (2020). Interactions between polygenic scores and environments: Methodological and conceptual challenges. *Sociol. Sci.* 7 (19), 465–486. doi:10.15195/v7.a19

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLOS Genet.* 9 (3), e1003348. doi:10.1371/journal.pgen.1003348

Elgart, M., Lyons, G., Romero-Brufau, S., Kurniansyah, N., Brody, J. A., Guo, X., et al. (2022). Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. *Commun. Biol.* 5 (1), 856. doi:10.1038/s42003-022-03812-z

European Commission (2021). *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts COM/2021/206 final*. European: European Commission.

European Parliament, Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official J. L* 119, 1–88.

Felt, U., Gugglberger, L., and Mager, A. (2009). Shaping the future e-patient: The citizen-patient in public discourse on e-health. *Sci. Technol. Stud.* 22 (1), 24–43. doi:10.23987/sts.55244

Fiske, A., Buyx, A., and Prainsack, B. (2019). Health information counselors: A new profession for the age of big data. *Acad. Med.* 94 (1), 37–41. doi:10.1097/ACM.0000000000002395

Floridi, L., and Cowls, J. (2021). "A unified framework of five principles for AI in society," in *Ethics, governance, and policies in artificial intelligence*. Editor L. Floridi (Cham: Springer International Publishing), 5–17.

Future AI (2022). FUTURE-AI: Best practices for trustworthy AI in medicine. Available at https://future-ai.eu (Accessed December 20, 2022).

Ganna, A., Verweij, K. J. H., Nivard, M. G., Maier, R., Wedow, R., Busch, A. S., et al. (2019). Large-scale GWAS reveals insights into the genetic architecture of same-sex sexual behavior. *Science* 365 (6456), eaat7693. doi:10.1126/science.aat7693

Ge, T., Chen, C. Y., Ni, Y., Feng, Y. A., and Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* 10 (1), 1776. doi:10.1038/s41467-019-09718-5

Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* 349 (6245), 273–278. doi:10.1126/science.aac6076

Gianfrancesco, M. A., Tamang, S., Yazdany, J., and Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern. Med.* 178 (11), 1544–1547. doi:10.1001/jamainternmed.2018.3763

Goodfellow, I., and Yoshua Bengio, A. C. (2016). *Deep learning*. Massachusetts: MIT Press.

Goodman, B., and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Mag.* 38 (3), 50–57. doi:10.1609/aimag.v38i3.2741

High-Level Expert Group on AI (2019). *Ethics guidelines for trustworthy AI*. Brussels: European Commission.

Ho, D. S. W., Schierding, W., Wake, M., Saffery, R., and O'Sullivan, J. (2019). Machine learning SNP based prediction for precision medicine. *Front. Genet.* 10, 267. doi:10.3389/fgene.2019.00267

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Min. Knowl. Discov.* 9 (4), e1312. doi:10.1002/widm.1312

Horst, M., Davies, S. R., and Irwin, A. (2017). "Reframing science communication," in *Handbook of science and technology studies*. Editors U. Felt, R. Fouché, C. A. Miller, and L. Smith-Doerr Fourth Edition (Cambridge/London: MIT Press), 881–907.

Hudec, M., Bednárová, E., and Holzinger, A. (2018). Augmenting statistical data dissemination by short quantified sentences of natural language. *J. Official Statistics* 34 (4), 981–1010. doi:10.2478/jos-2018-0048

Inouye, M., Abraham, G., Nelson, C. P., Wood, A. M., Sweeting, M. J., Dudbridge, F., et al. (2018). Genomic risk prediction of coronary artery disease in 480, 000 adults: Implications for primary prevention. *J. Am. Coll. Cardiol.* 72 (16), 1883–1893. doi:10.1016/j.jacc.2018.07.079

Janssens, A. C. (2019). Validity of polygenic risk scores: Are we measuring what we think we are? . *Hum. Mol. Genet.* 28 (R2), R143–50.

Kerminen, S., Martin, A. R., Koskela, J., Ruotsalainen, S. E., Havulinna, A. S., Surakka, I., et al. (2019). Geographic variation and bias in the polygenic scores of complex diseases and traits in Finland. *Am. J. Hum. Genet.* 104 (6), 1169–1181. doi:10.1016/j.ajhg.2019.05.001

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* 50 (9), 1219–1224. doi:10.1038/s41588-018-0183-z

Khera, A. V., Chaffin, M., Wade, K. H., Zahid, S., Brancale, J., Xia, R., et al. (2019). Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell* 177 (3), 587–596. e9. doi:10.1016/j.cell.2019.03.028

Khera, A. V., Emdin, C. A., Drake, I., Natarajan, P., Bick, A. G., Cook, N. R., et al. (2016). Genetic risk, adherence to a healthy lifestyle, and coronary disease. *N. Engl. J. Med.* 375 (24), 2349–2358. doi:10.1056/NEJMoa1605086

Knoppers, B. M., Bernier, A., Granados Moreno, P., and Pashayan, N. (2021). Of screening, stratification, and scores. *J. Personalized Med.* 11 (8), 736. doi:10.3390/jpm11080736

Kopp, W., Monti, R., Tamburrini, A., Ohler, U., and Akalin, A. (2020). Deep learning for genomics using Janggu. *Nat. Commun.* 11 (1), 3488. doi:10.1038/s41467-020-17155-y

Kozlov, M. (2022). The controversial embryo tests that promise a better baby. *Nature* 609.

Läll, K., Mägi, R., Morris, A., Metspalu, A., and Fischer, K. (2017). Personalized risk prediction for type 2 diabetes: The potential of genetic risk scores. *Genet. Med.* 19 (3), 322–329. doi:10.1038/gim.2016.103

Lambert, S. A., Abraham, G., and Inouye, M. (2019). Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* 28 (R2), R133–R42. doi:10.1093/hmg/ddz187

Lambert, S. A., Gil, L., Jupp, S., Ritchie, S. C., Xu, Y., Buniello, A., et al. (2021). The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* 53 (4), 420–425. doi:10.1038/s41588-021-00783-5

Lee, A., Mavaddat, N., Wilcox, A. N., Cunningham, A. P., Carver, T., Hartley, S., et al. (2019). Boadicea: A comprehensive breast cancer risk prediction modelincorporating genetic and nongenetic risk factors. *Genet. Med.* 21 (8), 1708–1718. doi:10.1038/s41436-018-0406-9

Lewis, A. C. F., and Green, R. C. (2021). Polygenic risk scores in the clinic: New perspectives needed on familiar ethical issues. *Genome Med.* 13 (1), 14. doi:10.1186/s13073-021-00829-7

Lewis, A. C. F., Perez, E. F., Prince, A. E. R., Flaxman, H. R., Gomez, L., Brockman, D. G., et al. (2022). Patient and provider perspectives on polygenic risk scores: Implications for clinical reporting and utilization. *Genome Med.* 14 (1), 114. doi:10.1186/s13073-022-01117-8

Lewis, C. M., and Vassos, E. (2020). Polygenic risk scores: From research tools to clinical instruments. *Genome Med.* 12 (1), 44. doi:10.1186/s13073-020-00742-5

Libbrecht, M. W., and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16 (6), 321–332. doi:10.1038/nrg3920

Loh, P-R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., et al. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* 47 (12), 1385–1392. doi:10.1038/ng.3431

Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., and Sham, P. C. (2017). Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* 41 (6), 469–480. doi:10.1002/gepi.22050

Markus, A. F., Kors, J. A., and Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J. Biomed. Inf.* 113, 103655. doi:10.1016/j.jbi.2020.103655

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019a). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 51 (4), 584–591. doi:10.1038/s41588-019-0379-x

Martin, A. R., Daly, M. J., Robinson, E. B., Hyman, S. E., and Neale, B. M. (2019b). Predicting polygenic risk of psychiatric disorders. *Biol. Psychiatry* 86 (2), 97–109. doi:10.1016/j.biopsych.2018.12.015

Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., et al. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am. J. Hum. Genet.* 104, 21–34.

McInnes, G., Sharo, A. G., Koleske, M. L., Brown, J. E. H., Norstad, M., Adhikari, A. N., et al. (2021). Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. *Am. J. Hum. Genet.* 108 (4), 535–548. doi:10.1016/j.ajhg.2021.03.003

McLennan, S., Fiske, A., Tigard, D., Müller, R., Haddadin, S., and Buyx, A. (2022). Embedded ethics: A proposal for integrating ethics into the development of medical AI. *BMC Med. Ethics* 23 (1), 6. doi:10.1186/s12910-022-00746-3

Meisner, A., Kundu, P., Zhang, Y. D., Lan, L. V., Kim, S., Ghandwani, D., et al. (2020). Combined utility of 25 disease and risk factor polygenic risk scores for stratifying risk of all-cause mortality. *Am. J. Hum. Genet.* 107 (3), 418–431. doi:10.1016/j.ajhg.2020.07.002

Meskó, B., Drobni, Z., Bényei, É., Gergely, B., and Győrffy, Z. (2017). Digital health is a cultural transformation of traditional healthcare. *mHealth* 3 (9), 38. doi:10.21037/mhealth.2017.08.07

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* 6, 26094. doi:10.1038/srep26094

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. *Briefings Bioinforma.* 19 (6), 1236–1246. doi:10.1093/bib/bbx044

Mitchell, B. L., Thorp, J. G., Wu, Y., Campos, A. I., Nyholt, D. R., Gordon, S. D., et al. (2021). Polygenic risk scores derived from varying definitions of depression and risk of depression. *JAMA Psychiatry* 78 (10), 1152–1160. doi:10.1001/jamapsychiatry.2021.1988

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Soc.* 3 (2), 205395171667967. doi:10.1177/2053951716679679

Moorthie, S., Hall, A., Janus, J., Brigden, T., Babb de Villiers, C., Blackburn, L., et al. (2021). *Polygenic scores and clinical utility*. Cambridge: PHG Foundation.

Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., et al. (2020). The ethics of AI in health care: A mapping review. *Soc. Sci. Med.* 260, 113172. doi:10.1016/j.socscimed.2020.113172

Murdoch, B. (2021). Privacy and artificial intelligence: Challenges for protecting health information in a new era. *BMC Med. Ethics* 22 (1), 122. doi:10.1186/s12910-021-00687-3

Murray, G. K., Lin, T., Austin, J., McGrath, J. J., Hickie, I. B., and Wray, N. R. (2021). Could polygenic risk scores be useful in psychiatry?: A review. *JAMA Psychiatry* 78 (2), 210–219. doi:10.1001/jamapsychiatry.2020.3042

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366 (6464), 447–453. doi:10.1126/science.aax2342

Palk, A. C., Dalvie, S., de Vries, J., Martin, A. R., and Stein, D. J. (2019). Potential use of clinical polygenic risk scores in psychiatry – ethical implications and communicating high polygenic risk. *Philosophy, Ethics, Humanit. Med.* 14 (1), 4. doi:10.1186/s13010-019-0073-8

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. 2 ed. Cambridge: Cambridge University Press.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. Cambridge: The MIT Press.

Pivovarov, R., Albers, D. J., Sepulveda, J. L., and Elhadad, N. (2014). Identifying and mitigating biases in EHR laboratory tests. *J. Biomed. Inf.* 51, 24–34. doi:10.1016/j.jbi.2014.03.016

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460 (7256), 748–752. doi:10.1038/nature08185

Ras, G., van Gerven, M., Haselager, P., et al. (2018). "Explanation methods in deep learning: Users, values, concerns and challenges," in *Explainable and interpretable models in computer vision and machine learning*. Editors H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, and U. Güçlü (Cham: Springer International Publishing), 19–36.

Reid, N. J., Brockman, D. G., Elisabeth Leonard, C., Pelletier, R., and Khera, A. V. (2021). Concordance of a high polygenic score among relatives: Implications for genetic counseling and cascade screening. *Circ. Genom Precis. Med.* 14 (2), e003262. doi:10.1161/CIRCGEN.120.003262

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?," *Explaining the predictions of any classifier* in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, August 2016.

Rigby, M. J. (2019). Ethical dimensions of using artificial intelligence in health care. *AMA J. Ethics* 21 (2), E121–E124.

Roth, A., and Dwork, C. (2013). The algorithmic foundations of differential privacy. *Found. Trends® Theor. Comput. Sci.* 9 (3-4), 211–407. doi:10.1561/0400000042

Rothstein, M. A. (2015). Ethical issues in big data health research: Currents in contemporary bioethics. *J. Law Med. Ethics* 43 (2), 425–429. doi:10.1111/jlme.12258

Shen, J., Zhang, C. J. P., Jiang, B., Chen, J., Song, J., Liu, Z., et al. (2019). Artificial intelligence versus clinicians in disease diagnosis: Systematic review. *JMIR Med. Inf.* 7 (3), e10010. doi:10.2196/10010

Simmons, S., Sahinalp, C., and Berger, B. (2016). Enabling privacy-preserving GWASs in heterogeneous human populations. *Cell Syst.* 3 (1), 54–61. doi:10.1016/j.cels.2016.04.013

Sirugo, G., Williams, S. M., and Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell* 177 (1), 26–31. doi:10.1016/j.cell.2019.02.048

Slunecka, J. L., van der Zee, M. D., Beck, J. J., Johnson, B. N., Finnicum, C. T., Pool, R., et al. (2021). Implementation and implications for polygenic risk scores in healthcare. *Hum. Genomics* 15 (1), 46. doi:10.1186/s40246-021-00339-y

Sohail, M., Maier, R. M., Ganna, A., Bloemendal, A., Martin, A. R., Turchin, M. C., et al. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife* 8, e39702. doi:10.7554/eLife.39702

Sud, A., Turnbull, C., and Houlston, R. (2021). Will polygenic risk scores for cancer ever be clinically useful? *npj Precis. Oncol.* 5 (1), 40. doi:10.1038/s41698-021-00176-1

Tellier, L. C. A. M., Eccles, J., Treff, N. R., Lello, L., Fishel, S., and Hsu, S. (2021). Embryo screening for polygenic disease risk: Recent advances and ethical considerations. *Genes* 12 (8), 1105. doi:10.3390/genes12081105

Torkamani, A., Wineinger, N. E., and Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* 19 (9), 581–590. doi:10.1038/s41576-018-0018-x

Treff, N. R., Eccles, J., Lello, L., Bechor, E., Hsu, J., Plunkett, K., et al. (2019). Utility and first clinical application of screening embryos for polygenic disease risk reduction. *Front. Endocrinol.* 10, 845. doi:10.3389/fendo.2019.00845

Trubetskoy, V., Pardinas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T. B., et al. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* 604, 502–508. doi:10.1038/s41586-022-04434-5

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., et al. (2021). *The ethics of algorithms: Key problems and solutions*. New York: AI & Society.

Turley, P., Meyer, M. N., Wang, N., Cesarini, D., Hammonds, E., Martin, A. R., et al. (2021). Problems with using polygenic scores to select embryos. *N. Engl. J. Med.* 385 (1), 78–86. doi:10.1056/NEJMsr2105065

Vayena, E., Blasimme, A., and Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLOS Med.* 15 (11), e1002689. doi:10.1371/journal.pmed.1002689

Vilhjalmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindstrom, S., Ripke, S., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 97 (4), 576–592. doi:10.1016/j.ajhg.2015.09.001

Wald, N. J., and Old, R. (2019). The illusion of polygenic disease risk prediction. *Genet. Med.* 21 (8), 1705–1707. doi:10.1038/s41436-018-0418-5

Wand, H., Lambert, S. A., Tamburro, C., Iacocca, M. A., O'Sullivan, J. W., Sillari, C., et al. (2021). Improving reporting standards for polygenic scores in risk prediction studies. *Nature* 591 (7849), 211–219. doi:10.1038/s41586-021-03243-6

Wang, Y., Tsuo, K., Kanai, M., Neale, B. M., and Martin, A. R. (2022). Challenges and opportunities for developing more generalizable polygenic risk scores. *Annu. Rev. Biomed. Data Sci.* 5, 293–320. doi:10.1146/annurev-biodatasci-111721-074830

Whittlestone, J., Nyrup, R., Alexandrova, A., and Cave, S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions, in the 2019 AAAI/ACM Conference, Honolulu, HI, USA, January 27–28, 2019.

Widen, E., Junna, N., Ruotsalainen, S., Surakka, I., Mars, N., Ripatti, P., et al. (2022). How communicating polygenic and clinical risk for atherosclerotic cardiovascular disease impacts health behavior: An observational follow-up study. *Circ. Genom Precis. Med.* 15 (2), e003459. doi:10.1161/CIRCGEN.121.003459

Wynne, B. (1993). Public uptake of science: A case for institutional reflexivity. *Public Underst. Sci.* 2 (4), 321–337. doi:10.1088/0963-6625/2/4/003

Xu, J., Yang, P., Xue, S., Sharma, B., Sanchez-Martin, M., Wang, F., et al. (2019). Translating cancer genomics into precision medicine with artificial intelligence: Applications, challenges and future perspectives. *Hum. Genet.* 138 (2), 109–124. doi:10.1007/s00439-019-01970-5

Zhang, T., He, C., Ma, T., Gao, L., Ma, M., and Avestimehr, S. (2021). "Federated learning for internet of things," in Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems, Coimbra, Portugal, 15 November 2021.

frontiers | Frontiers in Genetics

TYPE Review
PUBLISHED 13 March 2023
DOI 10.3389/fgene.2023.1086802

Check for updates

OPEN ACCESS

EDITED BY
Gabriele Werner-Felmayer,
Innsbruck Medical University, Austria

REVIEWED BY
Kay A. Robbins,
University of Texas at San Antonio, United States
Stephan Heunis,
Research Center Jülich, Germany

*CORRESPONDENCE
Aaron Reer,
aaron.reer@uol.de

SPECIALTY SECTION
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Genetics

RECEIVED 01 November 2022
ACCEPTED 21 February 2023
PUBLISHED 13 March 2023

CITATION
Reer A, Wiebe A, Wang X and Rieger JW (2023), FAIR human neuroscientific data sharing to advance AI driven research and applications: Legal frameworks and missing metadata standards. *Front. Genet.* 14:1086802. doi: 10.3389/fgene.2023.1086802

COPYRIGHT
© 2023 Reer, Wiebe, Wang and Rieger. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

FAIR human neuroscientific data sharing to advance AI driven research and applications: Legal frameworks and missing metadata standards

Aaron Reer[1]*, Andreas Wiebe[2], Xu Wang[2] and Jochem W. Rieger[1]

[1]Applied Neurocognitive Psychology Lab, Institute for Medicine and Healthcare, Department of Psychology, Oldenburg University, Oldenburg, Germany, [2]Chair for Intellectual Property and Information Law, Göttingen University, Göttingen, Germany

Modern AI supported research holds many promises for basic and applied science. However, the application of AI methods is often limited because most labs cannot, on their own, acquire large and diverse datasets, which are best for training these methods. Data sharing and open science initiatives promise some relief to the problem, but only if the data are provided in a usable way. The FAIR principles state very general requirements for useful data sharing: they should be findable, accessible, interoperable, and reusable. This article will focus on two challenges to implement the FAIR framework for human neuroscience data. On the one hand, human data can fall under special legal protection. The legal frameworks regulating how and what data can be openly shared differ greatly across countries which can complicate data sharing or even discourage researchers from doing so. Moreover, openly accessible data require standardization of data and metadata organization and annotation in order to become interpretable and useful. This article briefly introduces open neuroscience initiatives that support the implementation of the FAIR principles. It then reviews legal frameworks, their consequences for accessibility of human neuroscientific data and some ethical implications. We hope this comparison of legal jurisdictions helps to elucidate that some alleged obstacles for data sharing only require an adaptation of procedures but help to protect the privacy of our most generous donors to research . . . our study participants. Finally, it elaborates on the problem of missing standards for metadata annotation and introduces initiatives that aim at developing tools to make neuroscientific data acquisition and analysis pipelines FAIR by design. While the paper focuses on making human neuroscience data useful for data-intensive AI the general considerations hold for other fields where large amounts of openly available human data would be helpful.

KEYWORDS

data sharing and re-use, machine learning, open science, FAIR (findable accessible interoperable and reusable) principles, neuroimaging, metadata standards, data protection, privacy law

1 Introduction

Making data publicly available is considered beneficial for scientific research in many respects including improving reliability of results by increasing transparency and quality, increasing efficiency of the (public) money spent, accelerating innovation by enhancing interdisciplinarity, and enabling the use and development of new analysis techniques (Milham et al., 2018; Niso et al., 2022). For these reasons, opening up the currently mostly closed scientific research, e.g., by encouraging data sharing (among other research products and practices) is one of today's pressing issues.

Replication and reproducibility issues recently elicited growing concerns in the biomedical and life sciences over the credibility of claims raised in scientific studies and the economic efficiency of research. The OPEN SCIENCE COLLABORATION (2015) tried to replicate 100 highly influential studies published in top-tier psychology journals and found that in only 36% of these studies statistical significance of the results could be reproduced. Glasziou and Chalmers (2018) argue that due to fundamental deficiencies in the design and conduct of studies in clinical research, globally around 85% of the money being spent is wasted because many findings cannot be reproduced, nor can the respective data be re-used. Moreover, the authors concluded that many findings can or should not be implemented into practice due to their low reliability. Similarly, a meta-analysis of past studies on the cost of non-reproducible research has revealed that in the US over 50% of the preclinical research cannot be reproduced and therefore complicates cumulative knowledge acquisition (Freedman et al., 2015). According to the authors, this amounts to approximately 28 billion US dollars per year being misspent in the US alone. Today a common notion is that, among others, open sharing of data and research products is one important measure to make research more efficient in its resource use (Niso et al., 2022). The 2020 EU scoping report on "reproducibility of scientific results in the EU" (Europäische Kommission et al., 2020) and the 2019 report of the US National Academies of Sciences, Engineering, and Medicine on "Reproducibility and replicability in science" (National Academies of Sciences, 2019) list, among others, data sharing as one important scientific practice to enhance reproducibility and replicability. This includes, training for data sharing, the establishment and improvement of data sharing plans in publicly funded research, support for data sharing, the resolution of data sharing problems, and FAIRification of shared data. Sharing is also considered a measure to trigger a change in scientific practice from closed research to open sharing of research products to increase the quality and transparency of research practices.

One way to estimate an increase in efficiency of resource use by data sharing is to estimate potential monetary savings. This is of relevance as most research in public institutions is financed by public money. Employing a bibliometric analysis of the re-use of five openly shared large scale neuroimaging datasets provided by the International Neuroimaging Data-sharing Initiative (INDI, Mennes et al., 2013) Milham and others estimate savings of 900 million up to 1.7 billion US dollars compared to re-acquisition of the data for each of the approximately 900 papers published on the basis of these datasets (Milham et al., 2018). Likewise, the European Commission has issued a report in 2019 suggesting that better research data management would save 10.2 billion euros per year in Europe (European Commission and Directorate-General for Research and Innovation, 2019). They even argue that potentially the gain would be even bigger (up to an additional 16 billion euros) due to the generated innovation, e.g., faster accumulation of knowledge and potential savings of money spent on data acquisition.

Beyond improving credibility, reliability, and efficiency of research, individual researchers may personally benefit as well from sharing their data. Data sharing can increase their visibility and reputation by licensing the data and making it a citable object. This offers new opportunities for publications and can increase the number of citations, raise media attention, open new collaborations with researchers who do not belong to the narrow group of the individual research field, and finally can offer new funding and position opportunities (Markowetz, 2015; McKiernan et al., 2016; Allen and Mehler, 2019; Hunt, 2019; Niso et al., 2022; Nosek et al., 2022). However, it is important to note, that practices such as data sharing or proper description of the data through metadata imposes additional work for the individual researcher. For this reason, it is important to facilitate the implementation of these practices into workflows in order to lift some weight off the shoulders of the individual researcher. In other words, usability must be a critical aspect of tools for sharing or organizing data.

In the light of the issues with closed research and the potential advantages of sharing data and other research products the general reluctance of researchers to share their data appears surprising (Houtkoop et al., 2018). Recently, however, the importance of data sharing and research data management (RDM) moved from a small community of open science enthusiasts into the focus of funding agencies and journals as policy reinforcers to address these issues. Funding agencies are beginning to implement a top-down strategy for publicly funded research to expand data sharing for more efficient data use and accessibility of research results (de Jonge et al., 2021; Niso et al., 2022). Some funding agencies require RDM plans, openly accessible publications, and dissemination plans beyond journal publications. In addition, an increasing number of journals offer open access options and require authors to make their data publicly available (Niso et al., 2022). In parallel stakeholder institutions like the Organization for Human Brain Mapping (OHBM)[1], the International Neuroinformatics Coordinating Facility (INCF)[2] and the Chinese Open Science Network (COSN)[3], coordinate the development of data standards and best practices for open and FAIR research data management.

Data sharing in standardized data formats and enriched with metadata are important requirements for novel data-driven Artificial Intelligence (AI) analysis techniques. AI technologies are expected to propel and transform scientific research in the near future and are meanwhile key technologies in medical research, diagnostic procedures, etc. They learn generalizable structure in complex data which is unrecognizable to humans and make it possible to predict e.g., disease risks or cognitive

1 https://www.humanbrainmapping.org/i4a/pages/index.cfm?pageid=1; last accessed: 26.10.22.

2 https://www.incf.org/; last accessed: 26.10.22.

3 https://open-sci.cn/; last accessed: 26.10.22.

functions in new data. This development is supported by the increasing capacity of computing machines that enable more complex computations on ever-growing data sets. However, many AI algorithms estimate extremely complex models from the data. This requires huge amounts of data. The limited amount of available data within single labs (with known data structure and metadata) and the limited amount of well structured, meta-data annotated, and exhaustively documented publicly available data is a common bottleneck for the reliable application of complex but powerful AI methods. Therefore, data sharing and making experimental data interoperable (i.e., common data and annotation standards help to make the data computer interoperable) has become an important goal for the neuroimaging community.

Several fields in life science and medicine have recognized the potential of publicly available data early and started large scale initiatives to make data collected in individual labs accessible for other research groups in order to maximize the scientific benefits. The forerunners were the Human Genome Project[4], launched in 1990, in which the Bermuda principles were developed. These required the timely sharing of annotated sequence data (Collins et al., 2003). This policy initially boosted progress in genomic research and in related fields such as computer science and AI based data analysis (Rood and Regev, 2021). Hence it fostered interdisciplinary research approaches, digitization of life science research and the development of novel analysis methods (Gibbs, 2020). Over the years, the increase in size and complexity of available data, the lack of data standards, the scattering of data across various databases, and data privacy issues, in particular when the genetic data were enriched with "phenotype" metadata, have triggered a re-thinking of the current relatively unstructured sharing approach. This re-thinking was mainly due to the fact that it became more and more evident that this unstructured approach likely has a negative impact on the usability and usefulness of the shared data in current and future usage scenarios (Powell, 2021). Moreover, while the domain of genetics developed a relatively generous and open data culture, recent developments indicate a return to closed data policies with reluctance to share data or only under certain conditions, for example, data sharing policies in the commercial sector and in China (Koch and Todd, 2018; Chen and Song, 2018; PIPL Art. 38–43&53). This closed policy cuts international public genetics research off from huge data sources. In neuroscience, the later funded Human Connectome Project (launched in 2009) and the EU Human Brain Project (launched in 2013) also collect massive amounts of complex datasets consisting of diverse data types (e.g., brain imaging data recorded with different measuring techniques or devices, behavioral data, data about the experimental paradigm, genetic data, bio samples, clinical diagnostics, psychological testing, etc.). This was done to provide datasets, that enable tackling a range of research questions by different researchers, even questions unrelated to the original study. In general, acquiring more diverse data in an experiment, exceeding those needed for the original research question, would help to increase the efficiency of data re-use. While some efforts have been made to create publicly open databases to make the data accessible, common standards on how to store such datasets are only emerging (e.g., Teeters et al., 2015; Gorgolewski et al., 2016).

Publicly open databases which contain well described and standardized datasets help to make the data better understandable not only for humans but also for computers. Accordingly, such datasets can serve as training data for the development of new analysis approaches but also as realistic benchmark datasets to compare the performance of novel AI algorithms. Well-structured data enhanced with metadata and many accessory observational data are also attractive for researchers who have no access to the expensive experimental infrastructure, be they from different fields, like computational neuroscientists, developers of AI algorithms or experts from countries or research sites with less financial resources. In these cases, data sharing can make science more interdisciplinary and diverse by adding hitherto excluded modelers, methods developers, and researchers without access to neuroimaging resources to a research community.

In sum, data sharing offers benefits for the individual researcher as well as research communities besides improving transdisciplinary integration of research and thereby enhancing its development. So why is so little of the myriads of data produced in biomedical and life science publicly shared (Houtkoop et al., 2018)? There are many possible reasons, ranging from motivation and literacy to infrastructural problems at the level of FAIRification as well as legal and ethical issues, that create uncertainty under which conditions human research data can be shared and with whom (Paret et al., 2022). In this paper we will focus on two related issues. First, we want to outline the heterogeneous legal frameworks with respect to data privacy in different geopolitical zones. The focus of this analysis will be on comparing the goals of the frameworks and to explicate the constraints they impose on sharing of sensitive human data. Second, we discuss approaches for data organization and metadata annotation in the domain of neuroscience. In other words, standardized vocabularies or ontologies for turning data into meaningful and interpretable information. Finally, we will highlight initiatives and tools, that were developed to help the individual researcher to practically implement data sharing into their workflows.

2 Challenges for useful data sharing

Although data sharing is generally regarded as a good and desirable practice, it creates technical as well as ethical and legal challenges. Depending on how well these are met, the effects of data sharing can range from useful to harmful. As always, a good intent does not guarantee a good deed. Two big challenges to the useful sharing of human neuroimaging data will be highlighted in the following.

A first challenge for sharing data from humans arises when they include personal data or become personalizable (e.g., when biometric data such as genetic information or pictures of a person are included). Then legal and ethical restrictions may require higher control levels for data sharing. Complications arise from the fact that legal and ethical data protection levels vastly differ between states and cultures around the world and that it is often

4 https://www.genome.gov/human-genome-project; last accessed: 26.10.22.

unclear what combination of features can make the data personalizable. A second challenge arises from the fact that data from experiments with human participants are oftentimes complex, leaving the experimenter a lot of freedom with respect to organizing and describing them. Then metadata, data describing data, is required to make the data useful and interpretable for other researchers or automated analysis pipelines. We call this the metadata description challenge.

It should be noted that technical challenges like provision, maintenance and setting up of databases, and the technical implementation of safeguards for these repositories etc. is not in the scope of the paper. Moreover, in this article we focus on neuroimaging data. However, some of the points discussed, in particular the legal frameworks, also apply to other types of human data, like genetic data.

2.1 Legal and ethical frameworks around data sharing

Privacy issues can arise when the human neuroimaging data allow for re-identification of the person from whom they were recorded. By re-identification we mean, that the data may provide information, that makes it possible to tell whether it was recorded from let's say Jane Roe or Henry Wade. For example, anatomical magnetic-resonance-imaging (MRI) scans can contain an image of the face which might allow for re-identification of the person. Schwarz and others demonstrated that individual subjects could be re-identified by matching the faces reconstructed from MR-scans with pictures from the subjects that originated from social media (Schwarz et al., 2019). Another study showed that blurring the face in the MRI-scan may not be sufficient to prevent re-identification. Using Generative Adversarial Networks, it was shown that blurred faces could be reasonably well reconstructed to allow for re-identification. However, completely removing the facial features from the anatomical MRI scans greatly reduced the success of the method (Abramian and Eklund, 2019). In the field of neuroimaging this debate is most relevant for high resolution structural imaging techniques that can provide anatomical images, such as certain magnetic resonance imaging techniques. Electrophysiological recordings, such as EEG and MEG, or fNIRS do not provide detailed anatomical information. For that reason, data from these devices are less likely to be re-identifiable (Jwa and Poldrack, 2022; White et al., 2022). There is an ongoing discussion, however, to what extent neuroimaging data in general contain individual signatures, similar to genetic data. It has thus been suggested to consider them as a kind of biometric data, i.e., some data that is not alone identifiable but sufficient to single out data from an unidentified individual X in a group of datasets (Bannier et al., 2021). Whether the fear that human neurophysiological data allow for direct re-identification or singling out and subsequent identification of an individual is in general realistic or whether these are overly conservative assumptions still needs to be shown (Jwa and Poldrack, 2022). Moreover, it is not clear how future developments like increasing data availability, complexity, and progress in AI techniques contribute to the problem.

Such considerations are necessary because privacy and data protection laws across jurisdictions offer protection against processing of information from which a person may be re-identified. Therefore, a basic prerequisite of shared neuroimaging data and accompanying metadata is that the natural person from whom it was recorded cannot be re-identified. This can create a tension between the desire to have rich datasets with lots of metadata describing the individual (including phenotypic data), and privacy protection. Potential privacy breaches can have different consequences in different legal, ethical and cultural regions because data privacy and data protection is weighted very differently across regions and respective jurisdictions. This can make sharing of neuroimaging but also other "biometric" or "identifiable" data across borders very difficult (Eke et al., 2022). The main existing legal frameworks appear to revolve around three agents: a natural person who donates data, private institutions with commercial interests in the donated data, and governmental institutions with various goals regarding the data. Below, we provide an overview of the laws/regulations from three representative jurisdictions EU, United States, and China. These are the regions with the largest data resources and they span a spectrum of regulatory frameworks in which the different and potentially conflicting interests of the three agents are balanced and weighted differently. Readers looking for a quick overview of the regulations relevant for sharing human research data internationally can refer to Table 1. Details and sources to each point are provided in the text. A list of points researchers should consider when planning to acquire data for sharing is provided in Table 2 at the end of the chapter.

We would like to point out that the following section only provides an informative overview with pointers to regulations we considered relevant for the comparison. They should not be considered as legal advice.

2.1.1 The European Union

In the EU the General Data Protection Regulation (GDPR) came into effect in 2018. The GDPR was a major step to harmonize the legal regulations for acquisition, processing, and sharing of personal data across the jurisdictions of the Member States. This was necessary to ensure free movement of data between EU member states and states providing comparable data protection. It is based on *codified legal principles* relating to the protection of personality and most parts were implemented in the laws of the EU member states before. It is directly applicable as a regulation in all Member States without the need for further national implementation. However, there are supplementary national and local regulations specifying the rules and the GDPR is open for deviating national legislation in some cases. In a sense the GDPR follows the European tradition of the enlightenment as it aims to put the individual at the center and it follows the tradition of civil law. One motivation contributing to the design of the GDPR was to empower the individual against economic interests of companies which often consider the acquired data as their property which they can use without further accountability. The examples for questionable or unethical acquisition, use and (not-) sharing of user data by the big tech companies are legion (e.g., The European Data Protection Supervisor (EDPS)[5], 2020; Koch and Todd, 2018; Kurtz et al., 2022;

5 European Data Protection Supervisor. (2020). A Preliminary Opinion on Data Protection and Scientific Research. URL: https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf; last accessed: 25.10.2022.

TABLE 1 Overview of data protection regulations for publicly funded research.

Aspect	EU	United States	China
Relevant laws and regulations	GDPR and local data protection laws as instances of it.	Dependent on applicable regulator: e.g., HIPAA, Common Rule, special rules.	CSL, DSL, PIPL, CCC, and field specific regulations by MOST
What is protected	All processing of identifiable, pseudonymized, or special personal data. Anonymized data are exempted.	Common Rule: only identifiable private information collected during research.	Personal information (includes sensitive information, such as biometric and medical health data) and data sovereignty of China.
		HIPAA: personal health related data.	
What is personal information	Any information related to identified or identifiable natural person	Common Rule: Information from which the identity of the subject be readily ascertained	Information related to identified or identifiable natural persons.
		HIPAA: individually identifiable health information.	
Measures of responsible person (e.g., researchers) to protect personal data	Pseudonymization (e.g., replace identifiable information with code), anonymization (no identifiable and sensitive information)	Common Rule: Unclear	De-identification, anonymization (impossible to identify person)
		HIPAA: de-identification e.g., by Safe Harbor Method (similar to pseudonymization) and DUA	
Consent required for	All processing (here sharing) of personal, pseudonymized and sensitive data. Extended consent possible. For non-sensitive data other legal grounds Art. 6 I	Common Rule: Broad consent. Secondary use without consent.	Separate consent for different processing purposes. Sensitive data additionally require special purpose. New purposes require new consent.
		HIPAA: written informed consent for data sharing or DUA	
IRB required	Yes and legal assessment	Yes	IRB not mentioned. Sharing might be restricted by state institutions (e.g., genetic data).
With whom can adequately protected data be shared	Researchers in EU and adequacy region. Consent and DUA may allow widening scope.	Common Rule: policy evolving. HIPAA rules sufficient.	Sharing outside mainland requires several safety measures and local safeguard.
		HIPAA: With consent and/or DUA no restriction.	

Spector-Bagdady, 2021). In addition, the opaque handling of collected data, research practices and goals, created suspicions that these practices raise barriers for research and that egoistic economic goals of research can severely conflict with the interests of the individual as well as society. We will briefly review the regulations relevant for scientific data sharing in the following sections.

The GDPR's enormous impact is due to the broad scope that reaches beyond institutions established in the EU. It applies to any processing (e.g., analysis and sharing) of personal data in the context of the activities of a *data controller* (person who has control over the data) or a *data processor* (person who processes the data), regardless of whether these activities take place in the Union or not (GDPR Art. 3 (1)). The GDPR also restricts collection and processing of personal data by states. In short, the GDPR provides regulations for the protection of personal data of natural persons by establishing binding principles (e.g., transparency, purpose limitation and data minimization, GDPR Art. 5) and by defining a set of lawful processing purposes (GDPR Art. 6). One way to implement legal processing is to obtain consent from the person who donates data (data subject). The GDPR also defines rights of data subjects (GDPR Art. 12–23), and mechanisms to enforce their rights (GDPR Art. 77–84).

The GDPR defines *personal data* broadly as "any information relating to an identified or identifiable natural person", the data subject (GDPR Art. 4 (5)). One measure to protect personal data is to *pseudonymize* it (GDPR Art. 25), meaning that the data are processed in a way that they cannot be directly related to the data subject. This can be achieved, for example, by separating all personal information, that would allow re-identification, e.g., data to handle the compensation for participation like name, address, bank account etc., from the data to be processed. The link between data and personal information is stored in a coding list which is kept separate from the data. Pseudonymization is a safeguard for sharing that is provided in other regulations too (see Sections 2.1.2, 2.1.3) and in practice most neuroimaging labs already implement such a policy. Moreover, the coding list would be in the hands of the data controller, who determines the means and purposes of the data processing, but it would not be accessible to the data processor. This is not always possible, e.g., when the data controller and the data processor are the same person. However, there are ways to deal with such problems, e.g., by handing coding list and personal information to another trustworthy person. Importantly, the coding list must not be shared and a third-party data processor must not gain access to the content of the coding list. It should be noted that pseudonymized data are still in the scope of the GDPR as they can be associated with

a data subject by means of other information (e.g., the coding list, Recital 26). Conversely, *anonymous* data, which cannot be related to a natural person, is not covered by the GDPR (Recital 26), meaning that processing of anonymized data is outside the scope of the GDPR. However, this is not true for the processing up to the point of anonymization, for which a legal ground is still necessary. The GDPR does not suppose that means for personal data protection are perfect and unbreakable. Therefore, it adopts a risk-based protection assessment. The risk of re-identification or other misuse of the data should be minimized by considering state-of-the-art technology, but the data controller should also consider the costs for protection and re-identification, as well as the likelihood and severity of risks arising for the natural person from re-identification (GDPR Art. 25, 32, Recital 26).

As the GDPR promotes privacy by design and default (GDPR Art. 25) it has been argued that personal data cannot be shared with other researchers under the GDPR and that the GDPR therefore poses an obstacle for free international dataflow and hence scientific research (Eke et al., 2022). Unfortunately, this is a widely adopted misconception. The GDPR weights the value of scientific research and offers a range of derogations from the strict protection of personal data for *scientific research and academic expression* (GDPR Art. 85, Art. 5 (1) (b), (e)). However, some safeguards (GDPR Art. 89) must be met. The European Data Protection Supervisor (EDPS 2020)[6] lists transparency and being in the public interest as central features of scientific research. Moreover, the safeguards that need to be implemented include explicit informed consent to the sharing of personal data and independent ethical oversight, e.g., by an ethics committee. Personal data can be processed to make them suitable for archiving in public interest, meaning they can be pseudonymized and made available in research data repositories in pseudonymized form. Moreover, the data can be processed for scientific, historical, and statistical purposes (GDPR Art. 89) and for other purposes than those for which they were initially collected if consent was collected and recognized ethical standards for scientific research are met (GDPR Recital 33, 50). Privacy by design and default can be supported by Codes of Conduct like the "Code of Conduct on privacy for mobile health applications"[7] though that has not yet been adopted. Moreover, the position paper "A preliminary opinion on data protection and scientific research." by the EDPS (2020)[8] provides some advice for the interpretation of the GDPR in that respect.

The GDPR, like other regulations (see Sections 2.1.2, 2.1.3), puts particularly strong restrictions on the processing of *special categories of data* (e.g., health data or biometric data, Art. 9 GDPR). Some processing purposes are allowed and explicit consent for the processing is required (Art. 9 GDPR). But the GDPR also acknowledges the importance of science and research for society and provides some privileges for research purposes, to balance research with the rights of the individual (see Wiebe, 2020) and permits derogations to the prohibition of the processing of special data in accordance with GDPR Art. 89. Of special relevance are processing permissions for scientific purposes in Art. 9 (2) (j) GDPR that are specified by national legislation. For example, in Germany, the weighing of interests is a prerequisite for lawful processing (§ 27 German Data Protection Statute, BDSG). Article 7 (2) (h) of GDPR defines permissions for medical and (public) health related processing. However, specific measures to safeguard the fundamental rights and interests of the data subject must be implemented. For neuroimaging data sharing, explicit consent, mechanisms for access control and contracts in the form of data use agreements have been suggested (Bannier et al., 2021; Staunton et al., 2022). The exact scope of these permission with respect to the development and use of AI systems in the health sector has still to be developed, in connection with appropriate safeguards.

In the context of scientific research, data fulfilling the outlined requirements of the GDPR can and should be freely exchanged between researchers in the EU member states and those states with an adequacy decision, which means that they are recognized to offer data protection at a comparable level as the GDPR (see here[9] for a list of countries for which such adequacy decisions have been made). The Data Governance Act2022[10] seeks to enhance *data sharing* by removing technical and organizational obstacles to data sharing and provide a secure infrastructure for data sharing within the EU. It includes the promotion of the development of data intermediation services and the development of arrangements to facilitate data use on altruistic grounds, i.e., to make data available voluntarily, without reward, to be used in the public interest. E.g., Art. 25 of the Data Governance Act foresees the development of a European data altruism consent form, which shall allow the collection of consent or permission across Member States in a uniform format. Moreover, the European Commission issued plans to build a European Health Data Space which provides individual persons control over their health data in concordance with the GDPR (Directorate-General for Health and Food Safety, 2022)[11]. However, currently, due to a very restrictive decision of the European Court of Justice[12], transfer of personal data to third countries are very difficult to pursue lawfully with very high requirements on safeguards in the target country and their practical effectiveness. This applies to each country for which no adequacy decision, as stated above, exists, including countries like China and the United States. On the political level, efforts are underway to establish a renewed "safe harbor" for transfers to

6 European Data Protection Supervisor. (2020). A Preliminary Opinion on Data Protection and Scientific Research. URL: https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf; last accessed: 25.10.2022.

7 https://digital-strategy.ec.europa.eu/en/policies/privacy-mobile-health-apps; last accessed: 25.10.22.

8 European Data Protection Supervisor. (2020). A Preliminary Opinion on Data Protection and Scientific Research. URL: https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf; last accessed: 25.10.2022.

9 https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en; last accessed: 24.10.2022.

10 http://data.europa.eu/eli/reg/2022/868/oj; last accessed: 25.10.22.

11 https://health.ec.europa.eu/publications/communication-commission-european-health-data-space-harnessing-power-health-data-people-patients-and_en; last accessed: 21.10.2022.

12 https://www.europarl.europa.eu/RegData/etudes/ATAG/2020/652073/EPRS_ATA(2020)652073_EN.pdf; last accessed: 28.10.2022.

the U.S. However, this does not mean that personal or special category data cannot be shared with researchers in countries outside the EU or if no adequacy decision exists. For the transfer of special data to a third country without adequacy decision, explicit consent to the transfer by the data subject can be a potential legal basis if the transfer is not done on a regular basis (GDPR Art 49; EDPS, 2020[13]).

In sum, the GDPR defines a legal framework for the processing and transfer of personal data that aims to protect the individual and harmonize the legal frameworks across member states in order to simplify privacy protection and data exchange between states. It establishes as world-wide "gold standard" and serves as a blue print for most recently developed personal data protection laws (Greenleaf, 2022), among many others in multiple US-states (California, Wyoming, Ohio New York), in Canada, Brazil, and in some parts for the recently enacted Personal Information Protection Law of China.

2.1.2 The United States of America

In comparison to other nations the US has relatively weak personal data protection laws and data transfer legislations (Pernot-Leplay, 2020; Jwa and Poldrack, 2022). However, at the same time, in the United States the situation is complex and follows the tradition of *case law* that aims to regulate actions of agents. The regulations under which data are shared have been developed by several bodies with different fields of competence. Consequently, the regulation under which human data is shared might depend on the goal of the research (e.g., FDA for medical device development) and where it was collected (e.g., HIPAA for healthcare providers or the Common rule which defines a baseline standard for almost any government-funded research in the US). In addition, specific rules of funding bodies may apply. These regulations were developed to support the basic ethical principles of respect for persons (autonomy supported by informed consent), beneficence (assessment of risks and benefits), justice (selection of participants) stated in the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979)[14]. The regulations are not laws as they were developed by federal regulatory bodies but not by congress (Kulynych, 2007; Clayton et al., 2019). Therefore, sanctioning of violations of the regulations is done by the regulatory departments of the funding bodies. Individual research participants may have only limited means for legal action (Spector-Bagdady, 2021). Generally, this may be sufficient for publicly funded neuroimaging research but it has been questioned if the Common Rule is sufficient to guarantee privacy rights to research subjects in the private sector, for example, for companies who collect genetic data to build database for commercial secondary use (Koch and Todd, 2018; Meyer, 2020). The existing situation leaves a large space for a field of unregulated research on human subjects and data processing/brokering, e.g., in privately funded research (Price and Cohen, 2019; Price et al., 2019; Meyer, 2020). The situation is sufficiently complex that we can provide here only a coarse overview. More in depth reviews are provided, for example, by Kulynych (2007) and Spector-Bagdady (2021). In the following, we will briefly go through a few aspects of the above-mentioned regulations relevant for data sharing.

The most basic fallback regulation if no other specific regulation applies (see below) is the *Common Rule* (45 CFR 46), it was defined by the Department of Health and Human Services and has been adopted by a number of federal agencies that fund or conduct research. In addition, institutions not covered may voluntarily submit an assurance to comply with it. Virtually all academic research institutions in the US are covered by the Common Rule under these premises. However, there is research on humans that is not covered by the Common Rule because it is exempt, the institutions are not federally funded, do not want to provide an assurance, or because they are covered by a different regulation (Meyer, 2020). The Common Rule has a broad *action-oriented definition of research* as "a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge." (45 C.F.R. 46.102(l)). The definition of research on human subjects is also action oriented. It involves a living individual, about whom an investigator obtains information or biospecimens through intervention or interaction with the individual, and uses, studies, or analyzes it; or obtains, uses, studies, analyzes, or generates identifiable private information or identifiable biospecimens (45 CFR 46.102 (e)). *Identifiable private information* is private information for which the identity of the subject is or may readily be ascertained by the investigator. (45 CFR 46.102(e) (5)). As a consequence, research that neither interacts with the human subject (i.e., does not collect the data) nor uses data with identifiable personal information (i.e., de-identified data) does not fall under the Common Rule (Koch and Todd, 2018). Thus, secondary research on not individually identifiable data that has been obtained, for example, from a public database likely does not fall under the Common Rule. It may neither need IRB approval nor consent (Meyer, 2020). The Common Rule is not clear about the standards for what counts as identifiable personal information and acknowledges the risk that such information could be generated (e.g., by re-identification of non-identifiable data or by merging of information from different sources like coding lists). It therefore implements a regular process of re-examining the definition of identifiable data. The Common Rule suggests that "*broad consent*" should be collected from the participants if identifiable data will be stored, maintained, or processed in secondary research. However, there are also several conditions under which the requirement to obtain consent are waived for research on subjects performed in covered institutions (Koch and Todd, 2018). The control of adherence to the Common Rule of covered research is done by the Office for Human Research Protections (OHPR). Enforcement measures can range from termination of the research, including termination of funding, to the exclusion of the investigator from federal funding. However, the Common Rule does not implement options for legal action for

13 European Data Protection Supervisor. (2020). A Preliminary Opinion on Data Protection and Scientific Research. URL: https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf; last accessed: 25.10.2022.

14 National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). The Belmont report: Ethical principles and guidelines for the protection of human subjects of research. U.S. Department of Health and Human Services. Retrieved from https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html; last accessed: 26.10.22.

human participants, e.g., in case of privacy breaches or insufficient/inaccurate informed consent (Kulynych, 2007).

The Health Insurance Portability and Accountability Act (*HIPAA*) covers protected health information (PHI) collected by covered entities and their business associates. PHI means *individually identifiable health information* (45 CFR 160.103). This can include neuroimaging, genetic and other health related data. Covered entities can be hospitals (and their neuroimaging units), healthcare providers etc. Under HIPAA *research* is defined as "a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge" (45 CFR 160.501). This definition differs from the Common Rule as it does not require interaction with the participants and therefore secondary data use is not automatically out of the reach of HIPAA. Data protection is implemented by a privacy and a security rule. The latter comprises storing and handling of data while the former defines limits of data sharing and rights of individuals. PHI can be shared with business associates under a contract ensuring adherence to the HIPAA rules. HIPAA requires *written informed consent for data sharing* (Kulynych, 2007). In principle identifiable neuroimaging data could be shared if waiver was granted by an IRB on the basis that the research cannot be performed with de-identified data (Kulynych, 2007; Spector-Bagdady, 2021). However, de-identified neuroimaging data can be publicly shared (disclosed in HIPAA terminology). In contrast to the Common Rule HIPAA provides a set of *approaches to de-identify data*, of which at least one must be implemented (45 CFR 164.514). In concordance with GDPR it requires that "the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify the individual who is a subject of the information". The Expert Method requires some expert (e.g., a statistician) to confirm that the risk of identification is low. Alternatively, the Safe Harbor Method requires that faces, biometric, and a list of 16 other identifiers[15] are removed from the data. A code can be assigned for de-identification that allows a restricted number of persons, with access to the code, re-identification. This is similar to pseudonymization under GDPR. HIPAA also requires sparseness regarding people with access to PHI and the amount of information released. It grants the *release of de-identified PHI under a Data Use Agreement* (DUA) and defines a minimal set of requirements that must be included in these DUAs, such as the prohibition of re-identification. Moreover, participants have the right to access the stored data, to correct it, and the right to restrict the uses and the disclosure (sharing) of the data (Wolf and Evans, 2018). Thus, individuals may have access to raw data and interpreted results. This is in stark contrast to the Common Rule but similar to GDPR. Another important difference to the Common Rule is that individuals have the right to complain to the covered entity and the Secretary if HIPAA rules are violated and the Department of Health and Human Services (HHS) can sanction non-adherence with a civil monetary penalty.

Additional or other regulations hold for research with a different scope. For example, human data collected in NIH funded research fall under a Certificate of Confidentiality policy and FDA regulations apply if human data was collected in the context of medical device or drug development/testing (see e.g., regulation of the test kits for direct-to-consumer genetic testing, Spector-Bagdady, 2021). This multitude of regulations is not only a burden for researchers and human research participants. They also pose a problem for data scientist who want to make use of the data and become even more virulent when the neuroimaging data is augmented by meta- or other data. Rosati (2022) points out that the scopes and concepts of the definitions of de-identified data differ among the Common Rule, HIPAA, and the NIH Data Management and Sharing policy. As a consequence, the same data can be analyzed under different regulatory regimes depending on who analyses them, for what purpose and whether they are de-identified or identifiable.

In sum, a host of regulations exists in the US which cover different institutions and types of research. Despite that, the protection of data from humans voluntarily donating their data for research appears relatively weak. Even the fallback option "Common Rule" does not cover all research uncovered by other regulations. As a simple example the Common Rule does not apply to citizen scientists when they obtain human data (Meyer, 2020). The regulations create space for a field of unregulated research on human subjects and data processing/brokering gained in such research, e.g., privately funded research (Meyer, 2020). Also, research on publicly shared data obtained from open repositories often neither needs ethical review nor consent. Note that the GDPR would still cover secondary data use (e.g., downloaded from a database) and pseudonymized (de-identified) data. The current combination of weak protection of research participants by federal law and case law which favors data collection and access over participants' autonomy (Kulynych, 2007; Price et al., 2019; Spector-Bagdady, 2021) triggered the development of new data privacy laws like the California Consumer Privacy Act (CCPA), which is strongly oriented along the GDPR. Although CCPA explicitly excludes data regulated under HIPAA, this may be the starting point for a more principled regulation with a wider scope that closes gaps left by existing regulations (Price et al., 2019). On the federal level there is now the American Data Privacy and Protection Act (ADPPA)[16] in the legislative process, that will largely pre-empt state laws if it comes into force.

2.1.3 China

China's data protection has been suggested to implement a third way between EU's GDPR, which implements a basic right for protection of personal information and control by the individual data subject with extraterritorial reach, and the decentralized, application field and data processor oriented regulations issued by different authorities in the US (Pernot-Leplay, 2020). China builds on a *hierarchy of laws* of which the higher level ones, the Cyber Security Law (CSL), the Data Security Law (DSL), the Personal Information Protection Law (PIPL), and the Civil Code

15 https://www.law.cornell.edu/cfr/text/45/164.514; last accessed: 01.03.2023.

16 https://www.dataguidance.com/jurisdiction/usa-federal; last accessed: 28.10.22.

of the People's Republic of China (CCC), constitute a normative, systematic, and complete personal information framework that is supposed to guide regulations released by domain specific institutions (Pernot-Leplay, 2020; Wang et al., 2022). This is reminiscent of the EU approach. The *"lower level" regulations* are then supposed to provide the framework for the handling of data by specific actors in specific fields. This is reminiscent of the situation in the US, where regulations are flexibly defined within certain domains and are only valid there.

The CSL was enacted 2016, 5 years before PIPL, DSL, and CCC which were enacted in 2021. The CSL and the DSL focus on the protection of national security and public interest, while the PIPL and CCC (Art. 1034–1039) focus on the protection of *personal information*. The CSL and DSL implement the principle of *data sovereignty of China*, by giving the state control of over the data acquired on the mainland of China. The DSL categorizes data in the three groups of national core data, important data, and general data where national core data can be subject to cross border protection if they are relevant for national security or public interest (Creemers, 2022; S. Li and Kit, 2021).

The PIPL and the CCC (Art. 1034–1039) protect personal information rights and interests of natural persons and seek to promote the appropriate use of personal information (PIPL Art. 1; Cheng, 2022, see Presentation 1 in Supplementary Material for original Chinese version of this publication and Table 1 for a translation into English of the important sections). They distinguish private and non-private information; sensitive and non-sensitive personal information. PIPL is superficially reminiscent of the GDPR but has important differences as it puts more emphasis on the governance model under the principle of national sovereignty. PIPL considers it the state´s task to safeguard personal data at the national and international level and delegates protection to other laws, administrative regulations, and infrastructure programs.

In Article 4 PIPL[17] defines *personal information* as information related to identified or identifiable natural persons as opposed to anonymous information. *Anonymous* information is defined in a very strict sense as "impossible to distinguish specific natural persons and impossible to restore" (PIPL Art. 73 (4)). Data handlers must *de-identify* personal information to ensure it is impossible to identify specific natural persons without the support of additional information (PIPL Art. 73 (3)). This is similar to the concept of pseudonymization in the GDPR or de-identification under HIPAA.

PIPL does not distinguish data controller from data user and subsumes the concepts under the term data handler. The *data handler* is responsible for the security of the personal information they handle (PIPL Art. 9). Articles 51–59 define their duties and Articles 66–71 define legal punishments for violations of the laws and regulations on personal information handling, including monetary penalties. Interestingly, they also define penalties for the responsible person(s) for failures of state organs to protect personal information.

PIPL requires *informed consent* from the data subject for personal information handling (PIPL Art. 13) but provides many exceptions, including other laws and regulations. The consent must be detailed (e.g., purposes of data handling, transfer abroad etc.) and must be obtained again if new purposes of data handling are intended (PIPL Art. 14) but it can be withdrawn by the data subject (PIPL Art. 15). Interestingly, at the level of PIPL there is no mention of independent review boards in the sense of IRBs.

PIPL additionally defines *sensitive* personal information which includes, among others, biometric characteristics and medical health data (PIPL Art. 28 (1)). The handling of sensitive personal information should comply with the principle of "specific purpose" plus "separate consent" (Wang, 2022, see Presentation 2 in Supplementary Material for original Chinese version of this publication and Table 1 for a translation into English of the important sections). Firstly, the handling of sensitive personal information must be for a specific purpose and with sufficient necessity, as well as with strict safeguards (PIPL Art 28 (2)). Secondly, the *separate consent* must be obtained for handling sensitive information (PIPL Art. 29). However, the concept of a "specific purpose" is indistinct. In addition, many details of the practical implementation of handling sensitive information is delegated to other laws and regulations.

Article 36 of PIPL requires personal information handled by state organs to be *stored on mainland China*. This likely includes the majority of neuroimaging, genetic and other research data. Articles 38–43 regulate sharing of personal information across borders. They require justifications for *sharing abroad*, security assessments, standard contracts, notification of the data subjects, and put the burden to control adherence of the foreign receiving party to the regulations onto the data handler. In addition, Article 53 requires from the extraterritorial data handler the appointment of a *representative on China mainland* who must be reported to the relevant departments. This could mean that a collaborator from China is necessary when human research data from there are processed abroad. Articles 44–50 provide data subjects the right to require data handlers to provide, correct, transfer, or delete their data. Articles 60–65 define departments responsible for the oversight over the personal information protection, putting the Cybersecurity and Information Department at the top of the hierarchy. Here it is also stated that everyone has the right to complain about unlawful personal information handling activities.

PIPL is a relatively new law and the future will show which effects it has on sharing of neuroimaging data. Even before PIPL came into effect, several constraints on international research data exchange (e.g., the access of foreign researchers to genetic data collected on mainland China) were implemented in laws and regulations for state reasons. In March 2018, the State Council issued the "Measures for the Management of Scientific Data", or short "*The Measures*"[18]. The Measures are binding for research institutions. They state that a scientific data archive system should be

17 Creemers R. and Webster G. (2021) Translation: Personal Information Protection Law of the People's Republic of China–Effective Nov. 1, 2021. Retrieved from https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/; last accessed 24.10.2022.

18 https://www.sciping.com/33787.html; last accessed 24.10.2022.

established and that government-funded scientific data should be submitted to this data center (The Measures Art. 12 & 13). For data produced in government funded research the Ministry of Science and Technology (MOST) can decide whether data can be shared or not. Among the criteria for restricting sharing are whether the scientific data contain personal information or concern national security. The adherence to the privacy laws is supposed to be implemented at the level of the data center policies which are currently evolving (Li et al., 2022). The complex and domain specific regulatory framework and its consequences for international sharing has been mostly analyzed in the field of *human genetic data* where China has considerable data resources. Chen and Song (2018) provide an overview of the laws and regulations and conclude that while data privacy plays a role in the regulation of data transfers, the national interest and security became a main reason for their protection by restricting their processing to the China mainland and requiring researchers from abroad to collaborate with a Chinese researcher or institution if they want to process genetic data collected in China (Chen and Song, 2018; Mallapaty, 2022). Sharing of *neuroimaging* data may be less affected by national interests as long as they are not considered health data. Regulations like safety assessment, data use agreement, data protection impact assessments, and consent for transfer may apply only from a certain size of the data sets upwards (PIPL Art. 52, Mallapaty, 2022). Moreover, the MOST has recently released a very general set of ethical norms for the use of AI in China which also covers the use and protection of personal information (Dixon, 2022).

In sum PIPL has superficial similarities to the GDPR in that it provides data subjects similar protection mechanisms (personal data, special data, requirement of consent for processing, right to withdraw consent, right to obtain information, correction/deletion of data etc.) and mechanisms to enforce their rights. These protection rights are sometimes even stronger than in the GDPR. However, it is formulated in a very general way and relies on domain specific regulations implemented by the respective authorities, similar to the data protection regulations in the US. With the additional CSL and DSL it implements mechanisms that allow state authorities to control the transfer and processing of data collected in the mainland of China to researchers abroad, thereby establishing mechanisms to enforce data sovereignty of the state. These export restrictions already have some effects in the field of human genetics. As PIPL and DSL are relatively new laws and the specific regulations are currently emerging it remains to be seen what impact they will have on the exchange of neuroimaging data.

2.1.4 Summary

The review of the three systems must remain incomplete in breadth as well as in depth. However, it highlights some convergences and differences between the regulations in three geographic and cultural regions that can be considered as among the top scientific data generators and their regulations span a spectrum in which most of the currently emerging data privacy regulations may be contained. Convergent among the regulations is that they all have regulations to protect private information and sensitive data. They all require explicit informed consent for the acquisition and sharing of data and, in general, require that the data is at least de-identified/pseudonymized or anonymized before data processing and sharing. They all suggest or require some form of contractual agreement between the data supplier and the data receiver to ensure that the data is processed in concordance with the regulations of the country in which they were acquired. While there is agreement in the subject of protection and some general means for protection of human research data, there is a great diversity in the way how the regulations are implemented and in their reach of protection. While the EU GDPR puts the protection of the individual at the center and seeks to balance it with the interest of science, the US regulations tend to favor the accessibility of data for science and economy over privacy. Laws and regulations in China emphasize both the protection of the individual as well as state interest and data sovereignty. Moreover, the regulations differ considerably with respect to accountability, liability, and sanctioning with the most lenient regulations in the US and potentially the most comprehensive definitions of responsibilities in China's laws. Importantly care should be taken when matching terminology across jurisdictions (Eke et al., 2021). The same terms may have somewhat different meanings and some functional roles that may be distinguished in one context (e.g., data controller and data processors in GDPR) but lumped into one role in another (data handler in PIPL). It is, however, encouraging that there appears to be sufficient overlap that a limited set of measures could allow neuroimaging data sharing in a way that is compatible with all three sets of regulations for privacy protection. Table 2 provides an overview of such measures for the three jurisdictions. However, this should be further analyzed and corresponding procedures should be developed.

2.2 The (meta)data description challenge

Even when legal regulations are met and datasets are publicly shared, it is not guaranteed that the information in them is accessible and useful for a data processor. The FAIR principles (Wilkinson et al., 2016) state some general requirements of how scientific data should be handled and documented to make them useful for others. The acronym FAIR stands for Findable, Accessible, Interoperable, and Reusable. Findability means

TABLE 2 Some points researchers need to consider when sharing data or using shared data.

EU	United States	China
IRB, approval of lawfulness of processing, pseudonymization (anonymized data not covered by DPR). Specific consent important for legal sharing.	Under HIPAA: IRB, de-identification or anonymization (e.g., Safe Harbor Method), consent for sharing and/or DUA, depending on type of data.	De-identification. Very high standard for anonymization. Detailed consent for all forms of processing.
Lawful sharing possible within EU an states with adequacy decision. DUA to restrict processing purposes of data recipient outside EU.	No restriction for sharing into different countries.	Complex procedure for sharing outside mainland China. May require collaborator in China or be impossible depending on data classification.

that the data is either aggregated in a way (e.g., on a server) that the user knows where to look for them or that they are equipped with descriptive metadata such that some sort of search engine can retrieve their location. In addition, data should have a persistent identifier, such as a digital object identifier (doi), to assure findability over a long time period. Accessibility refers to the ability of a human or a computer to either retrieve the data from their storage location or to run them through an analysis pipeline on a remote server without retrieving them. Interoperability means that data should integrate into different data analysis ecosystems as well as the integration of data with other data. In particular with big data, interoperability is necessary to use the data on computers without or with minimal human interaction. Reusability aims at efficient data use which is of particular importance for data that are rare or expensive to produce. It means that data should not only be useful for the purpose they were originally collected for (Bigdely-Shamlo et al., 2020; Niso et al., 2022).

To match these requirements, research data need to be organized according to some standard. Using a standardized data structure alone, however does not suffice to ensure that shared data become findable, interoperable, or reusable, for example, in large-scale meta studies. Proper description of the data is another requirement. This has been pointed out, among others, by the European Commission expert group for FAIR data. The expert group recommended comprehensive documentation of research products, such as experimental data or analysis pipelines, through metadata (European Commission and Directorate-General for Research and Innovation, 2018). Ideally, these metadata are based on standard vocabularies or ontologies, which add semantics to the terms of the vocabulary.

The domains for metadata range from descriptions of the human participants, the experiment, the nature of the experimental data, additional tests and surveys, to consent and usage restrictions. Even though, many publicly shared datasets contain some metadata, these are likely not descriptive enough to effectively re-use them and working with such data can be error-prone and tedious (Niso et al., 2022). Additionally, metadata are often described in idiosyncratic terminology of the researchers, who share the dataset, making them hard to interpret for (other) humans and impossible for machines. This severely restricts findability, interoperability, and reusability. The latter particularly in the context of big data research efforts. One way to cope with this problem is to define vocabularies or even ontologies, which can then be used to annotate the data in a standardized manner. For example, a neuroimaging dataset with standardized event annotation can be re-used for purposes it was not originally collected for (Bigdely-Shamlo et al., 2020; Niso et al., 2022), simply because the experiment may include events, that were unrelated to the original research question but necessary for the structure of the experiment (e.g., button press events that require motor responses which might not have been in the scope of the original study). Ideally, if augmented by rich metadata, complex datasets can be used in many studies with different purposes (e.g., United Kingdom Biobank[19], Study Forrest[20])

Recently, the neuroimaging community elaborated open standards for data storage yielding common structural organizations of raw datasets from different modalities (Teeters et al., 2015; Gorgolewski et al., 2016; Niso et al., 2018; Pernet et al., 2019). The most commonly used is the Brain Imaging Data Structure (BIDS[21], Gorgolewski et al., 2016). Importantly, many neuroimaging data analysis tools have adopted the standard and interoperate on it to some degree. However, the standardization is still not comprehensive enough to guarantee the full FAIRification of datasets including derivatives. Moreover, other scientific communities may have different standards that may be less developed or lack standards at all. The reasons for that can be manifold, including but not restricted to the lack of a culture supporting sharing, the ubiquitous use of closed commercial systems, or particularly strong data protection constraints due to commercial interests, as in industry or in the health domain. Since we cannot cover the wide range of data standards in this paper, we focus on BIDS as a showcase for structured data storage enriched with some metadata.

2.2.1 BIDS

BIDS is a community driven project to abstract and standardize the representation of neuroimaging data. Essentially it breaks down to a hierarchical directory structure with specific data-file and folder naming conventions plus some standardized metadata for the description of the image acquisition and the event annotations of the experiment (given that the experiment deploys a task-based structure). Importantly, BIDS is not only defined as a human readable directory hierarchy but also as a computer interoperable schema, which allows for more flexibility, is less error-prone with respect to maintenance of the standard, and facilitates the usage of automated processing pipelines on BIDS datasets. Moreover, the metadata and some of the data (e.g., timing of events) are also human readable, which eases the understanding of the dataset. Such a unifying data structure carries the potential to make neuroscientific research more transparent and encourages data sharing between researchers and labs.

These advantages of BIDS only apply if the data structure is widely accepted and used. For this reason many experts from the neuroimaging community were consulted during the development of BIDS to create a data format which is intuitive and easy to use while being able to handle a variety of experimental data, e.g., from different modalities such as fMRI (Gorgolewski et al., 2016), EEG (Pernet et al., 2019), MEG (Niso et al., 2018), behavioral data, and many more. It can thus be used for most experiments and even across imaging techniques for the standardized storage of multimodal datasets. Since BIDS is a rather young development and open source, it is constantly evolving to describe more aspects of the data acquisition and the respective analyses applied.

BIDS defines some basic data acquisition related metadata and strongly recommends to include them in every dataset. Additionally, BIDS requires that metadata are stored in the Java Script Object Notation (JSON), an open and text-based file format consisting of attribute-value pairs that are both human and machine readable.

19 https://www.ukbiobank.ac.uk; last accessed: 26.10.22.

20 https://www.studyforrest.org/; last accessed: 26.10.22.

21 https://bids.neuroimaging.io/; last accessed: 26.10.22.

Even though these JSON files are not mandatory according to the BIDS specification, they are most often included in (publicly shared) BIDS datasets, simply because the tools that convert datasets from the vendor specific format to BIDS extract them from the former and write them to the JSON-files of the latter. These conversion tools are currently best developed in the MRI domain, e.g., HeuDiConv (Halchenko Y. et al., 2021) and dcm2bids[22], but there are ongoing community efforts to facilitate the conversion to BIDS for other modalities, such as MEEG (MNE-BIDS[23], Appelhoff et al., 2019). However, these basic metadata defined in BIDS do not suffice for exhaustive description of the raw data nor for the description of analyses employed to obtain data derivatives, e.g., results of an analysis. One of the reasons is that BIDS defines a framework for several data acquisitions modalities, all of which require domain specific metadata. Additionally, different fields of research may require different metadata which again adds complexity to the task of developing an exhaustive, overarching and modality agnostic metadata standard within BIDS.

2.2.2 HED tags and the neuroimaging data model (NIDM)

In the neuroimaging domain the Hierarchical Event Descriptor standard (HED[24], Bigdely-Shamlo et al., 2016; Robbins et al., 2021) is an infrastructure which defines rules for controlled and hierarchically organized vocabularies. Terms from these vocabularies can then be used to describe the nature and time course of an experiment, that was performed while brain data was recorded, by tagging the data with keywords while assuring findability of these tags during downstream analyses. The HED base schema defines a hierarchical vocabulary for the description of basic stimuli, responses, tasks and experimental conditions. However, more specialized or domain specific vocabularies/schemas can be added to the standard as long as they adhere to the rules for schemata defined by HED. One example is the SCORE vocabulary for clinical EEG annotation (Beniczky et al., 2013, 2017), which has been converted to an HED schema and is currently under community review. Moreover, existing vocabularies can be extended to cover a wider range of applications or use cases. HED was developed in a community effort, recently fully integrated into the BIDS ecosystem and since the release of BIDS 1.8.0. tagging data with terms from, multiple vocabularies is accepted[25].While far from being able to completely annotate all research products, like analysis pipelines, the HED vocabularies are an important ingredient to make data sets machine actionable and reduce ambiguity for human researchers. Moreover, tagging your data with these standardized HED-tags allows for better collation of separately recorded datasets.

The Neuroimaging Data Model (NIDM[26], Keator et al., 2013; Maumet et al., 2016) complements HED by providing additional functionality, such as the description of analysis workflows and results (though currently limited to MRI-data). Importantly it provides methods to describe the provenance of research products, i.e., the way they were generated. Provenance documentation is expected to increase reproducibility and to improve the usefulness of sharing analysis methods. NIDM employs different components to model different aspects of the data: NIDM Experiment for capturing and annotating experimental metadata (similar scope as HED), NIDM Workflow for the standardized description of analysis workflows, and NIDM Results (Maumet et al., 2016) for standardized description of results including provenance information. It should be noted, however, that these components are at different stages of development, with the NIDM Results being the most sophisticated. NIDM is a spin off from the US Brain Initiative and is based on Semantic Web technology. It is mainly based on the PROV (provenance) vocabulary (Moreau et al., 2015). However, it also incorporates terms from several other vocabularies or ontologies such as the Dublin Core[27] for file description and the STATisical Ontology (STATO)[28] for the annotation of statistical methods like General Linear Models. Additionally, the NIDM developers have started to map terms/study variables, commonly used in openly shared datasets, to concepts from existing ontologies/vocabularies, such as the Cognitive Atlas (Poldrack et al., 2011) or the InterLex information resource. This initiative is called the NIDM-Terms[29] and community efforts to expand this ontology are welcome.

In practice an immense amount of data and metadata standards exist even within such a small research field as neuroscience. Many of those standards are very narrow in their range of application, lack community/institutional support, and are potentially overlapping. This could lead to suboptimal use of human as well as financial resources. In an effort to integrate the different standardization approaches, the open Metadata Initiative for Neuroscience Data Structures (openMINDS[30]), which emerged from the EU Human Brain Project, aims to collect and integrate metadata standards into an overarching ontology to connect terminologies used in various fields of neuroscience. In addition, they also collect frequently used brain atlases and common coordinate spaces for neuroimaging data. Similar to NIDM, the openMINDS project is subdivided into several modules, which differ with respect to their level of development.

2.2.3 Metadata and privacy protection

Metadata annotations and privacy protection in legal frameworks may appear as two different challenges to the same problem, the lack of useful openly shared data. However, they are potentially connected. Data which is equipped with rich metadata is more likely to be de-identified and hence the developers of vocabularies or metadatmodels need to be cautious when

22 https://unfmontreal.github.io/Dcm2Bids/; last accessed:02.02.2023.

23 https://mne.tools/mne-bids/stable/index.html; last accessed: 30.11.22.

24 https://www.hedtags.org/; last accessed: 25.10.22.

25 https://bids-specification.readthedocs.io/en/stable/appendices/hed.html#hierarchical-event-descriptors; last accessed: 30.11.2022.

26 http://nidm.nidash.org/; last accessed: 25.10.2022.

27 https://www.dublincore.org/specifications/dublin-core/dcmi-terms/; last accessed: 30.11.2022.

28 http://stato-ontology.org/; last accessed: 30.11.2022.

29 https://nidm-terms.github.io/info.html?#about; last accessed: 26.10.2023.

30 https://ebrains.eu/news/new-openminds-metadata-models/; last accessed:25.10.2022.

including terms which could be mapped to identifiable information. More general, there is a trade-off between data which is perfectly described by metadata and minimizing the risk of re-identification. Additionally, the safeguards that need to be implemented and the metadata that need to be removed or "filtered" can vary depending on the legal regulations that apply to the data. However, little is currently known to what extend comprehensive sets of metadata may impact privacy protection in practice, how metadata could be exploited by future AI techniques, and how safety assessments would change with increasing volumes of findable, openly accessible, and properly annotated data.

3 Practical solutions

So far, we have covered important factors that may have a negative impact on useful data sharing, i.e., lawful sharing of data, that can be easily understood and interpreted. We also covered the benefits for the individual researcher and society. In this last section we want to introduce some tools, practices and initiatives that support the individual researcher in reducing the additional effort/labor associated with data sharing. Some of these may be specific to data from human neuroimaging but others might be more general, applying to a wide range of data types from different fields.

3.1 Consent and anonymization

A recent survey on open science practices in the functional neuroimaging community revealed that 41% of the researchers did not share their data due to the fact that their consent forms excluded the option for data sharing (Paret et al., 2022). Hence, researchers who plan to share data should take care to design the consent form in a way that data can be shared on a lawful basis or include a consent form that was specifically designed for that purpose. Obtaining explicit consent is one central building block for lawful data sharing. However, researchers should be aware that the informed consent to participate in the experiment does not entail consent to sharing the data with others. The explicit consent to sharing the data can be integrated into the informed consent form, though. This must be done in a way, such that the data subject clearly understands that their data might be shared with the research community in a pseudonymized form. Moreover, data subjects should understand the researcher's role in mitigating the risk of a privacy breach through re-identification. In order to simplify that step, the Open Brain Consent (OBC) working group (Bannier et al., 2021) provides template consent forms in many languages on their website[31]. They are designed to meet the requirements for explicit consent under the GDPR. Table 2 lists some points to consider for lawful data sharing in different jurisdictions. It should be noted here however, that the final decision whether obtaining informed consent for public sharing of pseudonymised data is in the hands of the data protection office of the research facility, and in practice their assessment may vary between institutions.

Besides obtaining consent, anonymization, de-identification or pseudonymization (in case anonymization is not possible) of the data are required in any of the legal frameworks covered here. There are numerous techniques for anonymization, de-identification and pseudonymization. If unsure which technique to use, the European Data Protection Working Party has issued an opinion on anonymization techniques[32] in 2014, highlighting benefits and potential pitfalls of several anonymization approaches including differential privacy, randomization, noise addition, permutation, generalization, and L-diversity/T-closeness. Additionally, several free and open-source tools exist to apply these techniques. For example, the ARX anonymization tool[33] (Prasser et al., 2014) provides functionality to anonymize data and additionally analyze the risk of re-identification for the chosen anonymization/de-identification technique. These general tools are useful for metadata. Neuroimaging data are more complex, since not only metadata need to be curated to achieve anonymization. In the case of fMRI all facial features need to be eliminated, a process called defacing. The OBC working group (Bannier et al., 2021) again provides links to some useful tools on their website[34], e.g., tools for sanitizing the DICOM header and tools for defacing. For example, BIDSonym[35] (Herholz et al., 2021) provides an interface for BIDS data which allows defacing using different techniques.

3.2 Data user agreements and databases

Data user agreements (DUA) are one option to bind the data processor (entity that receives the data) to some set of predefined conditions when accessing the shared data. This is particularly important when they belong to the category of sensitive data. DUAs have become a prominent way to mitigate the misuse of data and are applicable in different jurisdictions. A DUA is a contract between the data controller and an external entity or the person seeking to access the data. It defines a set of rules around the shared data. With such agreements a data controller can control with whom or for what purposes they want to share the data. For example, data can be shared under the constraint that no re-identification will be attempted, or for scientific research purposes only, thereby excluding the use of the shared data for economic purposes. DUA's are endorsed by the European government and are a step towards fulfilling the principle of privacy by design, as required by the GDPR. An exemplary template of a DUA is provided on the OBC's webpage[36].

31 https://open-brain-consent.readthedocs.io/en/stable/index.html; last accessed 22.01.2023.

32 https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf; last accessed: 24.01.2023.

33 https://arx.deidentifier.org/anonymization-tool/; last accessed 22.01.2023.

34 https://open-brain-consent.readthedocs.io/en/stable/index.html; last accessed 22.01.2023.

35 https://github.com/PeerHerholz/BIDSonym; last accessed: 12.02.2023.

36 https://open-brain-consent.readthedocs.io/en/stable/index.html; last accessed: 22.01.2023.

Providers of public data platforms or repositories need to implement a mechanism to handle and store such (digital) contracts. Moreover, these platforms need some kind of access control and identification mechanisms, since DUAs are legally binding contracts. Unfortunately, many well-known public and open neuroimaging data repositories, e.g., OpenNeuro (Markiewicz et al., 2021), Distributed Archives for Neurophysiology Data (DANDI)[37], and the International Data-Sharing Initiative (INDI)[38] have hitherto not or only partially implemented infrastructure for DUAs or access control mechanisms. While this might be sufficient to share some data acquired in the US, it may not suffice for data acquired under HIPAA, the GDPR and Chinese laws and regulations. However, there are also several data platforms that allow lawful sharing under the GDPR with the required safeguards. For example, Ebrains[39] provides a platform for sharing several kinds of data. There, uploading data is only possible if certain technical and organizational measures for safeguarding the individual's right to privacy are met. There the data needs to be de-identified, anonymized or pseudonymized and is additionally safeguarded (encrypted) *via* the Human Data Gateway. Moreover, the users who want to access data need to comply to a given set of conditions, one of which is the acceptance of an additional DUA. Another example is the Open MEG Archive (OMEGA, Niso et al., 2016). This is a data repository specialized on MEG data. It implements a controlled access mechanism (institutional credentials are necessary to create an account) and requires signing a DUA before data access. A list of some online data repositories with information on the safeguards that these databases have implemented can be found in Eke et al. (2021).

3.3 User-to-data

The concept of user-to-data describes an alternative approach to data custodianship to avoid legal issues revolving around shared data. The idea behind this concept is that data does not need to change its location (the server or computer it is stored on) to be useful to many people. Instead, users can be "moved" to the data by giving them means to work on the data and run analyses on them without having full access rights, e.g., researchers can not see or copy the data. Consequently, this requires the host websites to provide some kind of interface for working with the data on their servers. One example of this approach is brainlife[40]. This platform also provides sufficient computing power to run analyses, test algorithms or to benchmark software and has streamlined access to data from various open databases. However, brainlife does not entirely exclude the option to download data processed on their servers. A Data Safe Haven provides a secure environment for the analysis of sensitive data with appropriate technical and informational governance mechanisms. Data Safe Havens have been developed at several institutions and universities, such as the UCL[41], or the university of Hull[42]. The Turing Data Safe Haven[43] is a resource that comprises general information on Data Safe Havens as well as scripts and templates to set-up and maintain such secure environments. Moreover and very recently, several initiatives have emerged targeting the facilitation of setting up privacy preserving frameworks for the analysis of sensitive data, such as Vantage6[44] or OpenMined[45]. Vantage6 is an open source infrastructure for privacy preserving analysis. It provides functionality for servers, which allow setting up "data stations" which securely store the data. Algorithms can be delivered to these "data stations" and results will be sent back to the user. OpenMined is a movement, which is composed of three programs: the build, the educate, and the impact program. The build program is about developing tools to help setting up privacy preserving data analysis environments. This is similar to Vantage6, though with a strong focus on running AI methods on the data. The educate program clearly is about education of remote data science, especially since this is a comparatively new field of data science. They provide several courses to learn more about remote data science and working with their PySyft[46] library. The impact program is about showing that the developed tools work by teaming up with partners from public and private organizations to test the generalizability and usability of them in a variety of use cases. The user-to-data approach seems to be promising to enable data access for many people with minimal legal constraints, however, it needs to be considered that limited compute and storage capacities might be the bottleneck of this approach. Additionally, the maintenance of the infrastructure is complex and expensive. Smart data management tools, such as Datalad (Halchenko Y. O. et al., 2021), can promise some relief to the resource problem by employing a decentralized structure (Hanke et al., 2021), e.g., servers for databases need not be at the same physical location. Detailed information on Datalad, e.g., its usage and range of application, can be found in the Datalad Handbook (Wagner et al., 2021). Finally, the speed of technical development might also mitigate the issues with resources.

3.4 Tools for data and (meta)data handling

Making a dataset useful for other researchers can be costly. Data and metadata standards support this task. Fortunately, tools exist that help implementing these standards in everyday scientific practice. They support data transformation, metadata annotation, and data handling in general. This can include software for the conversion into a given data storage standard or file format, software for data management, parsers for specific file formats, tools to filter

37 https://dandiarchive.org/; last accessed 24.01.2023.

38 http://fcon_1000.projects.nitrc.org/; last accessed: 24.01.2023.

39 https://ebrains.eu/service/share-data; last accessed: 21.10.2022.

40 https://brainlife.io/projects; last accessed:21.10.22.

41 https://www.ucl.ac.uk/isd/services/file-storage-sharing/data-safe-haven-dsh; last accessed: 12.02.2023.

42 https://datasafehaven.hull.ac.uk; last accessed: 12.02.2023.

43 https://alan-turing-institute.github.io/data-safe-haven/develop/overview/index.html; last accessed: 12.02.2023.

44 https://distributedlearning.ai; last accessed: 12.02.2023.

45 https://www.openmined.org; last accessed: 14.02.2023.

46 https://github.com/OpenMined/PySyft; Last accessed: 14.02.2023.

the dataset for specific metadata, ideally with many options for queries, tools for validation of adherence to a given standard, and tools for metadata extraction or editing of metadata files.

In Section 2.2.1 we mentioned some tools that help with the conversion of rawdata into BIDS, covering several modalities and programming languages. In addition, the BIDS community offers a web-based tool for the validation process. For interaction with the BIDS converted data stored locally, BIDS-Matlab (Gau et al., 2022) and PyBIDS/ancpBIDS (Yarkoni et al., 2019) are commonly used tools. Both allow for complex queries on the data, hence many filtering options and provide an API for their integration into custom workflows or pipelines. Moreover, the DataLad (Halchenko Y. et al., 2021) family provides useful functionality for decentralized data management (i.e., data that is stored on several servers or repositories), while additionally tracking the provenance of all files in a dataset. Extensions to DataLad target more specific aspects of data handling. For example, MetaLad[47] is a tool which is specifically designed to facilitate the handling of metadata. It can deal with various file formats and provides useful functionality, such as filtering existing metadata, e.g., for specific keys, or the extraction and aggregation of metadata. On top of that, DataCat[48] is another DataLad extension, which eases user interaction with the metadata by providing browser-based and easy-to-navigate-through metadata catalogues, i.e., a user interface which facilitates metadata inspection and handling. Note, that DataCat is still under development and no stable version exists yet.

Additional tools are available for working with the metadata standards mentioned in Section 2.2.2. The NIDM team has developed a python-based command line tool (PyNIDM[49]) and an additional web application which allow the user to convert BIDS data into NIDM files, interactively map terms (e.g., study variables from a tabular sidecar file) to concepts in existing ontologies/vocabularies or to define new terms. These tools also allow the creation of JSON-formatted data dictionaries, e.g., with provenance information, which are then stored as sidecar files alongside the data. Additionally, the developers of HED provide several online tools[50]. They include tools for validation, summarization and generation of BIDS compatible events-files, tools for the generation, validation, transformation, extraction and merging of respective JSON sidecar files, which are designed to semantically describe the columns of the events-files. Moreover, HED offers a tool to validate and convert new schemas or extensions to existing schemas (vocabularies). All of these tools are intuitive and easy to use and provide a self-explaining browser-based user interface and unlike command line tools, the HED online tools do not require any prior experience in programming or any operation system specific knowledge since they are browser based. Technically, this should also enable the user to make use of these tools on mobile devices, such as tablets.

The scope of this paper does not allow for an exhaustive list of tools and practices for open neuroimaging. Therefore, we refer the interested reader to Niso et al. (2022) and, in particular, the table in the supplementary material, for a more detailed overview of available open science tools and practices, that support transparent and reproducible research at every stage of the research cycle.

4 Conclusion

Despite the manifold benefits of shared data for individual researchers, the scientific community and society, only a small fraction of data generated in life sciences is made openly available (Houtkoop et al., 2018). Moreover, the data, that is openly shared, is often of limited use because it is not saved in a standardized way and/or insufficiently described. This renders them hardly understandable for humans and prevents automated computer interoperability. Here, we cover the two important factors contributing to these problems: insecurities around the lawfulness of data sharing as well as missing metadata and standardized data organization. Many individual researchers withhold their data because they lack knowledge about options for sharing and are afraid of legal implications of privacy protection laws (Eke et al., 2022). In order to shed light on options and constraints for sharing human neuroimaging and comparable human data, we provided an overview of relevant legal frameworks in the three geographic regions with the largest data resources, provided an accessible tabular overview, provided a concise overview of points to consider when planning to share data, and introduced platforms and procedures that support lawful human research data sharing. In order to ease the burden of standardizing data organization and annotation we introduced initiatives, that develop standardized data structures and vocabularies for the description of neuroimaging data. Additionally, we provided an overview of free, community developed, and open source tools and databases that simplify the construction and reproduction of analysis pipelines by integrating standards and practices, covered here, into the research workflow. The mentioned tools/initiatives/practices can drastically reduce the over-head for FAIR and lawful data sharing for the individual researcher, increase the efficiency of data handling, and increase the reusability of the data and thereby their value for the individual researcher, the scientific community, and society.

At a first glance, the three legal frameworks covered here appear very different and they are, when scrutinizing details like the definitions of terminologies, their reach of protection and the implemented mechanisms for sanctioning. However, at a practical level, there is quite some overlap among the requirements for research data sharing: A combination of IRB, detailed explicit consent, and pseudonymization is at the core of all regulations and established practice in the majority of (neuroimaging) labs handling human data. Additionally, DUAs help with sharing data requiring special protection. However, there are several domains, in which further improvements are desirable. In the foreseeable future, DUAs and user-to-data platforms may play a bigger role if the volume of internationally shard data increases. More and better tools are required to support

47 https://github.com/datalad/datalad-metalad; last accessed: 03.02.2023.

48 https://github.com/datalad/datalad-catalog; last accessed: 03.02.2023.

49 https://github.com/incf-nidash/PyNIDM; last accessed: 27.01.2023.

50 https://hedtools.ucsd.edu/hed; last accessed: 26.01.2023.

this development as only few and often local user-to-data platforms exist and the handling of DUAs is still in its infancy and not really useful in AI applications aiming to include datasets from distributed sources in addition to, or instead of, centralized large databanks. Moreover, the assessment of risk for re-identification seems underdeveloped for neuroimaging data compared to some common metadata, for which risk-assement procedures and tools already exist. However, the interactions between neuroimaging and metadata in risk assessment seems unexplored although such interactions can be expected. At the level of legal regulations, it has been reported that the GDPR serves as a blueprint for many privacy protection laws that are currently developed or updated in countries around the world (Greenleaf, 2022). This trend may support the homogenization of privacy protection laws across jurisdictions and as a consequence allow the development of some generalizable core practices for sharing, although local regulatory idiosyncrasies, that need to be met, will likely continue to exist.

Shared data must meet some requirements to be useful. Among others are adherence to a well-established open data standard that is supported by tools for data conversion, data handling and frequently used analysis tools. Moreover, standardized metadata are necessary to make them understandable. So far only few tools exist to augment the core data with metadata and to process them. Standardization of data storage formats and metadata is core to make a dataset FAIR and useful for humans and machines. Most researchers may have searched for a data reader because the favorite analysis tool cannot processes the format of the desired data. Many may be familiar with the guessing whether "RT" in one dataset may mean the same as "index" in another, and "button press" in a third. Such obstacles can, in principle, be removed when open data standards are used. However, when it comes to choosing a standard the blessing of many options can turn into a burden. Our own approach to the choice problem is to consider a) wide acceptance and adoption in the community, b) the existence of tools that support the application to the data, c) support of the standard by tools used in the analysis workflow or even automation of it, d) sustainability supported by a strong community that continuously develops the standard and respective tools, e) that time to develop idiosyncratic solutions for an individual lab is often wasted and better invested in the support of community developments.

Author contributions

AR: conception, design, draft, content and revision; AW: draft, content and revision; XW: draft, content and revision; JR: conception, design, draft, content and revision. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This research was supported by the DFG Device Center Grant INST 184/216-1"Tools and infrastructure for open and reproducible neuroimaging".

Acknowledgments

We thank Stephan Heunis and an anaonymous reviewer for their constructive comments that greatly helped to improve the manuscript.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1086802/full#supplementary-material

References

Abramian, D., and Eklund, A. (2019). "Refacing: Reconstructing anonymized facial features using GANS," in 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) (Venice, Italy: IEEE), 1104–1108. doi:10.1109/ISBI.2019.8759515

Allen, C., and Mehler, D. M. A. (2019). Open science challenges, benefits and tips in early career and beyond. *PLOS Biol.* 17 (5), e3000246. doi:10.1371/journal.pbio.3000246

Appelhoff, S., Sanderson, M., Brooks, T., van Vliet, M., Quentin, R., Holdgraf, C., et al. (2019). MNE-BIDS: Organizing electrophysiological data into the BIDS format and facilitating their analysis. *J. Open Source Softw.* 4 (44), 1896. doi:10.21105/joss.01896

Bannier, E., Barker, G., Borghesani, V., Broeckx, N., Clement, P., Emblem, K. E., et al. (2021). The Open Brain Consent: Informing research participants and obtaining consent to share brain imaging data. *Hum. Brain Mapp.* 42 (7), 1945–1951. doi:10.1002/hbm.25351

Beniczky, S., Aurlien, H., Brøgger, J. C., Fuglsang-Frederiksen, A., Martins-da-Silva, A., Trinka, E., et al. (2013). Standardized computer-based organized reporting of EEG: Score. *Epilepsia* 54 (6), 1112–1124. doi:10.1111/epi.12135

Beniczky, S., Aurlien, H., Brøgger, J. C., Hirsch, L. J., Schomer, D. L., Trinka, E., et al. (2017). Standardized computer-based organized reporting of EEG: SCORE – second version. *Clin. Neurophysiol.* 128 (11), 2334–2346. doi:10.1016/j.clinph.2017.07.418

Bigdely-Shamlo, N., Cockfield, J., Makeig, S., Rognon, T., La Valle, C., Miyakoshi, M., et al. (2016). Hierarchical event descriptors (HED): Semi-structured tagging for real-world events in large-scale EEG. *Front. Neuroinformatics* 10. doi:10.3389/fninf.2016.00042

Bigdely-Shamlo, N., Touryan, J., Ojeda, A., Kothe, C., Mullen, T., and Robbins, K. (2020). Automated EEG mega-analysis I: Spectral and amplitude characteristics across studies. *NeuroImage* 207, 116361. doi:10.1016/j.neuroimage.2019.116361

Chen, Y., and Song, L. (2018). China: Concurring regulation of cross-border genomic data sharing for statist control and individual protection. *Hum. Genet.* 137 (8), 605–615. doi:10.1007/s00439-018-1903-2

Cheng, X. (2022). Discussion of the relationship between the civil code and the personal information protection law (论《民法典》与《个人信息保护法》的关系). *Sci. Law (法 律科学(西北政法大学学报))* 19 (20).

Clayton, E. W., Evans, B. J., Hazel, J. W., and Rothstein, M. A. (2019). The law of genetic privacy: Applications, implications, and limitations. *J. Law Biosci.* 6 (1), 1–36. doi:10.1093/jlb/lsz007

Collins, F. S., Morgan, M., and Patrinos, A. (2003). The human genome project: Lessons from large-scale biology. *Science* 300 (5617), 286–290. doi:10.1126/science.1084564

Creemers, R. (2022). China's emerging data protection framework. *J. Cybersecurity* 8 (1), tyac011. doi:10.1093/cybsec/tyac011

de Jonge, H., Cruz, M., and Holst, S. (2021). Funders need to credit open science. *Nature* 599 (7885), 372. doi:10.1038/d41586-021-03418-1

Dixon, R. B. L. (2022). A principled governance for emerging AI regimes: Lessons from China, the European Union, and the United States. *AI Ethics*. doi:10.1007/s43681-022-00205-0

Eke, D., Aasebø, I. E. J., Akintoye, S., Knight, W., Karakasidis, A., Mikulan, E., et al. (2021). Pseudonymisation of neuroimages and data protection: Increasing access to data while retaining scientific utility. *Neuroimage Rep.* 1 (4), 100053. doi:10.1016/j.ynirp.2021.100053

Eke, D. O., Bernard, A., Bjaalie, J. G., Chavarriaga, R., Hanakawa, T., Hannan, A. J., et al. (2022). International data governance for neuroscience. *Neuron* 110 (4), 600–612. doi:10.1016/j.neuron.2021.11.017

European CommissionDirectorate-General for Research and Innovation (2019). *Cost-benefit analysis for FAIR research data: Cost of not having FAIR research data*. Publications Office. doi:10.2777/02999

European CommissionDirectorate-General for Research and Innovation (2018). *Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data*. Publications Office. doi:10.2777/1524

Freedman, L. P., Cockburn, I. M., and Simcoe, T. S. (2015). The economics of reproducibility in preclinical research. *PLOS Biol.* 13 (6), e1002165. doi:10.1371/journal.pbio.1002165

Gau, R., Flandin, G., Janke, A., Tanguyduval Oostenveld, R., Madan, C., Niso Galán, G., et al. (2022). Bids-matlab. *Zenodo*. doi:10.5281/zenodo.5910585

Europäische Kommission; Generaldirektion Forschung und InnovationBaker, L., Cristea, I., Errington, T., Jaśko, K., et al. (2020). in *Reproducibility of scientific results in the EU: Scoping report*. Editor W. Lusoli (Publications Office). doi:10.2777/341654

Gibbs, R. A. (2020). The human genome project changed everything. *Nat. Rev. Genet.* 21 (10), 575–576. Article 10. doi:10.1038/s41576-020-0275-3

Glasziou, P., and Chalmers, I. (2018). Research waste is still a scandal—an essay by Paul glasziou and iain Chalmers. *BMJ* 363, k4645. doi:10.1136/bmj.k4645

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* 3 (1), 160044. Article 1. doi:10.1038/sdata.2016.44

Greenleaf, G. (2022). Now 157 countries: Twelve data privacy laws in 2021/22 (SSRN scholarly paper No. 4137418). Available at: https://papers.ssrn.com/abstract=4137418.

Halchenko, Y., Goncalves, M., Castello, M. V., di, O., Ghosh, S., Salo, T., et al. (2021). Nipy/heudiconv. *Zenodo*. doi:10.5281/zenodo.5557588

Halchenko, Y. O., Meyer, K., Poldrack, B., Solanky, D. S., Wagner, A. S., Gors, J., et al. (2021). DataLad: Distributed system for joint management of code, data, and their relationship. *J. Open Source Softw.* 6 (63), 3262. doi:10.21105/joss.03262

Hanke, M., Pestilli, F., Wagner, A. S., Markiewicz, C. J., Poline, J.-B., and Halchenko, Y. O. (2021). In defense of decentralized research data management. *Neuroforum* 27 (1), 17–25. doi:10.1515/nf-2020-0037

Herholz, P., Ludwig, R. M., and Poline, J.-B. (2021). *BIDSonym—a BIDSapp for the pseudo-anonymization of neuroimaging datasets*. PsyArXiv. doi:10.31234/osf.io/3aknq

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E., and Wagenmakers, E.-J. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Adv. Methods Pract. Psychol. Sci.* 1 (1), 70–85. doi:10.1177/2515245917751886

Hunt, L. T. (2019). The life-changing magic of sharing your data. *Nat. Hum. Behav.* 3 (4), 312–315. Article 4. doi:10.1038/s41562-019-0560-3

Jwa, A. S., and Poldrack, R. A. (2022). The spectrum of data sharing policies in neuroimaging data repositories. *Hum. Brain Mapp.* 43 (8), 2707–2721. doi:10.1002/hbm.25803

Keator, D. B., Helmer, K., Steffener, J., Turner, J. A., Van Erp, T. G. M., Gadde, S., et al. (2013). Towards structured sharing of raw and derived neuroimaging data across existing resources. *NeuroImage* 82, 647–661. doi:10.1016/j.neuroimage.2013.05.094

Koch, V. G., and Todd, K. (2018). *Research revolution or status quo?: The new common rule and research arising from direct-to-consumer genetic testing (SSRN scholarly paper No. 3132849)*. doi:10.2139/ssrn.3132849

Kulynych, J. J. (2007). The regulation of MR neuroimaging research: Disentangling the gordian knot. *Am. J. Law Med.* 33 (2–3), 295–317. doi:10.1177/009885880703300207

Kurtz, C., Wittner, F., Semmann, M., Schulz, W., and Böhmann, T. (2022). Accountability of platform providers for unlawful personal data processing in their ecosystems–A socio-techno-legal analysis of Facebook and Apple's iOS according to GDPR. *J. Responsible Technol.* 9, 100018. doi:10.1016/j.jrt.2021.100018

Li, C., Zhou, Y., Zheng, X., Zhang, Z., Jiang, L., Li, Z., et al. (2022). Tracing the footsteps of open research data in China. *Learn. Publ.* 35 (1), 46–55. doi:10.1002/leap.1439

Li, S., and Kit, C. (2021). Legislative discourse of digital governance: A corpus-driven comparative study of laws in the European union and China. *Int. J. Leg. Discourse* 6 (2), 349–379. doi:10.1515/ijld-2021-2059

Mallapaty, S. (2022). China expands control over genetic data used in scientific research. *Nature* 605 (7910), 405. doi:10.1038/d41586-022-01230-z

Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., et al. (2021). The OpenNeuro resource for sharing of neuroscience data. *ELife* 10, e71774. doi:10.7554/eLife.71774

Markowetz, F. (2015). Five selfish reasons to work reproducibly. *Genome Biol.* 16 (1), 274. doi:10.1186/s13059-015-0850-7

Maumet, C., Auer, T., Bowring, A., Chen, G., Das, S., Flandin, G., et al. (2016). Sharing brain mapping statistical results with the neuroimaging data model. *Sci. Data* 3 (1), 160102. Article 1. doi:10.1038/sdata.2016.102

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., et al. (2016). How open science helps researchers succeed. *ELife* 5, e16800. doi:10.7554/eLife.16800

Mennes, M., Biswal, B. B., Castellanos, F. X., and Milham, M. P. (2013). Making data sharing work: The FCP/INDI experience. *NeuroImage* 82, 683–691. doi:10.1016/j.neuroimage.2012.10.064

Meyer, M. N. (2020). There oughta Be a law: When does(n't) the U.S. Common rule apply? *J. Law, Med. Ethics* 48 (S1), 60–73. doi:10.1177/1073110520917030

Milham, M. P., Craddock, R. C., Son, J. J., Fleischmann, M., Clucas, J., Xu, H., et al. (2018). Assessment of the impact of shared brain imaging data on the scientific literature. *Nat. Commun.* 9 (1), 2818. Article 1. doi:10.1038/s41467-018-04976-1

Moreau, L., Groth, P., Cheney, J., Lebo, T., and Miles, S. (2015). The rationale of PROV. *J. Web Semant.* 35, 235–257. doi:10.1016/j.websem.2015.04.001

National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. doi:10.17226/25303

Niso, G., Botvinik-Nezer, R., Appelhoff, S., De La Vega, A., Esteban, O., Etzel, J. A., et al. (2022). Open and reproducible neuroimaging: From study inception to publication. *NeuroImage* 263, 119623. doi:10.1016/j.neuroimage.2022.119623

Niso, G., Gorgolewski, K. J., Bock, E., Brooks, T. L., Flandin, G., Gramfort, A., et al. (2018). MEG-BIDS, the brain imaging data structure extended to magnetoencephalography. *Sci. Data* 5 (1), 180110. Article 1. doi:10.1038/sdata.2018.110

Niso, G., Rogers, C., Moreau, J. T., Chen, L.-Y., Madjar, C., Das, S., et al. (2016). Omega: The open MEG archive. *NeuroImage* 124, 1182–1187. doi:10.1016/j.neuroimage.2015.04.028

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., et al. (2022). Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol.* 73 (1), 719–748. doi:10.1146/annurev-psych-020821-114157

Open Science Collaboration (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* 349 (6251), aac4716. doi:10.1126/science.aac4716

Paret, C., Unverhau, N., Feingold, F., Poldrack, R. A., Stirner, M., Schmahl, C., et al. (2022). Survey on open science practices in functional neuroimaging. *NeuroImage* 257, 119306. doi:10.1016/j.neuroimage.2022.119306

Pernet, C. R., Appelhoff, S., Gorgolewski, K. J., Flandin, G., Phillips, C., Delorme, A., et al. (2019). EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Sci. Data* 6 (1), 103. Article 1. doi:10.1038/s41597-019-0104-8

Pernot-Leplay, E. (2020). China's approach on data privacy law: A third way between the US and the EU? *Penn St. JL Int'l Aff.* 8, 49.

Poldrack, R. A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., et al. (2011). The cognitive atlas: Toward a knowledge foundation for cognitive neuroscience. *Front. Neuroinformatics* 5, 17. doi:10.3389/fninf.2011.00017

Powell, K. (2021). The broken promise that undermines human genome research. *Nature* 590 (7845), 198–201. doi:10.1038/d41586-021-00331-5

Prasser, F., Kohlmayer, F., Lautenschläger, R., and Kuhn, K. A. (2014). Arx - a comprehensive tool for anonymizing biomedical data. *AMIA Annu. Symp. Proc.*, 2014, 984–993.

Price, W. N., and Cohen, I. G. (2019). Privacy in the age of medical big data. *Nat. Med.* 25 (1), 37–43. Article 1. doi:10.1038/s41591-018-0272-7

Price, W. N., Kaminski, M. E., Minssen, T., and Spector-Bagdady, K. (2019). Shadow health records meet new data privacy laws. *Science* 363 (6426), 448–450. doi:10.1126/science.aav5133

Robbins, K., Truong, D., Appelhoff, S., Delorme, A., and Makeig, S. (2021). Capturing the nature of events and event context using hierarchical event descriptors (HED). *NeuroImage* 245, 118766. doi:10.1016/j.neuroimage.2021.118766

Rood, J. E., and Regev, A. (2021). The legacy of the human genome project. *Science* 373 (6562), 1442–1443. doi:10.1126/science.abl5403

Rosati, K. B. (2022). Legal compliance and good data stewardship in data sharing plans. *Harv. Data Sci. Rev.* 4 (3). doi:10.1162/99608f92.5ff070bf

Schwarz, C. G., Kremers, W. K., Therneau, T. M., Sharp, R. R., Gunter, J. L., Vemuri, P., et al. (2019). Identification of anonymous MRI research participants with face-recognition software. *N. Engl. J. Med.* 381 (17), 1684–1686. doi:10.1056/NEJMc1908881

Spector-Bagdady, K. (2021). Governing secondary research use of health data and specimens: The inequitable distribution of regulatory burden between federally funded and industry research. *J. Law Biosci.* 8 (1), lsab008. doi:10.1093/jlb/lsab008

Staunton, C., Slokenberga, S., Parziale, A., and Mascalzoni, D. (2022). Appropriate safeguards and article 89 of the GDPR: Considerations for Biobank, databank and genetic research. *Front. Genet.* 13, 719317. doi:10.3389/fgene.2022.719317

Teeters, J. L., Godfrey, K., Young, R., Dang, C., Friedsam, C., Wark, B., et al. (2015). Neurodata without borders: Creating a common data format for Neurophysiology. *Neuron* 88 (4), 629–634. doi:10.1016/j.neuron.2015.10.025

Wagner, A. S., Waite, L. K., Meyer, K., Heckner, M. K., Kadelka, T., Reuter, N., et al. (2021). The DataLad Handbook (v0.14) [computer software]. *Zenodo*. doi:10.5281/zenodo.4495560

Wang, C., Zhang, J., Lassi, N., and Zhang, X. (2022). Privacy protection in using artificial intelligence for healthcare: Chinese regulation in comparative perspective. *Healthcare* 10 (10), 1878. Article 10. doi:10.3390/healthcare10101878

Wang, L. (2022). Fundamental issues in the protection of sensitive personal information in the context of the interpretation of the civil code and the personal information protection law (敏感个人信息保护的基本问题-以《民法典》和《个人信息保护法》的解释为背景). *Contemp. Law Rev. (*当代法学*)* 3 (10), 1.

White, T., Blok, E., and Calhoun, V. D. (2022). Data sharing and privacy issues in neuroimaging research: Opportunities, obstacles, challenges, and monsters under the bed. *Hum. Brain Mapp.* 43 (1), 278–291. doi:10.1002/hbm.25120

Wiebe, A. (2020). "Datenschutz, Big Data und KI im Gesundheitswesen," in *Festschrift für Jürgen Taeger*. Editor U. A. Specht (Oldenburg: RuW-Suche).

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3 (1), 160018. Article 1. doi:10.1038/sdata.2016.18

Wolf, S. M., and Evans, B. J. (2018). Return of results and data to study participants. *Science* 362 (6411), 159–160. doi:10.1126/science.aav0005

Yarkoni, T., Markiewicz, C. J., de la Vega, A., Gorgolewski, K. J., Salo, T., Halchenko, Y. O., et al. (2019). PyBIDS: Python tools for BIDS datasets. *J. Open Source Softw.* 4 (40), 1294. doi:10.21105/joss.01294

frontiers | Frontiers in Genetics

TYPE Opinion
PUBLISHED 30 May 2023
DOI 10.3389/fgene.2023.1169580

Check for updates

OPEN ACCESS

EDITED BY
Manuel Corpas,
University of Westminster,
United Kingdom

REVIEWED BY
Stuart McLennan,
Technical University of Munich, Germany
Marie-Christine Fritzsche,
Technical University of Munich Munich,
Germany, in collaboration with
reviewer SM

*CORRESPONDENCE
Aviad Raz,
aviadraz@bgu.ac.il

RECEIVED 19 February 2023
ACCEPTED 17 May 2023
PUBLISHED 30 May 2023

CITATION
Raz A and Minari J (2023), AI-driven risk scores: should social scoring and polygenic scores based on ethnicity be equally prohibited?
Front. Genet. 14:1169580.
doi: 10.3389/fgene.2023.1169580

COPYRIGHT
© 2023 Raz and Minari. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

AI-driven risk scores: should social scoring and polygenic scores based on ethnicity be equally prohibited?

Aviad Raz[1]* and Jusaku Minari[2]

[1]Department of Sociology and Anthropology, Ben-Gurion University of the Negev, Beer-Sheba, Israel, [2]Uehiro Research Division for iPS Cell Ethics, Center for iPS Cell Research and Application (CiRA), Kyoto University, Kyoto, Japan

KEYWORDS

polygenic risk score (PRS), artificial intelligence, stratificati on, ethnicity, transparency

Introduction

Big data, juxtaposing genetic, clinical, and socio-demographic information, forms the basis for research on various health risk correlations in precision/personalized medicine. In this context, artificial intelligence (AI) has recently been used to improve polygenic risk score (PRS). Polygenic risk scores provide a measure of individual disease risk based on one's genome-wide information, with a particular focus on a statistical calculation of multiple genomic variants[1]. The development of population-level genetic studies, such as genome wide association studies (GWAS), has accelerated the development of PRS as part of genomic research. This characteristic, where PRS is based on a particular population, leads to an inherent need to avoid overfitting and underfitting and to address diversity in the development of the scores. Previous studies comparing PRS predictive accuracy for biobank data from different countries have shown that genetic prediction accuracy (based on UK biobank data) was far lower in non-European populations. Indeed, it was 2.5-fold lower in East Asians and 4.9-fold lower in Africans, on average (Martin et al., 2019). This poorer predictive power of PRS in non-European populations, particularly among African ancestry individuals, is most likely due to them being underrepresented within the training data. In the same vein, PRS for breast cancer in African American women based largely on variants identified in European-ancestry populations show poor performance, as DNA susceptibility loci are not similar across race/ethnicity, and have indeed been shown to differ most often for individuals of African ancestry because of their considerably greater genetic diversity (Feng et al., 2017). The way in which each individual variant affects the polygenic score can vary from study to study, adding to the complexity. In addition, using AI for PRS increases the complexity of ethical and social challenges, especially when electronic health records are integrated (Fritzsche et al., 2023). While research on PRS is ongoing, its clinical validity is still debated (Slunecka et al., 2021). Nevertheless, commercial genomic sequencing laboratories are already offering an array of both clinical and direct-to-consumer tests that include PRS as part of their risk prediction products for a variety of diseases and conditions (James et al., 2021).

1 https://www.genome.gov/genetics-glossary/Polygenic-Risk-Score.

We wish to draw attention to the importance and timeliness of comparing some key issues that are more broadly emerging from the proposed EU AI Act, especially regarding the banning of "social scoring" in AI systems, with the ethical concerns related to PRS. In particular, when used as a form of ethnicity-related genomic scoring, PRS with poorer predictive power in underrepresented populations could exacerbate ethnically based health discrimination as well as reinforce a reckoning with the relevance of self-reported race, ethnicity, and ancestry, and the relationship of such biomarkers and risk factors to disease diagnoses. PRS is geared primarily toward healthcare/medicine whereas social scoring is used in various areas (e.g., education, finance, insurances, migration, etc.), as well as in healthcare/medicine. While PRS are also developed for other areas than healthcare, e.g., for educational purposes (Merz et al., 2022), such educational attainment polygenic scores are similarly vulnerable to biases due to stratification, thus again highlighting the need for the critical reflection raised in this opinion. While PRS is not the same as social scoring, highlighting the differences and similarities will open up the interface of AI and health risk construction to an even wider audience.

Criticism of PRS in the context of ethnic/ancestry traits

There is no well-established *genetic* basis for distinctly stratifying human populations by ethnicity (Mersha and Beck, 2020). However, adding parameters of ethnicity to the calculation of polygenic risk scores may reveal statistical correlations and thus interest researchers. It is now widely accepted that most of the genetic diversity in the human species exists between individuals within populations and that only a small fraction of the total genetic diversity is related to variation between ethnic populations (Kaplan & Fullerton, 2022). As geneticist Richard Lewontin (1972) famously asserted, these features of human genetic variation mean that racial classification is of "virtually no genetic or taxonomic significance" and hence should be abandoned. Recently, there are calls for building genetic literacy through education that uses population thinking and multifactorial genetics to refute genetic essentialist beliefs about race (Little et al., 2022). However, with PRS targeting "risk groups," we are currently witnessing the resurfacing of traditional social groupings like ethnicity and race, re-charged by genomic designations. When risk estimates are applied to patients stratified by self-identified race and/or ethnicity, it may result in a range of consequences, despite the often-unprecise designation of "ethnicity" and its confluence with ancestry (James et al., 2021). Clinical use of PRS could exacerbate race-based health disparities and reinforce systemic biases of self-reported race, ethnicity, and ancestry as biomarkers and risk factors to disease diagnoses (Lewis and Green, 2021). While many common complex traits and diseases differ in their prevalence between racial and/or ethnic groups, particularly in the United States, this has been shown to be the result of pronounced racial and ethnic health disparities rather than genetic differences (Yearby et al., 2022). These concerns regarding the social/ethnic aspects of PRS echo recent concerns about AI-driven social scoring.

Criticism of AI-driven social scoring

The proposed EU AI Act (2021) explicitly bans AI system use by public authorities (expected to later include also the private sectors) for social scoring purposes. Social scoring in this context means using an AI system to evaluate the trustworthiness of individuals based on their behaviors or personal characteristics, leading to stratified treatment of individuals. Adherence to public health measures can affect social scoring, for example, following quarantine measures or receiving vaccinations (Meszaros et al., 2022). The proposed Act explains this as follows:

AI systems providing social scoring of natural persons for general purpose by public authorities or on their behalf may lead to discriminatory outcomes and the exclusion of certain groups. They may violate the right to dignity and non-discrimination and the values of equality and justice. Such AI systems evaluate or classify the trustworthiness of natural persons based on their social behaviour in multiple contexts or known or predicted personal or personality characteristics. The social score obtained from such AI systems may lead to the detrimental or unfavourable treatment of natural persons or whole groups thereof in social contexts, which are unrelated to the context in which the data was originally generated or collected or to a detrimental treatment that is disproportionate or unjustified to the gravity of their social behaviour. Such AI systems should be therefore prohibited. (EU, 2021, article 17, p. 21).

The proposed AI Act lists high-risk AI systems in areas that include, for example, biometric identification and categorization of natural persons, law enforcement, as well as migration, asylum, and border control management. The China social credit system, which allegedly rates individuals based on the aggregation and analysis of data concerning their past behaviors, would be banned by the EU Act, if it indeed uses social scoring.

Discussion

Polygenic risk scores (PRS) and social scoring are two different concepts. PRS are used primarily in medical research and do not involve any evaluation of an individual's behavior or personal characteristics, but rather are based solely on genetic data. Social scoring, on the other hand, refers to a system of evaluating individuals based on various social and behavioral factors, such as their credit score, online activity, criminal record, or other personal data. However, both may reproduce biases. The concerns raised here could be used to develop a critique of how AI for genomic risk stratification in healthcare/medicine should not only be regulated for representativeness of human diversity but perhaps also for potential amplification of social scoring. This is especially important, as there may be a risk of drawing conclusions from PRS about causal relationships too quickly and with insufficient knowledge of statistics and causality/correlation claims (Fritzsche et al., 2023). By lowering the statistical standards for regarding a marker as trait-associated, weighting associations by estimated effect sizes, and aggregating associations over a larger number of variants, predictive accuracy may be increased at the expense of explainability, as any clear

etiological link between specific genetic changes and the phenotype of interest is obscured.

By banning social scoring as an unacceptable risk, the proposed AI Act aims to go beyond the technical robustness, privacy, and safety required by the General Data Protection Regulation (GDPR) to prevent or minimize the probability of unintentional harm in processing personal data by AI systems. The AI Act does not directly mention AI-driven PRS. Nevertheless, in addition to specifying several unacceptable risks, it establishes the goal of minimizing the risk of erroneous or biased AI-assisted decisions in critical areas, including healthcare. We must hence carefully consider PRS in the light of minimizing the risk of erroneous or biased AI-assisted decisions. Arguably, there are three major foci in the proposed AI Act that are relevant to both social scoring and polygenic scores: transparency, non-discrimination, and accountability.

(1) Transparency: The "right to explanation" formulated in the GDPR and the proposed EU AI Act require that AI systems be explainable for high-risk decision making (EU, 2021). The "black box" conundrum is manifested in the context of scoring through the non-explainable relationships between individual genomic variants, PRS and diseases phenotype, similar to the relationships between individual "accountability", obtained/accessible personal data, and social scoring.
(2) Non-discrimination: AI systems must collect diverse data to avoid bias and prevent the uncertain decision-making and unjust use of such data toward different populations. This requirement is critical in the case of ethnicity-based PRS due to careful consideration of the diversity of the ethnicity.
(3) Accountability: Certain actors, such as the government, health maintenance organizations, or health insurance companies, should be held responsible for the unintended consequences of individual's actions using PRS. For example, who is responsible if a PRS-based model for breast cancer screening leads to precluding a patient from accessing screening, or has the responsibilities of harm by improper screening due to risk scores that are wrongfully produced based on race and ethnicity?

Social scoring is used in various areas as well as in healthcare/medicine, but for the sake of comparison we focus here on its use in healthcare/medicine, which is the primary area of PRS. If AI-derived PRS evaluates or classifies the risk of natural persons based on their ethnic/racial self-designation (or practitioner-designated), it would be akin to AI-derived social scoring based on previous social behaviours in multiple contexts or known or predicted personal or personality characteristics. The ethnicity related PRS obtained from such AI systems may therefore lead to the detrimental or unfavourable treatment of natural persons or whole groups of persons in healthcare contexts. Further, if the model of PRS-based screening is adopted as standard clinical practice, and if risk scores are produced based on race and ethnicity, it could lead to under- or over-screening. The purpose and implications of the classification must be clear to both those making the classification and those being classified. Social scoring can be wrong due to being based on previous behaviours that are unrelated to the context of scoring or to a detrimental treatment. Ethnicity-related PRS can be wrong because of being based on ethnic/ancestry traits that are similarly unrelated to the context of scoring or to a detrimental treatment. In this case, both AI systems should thus be equally prohibited.

Author contributions

AR and JM both made substantial, direct contributions and approved the final version of the manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

We are grateful for the funding provided by the JSPS–ISF Joint Program, grant 62/22 (ISF), JPJSBP120228404 (JSPS), "Biobanks for genomic medicine in Israel and Japan: An analysis of ethics and policy".

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

EU (2021). Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

Feng, Y., Rhie, S. K., Huo, D., Ruiz Narvaez, E. A., Haddad, S. A., Ambrosone, C. B., et al. (2017). Characterizing genetic susceptibility to breast cancer in women of African ancestry. *Cancer Epidemiol. Biomarkers Prev.* 26 (7), 1016–1026.

Fritzsche, M. C., Akyüz, K., Cano Abadía, M., McLennan, S., Marttinen, P., Mayrhofer, M. T., et al. (2023). Ethical layering in AI-driven polygenic risk scores – new complexities, new challenges. *Front. Genet.* 14, 1098439. doi:10.3389/fgene.2023.1098439

James, J. E., Riddle, L., Koenig, B. A., and Joseph, G. (2021). The limits of personalization in precision medicine: Polygenic risk scores and racial categorization in a precision breast cancer screening trial. *PLoS ONE* 16 (10), e0258571. doi:10.1371/journal.pone.0258571

Kaplan, J. M., and Fullerton, S. M. (2022). Polygenic risk, population structure and ongoing difficulties with race in human genetics. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 2022. doi:10.1098/rstb.2020.0427

Lewis, A. C. F., and Green, R. C. (2021). Polygenic risk scores in the clinic: New perspectives needed on familiar ethical issues. *Genome Med.* 13, 14. doi:10.1186/s13073-021-00829-7

Lewontin, R. C. (1972). "The apportionment of human diversity," in *Evolutionary biology*. Editors T. Dobzhansky, M. K. Hech, and W. C. Steere (New York: NY: Springer), 381–398.

Little, I. D., Koehly, L. M., and Gunter, C. (2022). Understanding changes in genetic literacy over time and in genetic research participants. *Am. J. Hum. Genet.* 109 (12), 2141–2151. doi:10.1016/j.ajhg.2022.11.005

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 51 (4), 584–591. doi:10.1038/s41588-019-0379-x

Mersha, T. B., and Beck, A. F. (2020). The social, economic, political, and genetic value of race and ethnicity in 2020. *Hum. Genomics* 14 (1), 37. doi:10.1186/s40246-020-00284-2

Merz, E. C., Strack, J., Hurtado, H., Vainik, U., Thomas, M., Evans, A., et al. (2022). Educational attainment polygenic scores, socioeconomic factors, and cortical structure in children and adolescents. *Hum. Brain Mapp.* 43 (16), 4886–4900. doi:10.1002/hbm.26034

Meszaros, J., Minari, J., and Huys, I. (2022). The future regulation of artificial intelligence systems in healthcare services and medical research in the European Union. *Front. Genet.* 13, 927721. doi:10.3389/fgene.2022.927721

Slunecka, J. L., van der Zee, M. D., Beck, J. J., Johnson, B. N., Finnicum, C. T., Pool, R, et al. (2021). Implementation and implications for polygenic risk scores in healthcare. *Hum. Genomics* 15 (1), 46. doi:10.1186/s40246-021-00339-y

Yearby, R., Clark, B., and Figueroa, J. F. (2022). Structural racism in historical and modern US health care policy. *Health Aff.* 41 (2), 187–194.

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 26 June 2023
DOI 10.3389/fgene.2023.1039839

Check for updates

OPEN ACCESS

EDITED BY
Sivia Barnoy,
Tel Aviv University, Israel

REVIEWED BY
Karsten Weber,
Regensburg University of Applied Sciences, Germany
Giovanni Rubeis,
Karl Landsteiner University of Health Sciences, Austria

*CORRESPONDENCE
Silke Schicktanz,
sschick@gwdg.de

RECEIVED 08 September 2022
ACCEPTED 23 May 2023
PUBLISHED 26 June 2023

CITATION
Schicktanz S, Welsch J, Schweda M, Hein A, Rieger JW and Kirste T (2023), AI-assisted ethics? considerations of AI simulation for the ethical assessment and design of assistive technologies. *Front. Genet.* 14:1039839. doi: 10.3389/fgene.2023.1039839

COPYRIGHT
© 2023 Schicktanz, Welsch, Schweda, Hein, Rieger and Kirste. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

AI-assisted ethics? considerations of AI simulation for the ethical assessment and design of assistive technologies

Silke Schicktanz[1,2]*, Johannes Welsch[1], Mark Schweda[3], Andreas Hein[4], Jochem W. Rieger[5] and Thomas Kirste[6]

[1]University Medical Center Göttingen, Department for Medical Ethics and History of Medicine, Göttingen, Germany, [2]Hanse-Wissenschaftskolleg, Institute of Advance Studies, Delmenhorst, Germany, [3]University of Oldenburg, Department of Health Services Research, Division for Ethics in Medicine, Oldenburg, Germany, [4]University of Oldenburg, Department of Health Services Research, Division Assistance Systems and Medical Device Technology, Oldenburg, Germany, [5]University of Oldenburg, Applied Neurocognitive Psychology Lab, Oldenburg, Germany, [6]University of Rostock, Institute for Visual and Analytic Computing, Rostock, Germany

Current ethical debates on the use of artificial intelligence (AI) in healthcare treat AI as a product of technology in three ways. First, by assessing risks and potential benefits of currently developed AI-enabled products with ethical checklists; second, by proposing ex ante lists of ethical values seen as relevant for the design and development of assistive technology, and third, by promoting AI technology to use moral reasoning as part of the automation process. The dominance of these three perspectives in the discourse is demonstrated by a brief summary of the literature. Subsequently, we propose a fourth approach to AI, namely, as a methodological tool to assist ethical reflection. We provide a concept of an AI-simulation informed by three separate elements: 1) stochastic human behavior models based on behavioral data for simulating realistic settings, 2) qualitative empirical data on value statements regarding internal policy, and 3) visualization components that aid in understanding the impact of changes in these variables. The potential of this approach is to inform an interdisciplinary field about anticipated ethical challenges or ethical trade-offs in concrete settings and, hence, to spark a re-evaluation of design and implementation plans. This may be particularly useful for applications that deal with extremely complex values and behavior or with limitations on the communication resources of affected persons (e.g., persons with dementia care or for care of persons with cognitive impairment). Simulation does not replace ethical reflection but does allow for detailed, context-sensitive analysis during the design process and prior to implementation. Finally, we discuss the inherently quantitative methods of analysis afforded by stochastic simulations as well as the potential for ethical discussions and how simulations with AI can improve traditional forms of thought experiments and future-oriented technology assessment.

KEYWORDS

intelligent assistive technology, ethical reflection, simulation, person with dementia, conceptual approach

1 Introduction

In the science fiction movie Dark Star (1974, director John Carpenter), the captain of a starship argues with an artificial intelligence (AI)-controlled bomb about whether it should detonate. The dispute is whether the bomb's decision to detonate itself is based on correct data, namely, a correct order perceived by the bomb's sensory input. The captain's arguments regarding the limits of what the bomb can know about its own existence or intelligence, however, just serve to convert the bomb into a nihilist. Finally, it detonates itself and kills the ship's human crew.

Such movie scenes function like thought experiments, a common methodology in philosophy and science, and can help us anticipate implications or test for the ethical or epistemic coherence (Walsh, 2011) of an argumentation or idea. In the Dark Star case, the risk is that future AI could develop its own morality, with harmful outcomes for humans. However, it is never clear whether the argumentation generated by thought experiments translates to real situations, for example, to AI systems currently under development for dementia care (Schweda et al., 2019). Our goal in this paper is to expand the thought-experiment approach by exploring the opportunities and rationales for using computational simulation as a tool for ethical reflection on human-AI interaction. Our idea is that such algorithmic simulations can augment ethical reflection with empirical, simulated data during the design phase of systems, thereby improving the anticipation of ethical problems in the use of AI technology in various settings.

Following the High Level Expert Group of Artificial Intelligence (2019), we define AI-systems as software systems that analyze their environment and take actions to achieve some goals independently. This general definition does not predefine the type of mathematics and algorithms implemented—e.g., symbolic rule-based or sub-symbolic AI-like neuronal networks—nor does it require specifics on the level of automation (Shneiderman, 2021, *p*. 48).

Ethics in technology development is traditionally guided by general principles that can be employed in thought experiments to test which main principles seem to have what consequences or are more likely to gain public acceptance. The employment of such thought experiments, e.g., trolley-dilemma experiments for exploring ethical aspects of automated vehicles, recently have been criticized as too narrow or abstract (e.g., Goodall, 2019; De Freitas et al., 2021). Furthermore, empirical validity in such thought experiments is often low, and reasoning can be biased by the prejudice of ethicists or technology developers.

Overall, mainstream ethical evaluation approaches regarding new technologies, such as biotech, nanotech or artificial intelligence, tend to conceptualize technology as a mere object of ethical reflection in terms of the "ethics of AI." While this makes sense for biosciences or nanotechnology, it need not be the only way to reflect on AI.

Empirically informed ethical reasoning is a more recently established standard with the potential to significantly reduce bias–including expert bias–and to improve generalizability to real-world situations (Schicktanz et al., 2012; Mertz et al., 2014). Therefore, the social and moral perspectives about, for example, genetic, biological, nano, or AI technology as held by practitioners, stakeholders, and affected persons are collected empirically using qualitative methods. Although this empirically informed approach has some advantages in comparison to traditional, non-empirical methods, it also has some epistemic limitations.

In applied ethics, experts often think of AI as a feature of specific products, be it a feature that analyzes the environment and adapts actions to reach a particular goal or a feature that helps to make (moral) decisions. The dominance of this approach in the discourse is evident from a review of the literature (see below, Section 2). As AI technology often has complex or even hidden outcomes, it has been argued recently that "explainability" and "trust" are essential criteria for ethical evaluation (High Level Expert Group of Artificial Intelligence, 2019; Amann et al., 2020; Coeckelbergh, 2020; Markus et al., 2021; Border and Sarder, 2022). However, explainability and trust focus again on a human-AI interaction, again conceptualizing AI mainly as an end-product and humans as being capable of understanding it. This assumption does not always hold, e.g., when in healthcare and disability settings. Here, one cannot guarantee that the users of technology and the people it affects are able to monitor, interact with, or to understand an AI system's outputs. These individuals cannot "trust" the system because trust requires specific cognitive and emotional features.[1] The aim of our "AI-Assisted Ethics" approach is to anticipate ethical trade-offs and social implications in complex, contextualized settings where criteria such as trust or explainability might not function or are not appropriate. Furthermore, complexity is particularly important in situations where direct anticipation of outcomes and implications is limited, e.g., because the characteristics of people involved are very heterogeneous. In these cases, individuals might interact with the AI very differently, which may turn greatly restrict the generalizability of empirical observations of behavior and values to other individuals.

Our article results from a truly interdisciplinary cooperation between ethicists, social scientists, engineers, and machine learning specialists. It combines insights from various sub-studies that built on each other, with a specific focus on intelligent assistive technologies (IAT) in healthcare settings and especially for the care of persons with dementia. The following proposal has been developed from the comprehensive exchange between these sub-studies. First, we summarize the main strands of the general discussion regarding AI ethics and machines ethics as well as insights regarding the ethics of human-AI interaction in the particular setting of care for older people and persons with dementia (2). As the assessment of impacts of AI-technology in cases such as dementia care is often difficult or risks to neglect the complexity of values and interactions of involved agents, we developed in a next step a conceptual approach to consider AI as a tool (in a simulation) to anticipate ethical and social issues of implementing IAT (3). Here, we focus on the context of dementia care (3.1., see also info box). The concept of such an *in silico* simulation (3.2) considers multiple agents who can interact in various ways with different ethical values. By an 'Ethical Compliance Quantification' evaluation different design alternatives can be quantitatively compared and can inform stakeholder discussions. Hence, results from an exemplary simulation model to test for the developed ethical compliance quantification are presented to illustrate the conceptual approach (3.3). Hereby we construct an example from research in technology-assisted dementia care to discuss the advantages and challenges of this approach. This simulation is informed by value statements drawn from interviews. It utilizes stochastic human behavior models that encompass behavioral data,

1 Therefore, language-based assistive systems to foster a "dialog" between human and machine are neither meaningful nor appropriate.

pre-set values for simulating realistic settings, variables, data sets and setting variables. Finally, we discuss the differences between qualitative and inherently quantitative methods of ethical reasoning and how the simulation approach can enhance ethical reasoning for technology assessment (4.), and provide a short conclusion section (5).

2 AI as a product of technology and as an object of ethical reflection

In the following, we summarize the main strands of discussion regarding AI ethics and ethical machines. We then focus on the ethics of human-AI interaction in a particular setting: "intelligent assistive technology" for the care of older people and persons with dementia.[2]

Various authors have developed catalogs of values and ethical principles to aid in this kind of ethical assessment (Currie et al., 2020; Spiekermann, 2016; Umbrello and van de Poel, 2021; van Wynsberghe, 2013; for an overview, see Schicktanz and Schweda, 2021). Prevalent ethical criteria include self-determination, not harming or actively promoting human welfare, privacy, and sustainability (Hofmann, 2013; Novitzky et al., 2015; Ienca et al., 2018; Vandemeulebroucke et al., 2018). In many of these approaches, some ethical principles are prioritized over others. Although these values are sometimes proposed as guides for design processes, more often they are treated as criteria for assessing existing technologies.[3]

This latter fact might have motivated Ienca and colleagues to call for "a coordinated effort to proactively incorporate ethical considerations early in the design and development of new products" (Ienca et al., 2018:1035). In a recent paper on "embedded ethics," McLennan and colleagues also highlight the need for an "ongoing practice of integrating ethics into the entire development process" (McLennan et al., 2022:3) based on a "truly collaborative, interdisciplinary enterprise" (ibid). This is reminiscent of approaches dating back to the 1990s, when engineers and philosophers started to develop strategies for considering ethical issues and values for the design of human-machine interaction. This has been called computer ethics, social informatics, participatory design, and value-in-design. As Friedemann and Kahn 2007 distinguishes, there exist three main ideas about how values and ethical principles are related to the development of new technologies. In the embodied approach, values are incorporated in technology by the designers. In the exogenous approach, values are determined and imposed by the users after a technology is implemented. Interactional approaches focus on the interaction of designers and users; these include approaches like value-in-design and participatory design. Interestingly, all three approaches can be found in current AI-technology design.

Pertinent questions for this kind of technology assessment are as follows. How should we (or how should we not) use AI/IAT technologies? Are there ethically acceptable risks, do opportunities outweigh risks, or might the use of AI/IAT technologies create conflict with basic human rights and ethical principles such as human dignity, self-determination, or justice? In some fields, such as technologies for the care of older people, it seems that this assessment often takes place after a prototype of the technology has been developed, but not during the design process.

By contrast, the central question in machine ethics is whether AI-technologies that can operate more or less autonomously can and should be constructed to operate in a morally acceptable way. This touches upon ethical questions regarding adequate concepts and standards, as well as on the criteria of morality as such. It encompasses issues of moral agency and responsibility, as well as informatics and engineering questions regarding effective technological implementation through algorithms and "training" (Anderson and Anderson, 2007). This debate differentiates between top-down and bottom-up approaches to the problem of implementing morality-sensitive technology (Wallach et al., 2009). Top-down approaches try to specify moral precepts in a deductive manner by means of the successive specification and application of a set of general moral norms. In this vein, fundamental moral philosophical principles such as the utilitarian principle of utility (maximization of utility) or the Kantian categorical imperative (principle of universalizability of maxims) are operationalized in terms of algorithms that constitute the procedural rules of the autonomous technical system, its "moral modus operandi." For example, van Wynsberghe (2013:411-413) sees a fundamental need to endow care robots (which can be considered a special case of assistive systems) with moral values during the development process. By contrast, bottom-up approaches try to specify moral precepts in an inductive manner by developing moral competences through a series of pertinent moral experiences. An example of this can be seen by the MIT moral machine experiment (Awad et al., 2018) by gathering large data sets of

2 Weber states in a recent article that "There is no good and generally accepted definition for age-appropriate assistive systems." (Weber, 2021: 29, own translation). Kunze and König (2017):1), Hofmann (2013:390) and the umbrella association of German health insurers (Spitzenverband der Gesetzlichen Krankenversicherungen in Deutschland, 2019:22) report similar findings. In general, according to Ienca and colleagues, the following definition has become accepted: "Assistive technology is the umbrella term used to describe devices or systems which allow to increase, maintain or improve capabilities of individuals with cognitive, physical or communication disabilities" (Ienca et al., 2017:1302; cf. World Health Organization, 2018; Endter, 2021:15; Novitzky et al., 2015:709; Manzescheke et al., 2013:8). Based on this, Ienca and colleagues describe as "intelligent" those assistive systems "with [their] own computation capability and the ability to communicate information through a network" (Ienca et al., 2017:1302). For our purpose here, we assume that the more complex such IAT systems are, the more relevant our considerations regarding the usage of AI for ethical consideration become.

3 In this context, it is important to distinguish between ethics and morality. Ethics is understood as a philosophical reflection about the meaning and justification of various kinds of normative statements, legal practices, or everyday judgements. By contrast, morality is understood as the everyday application of a set of moral principles, e.g., norms and values, in judgment and decision making. Although this underlying set of principles often remains implicit and unarticulated, human agents are usually able to provide a simple explanation of such norms and values upon request (so it is not fully opaque). From these definitions, it follows that if the artificial, automated system cannot reflect and explain its decisions in an appropriate way in varying situations, it should rather be labeled as a moral machine because while it fulfills the criteria of moral decision-making, it does not fulfill the criteria of ethical reflection. By contrast, to describe a machine truly as an ethical machine would, in analogy to human ethical thinking, require that the criteria of "reflection" are fulfilled. This includes at least four components or stages: a) the potential to revise pre-implemented norms, b) the availability of a set of alternative approaches with an understanding of how they differ, c) discussing the pros and cons of revision, and d) providing a final justification of the final conclusion. It is an open question whether the new standard of explainability in AI would satisfy the criteria of ethical reflection or whether it would remain on the level of just making moral criteria comprehensible.

humans answering online moral dilemma. For AI, this means learning morality. Here, the technical system is not equipped with general rules and a basic moral orientation but is trained rather by repeated confrontations with a variety of pertinent "cases," i.e., moral problems and their solutions, thus emulating the process of human moral development. One might expect such learning processes to result mainly in punishment avoidance in standard learning techniques or to level out practical compromises between different moral opinions. To go beyond this level and reach a coherent ethical framework would require the intellectual capacity to identify new top categories, rules for consistency, and inductive theoretical reflection. This might be beyond the capacities of AI according to some scholars (Brundage, 2014).

Whether moral precepts can be derived through technology and whether deriving moral precepts is a proper and feasible objective of AI has been debated over the last 2 decades (Anderson et al., 2004; Nallur, 2020). Misselhorn (2021) who talks of "algorithm morality" or "artificial morality," favors a "hybrid approach" combing fundamental moral rules (e.g., never harm or kill a human) with AI-based learning of contextual moral rules for interacting with humans (e.g., respecting privacy for person X, and favoring safety issues for person Y). This also allows for the integration of empirical information on actual user preferences. Furthermore, some have proposed to use AI-tools for enhancing human ethical judgment, hence the idea is not to make machines more ethical, but to use AI to improve ethical judgements of humans. For example, Walton (2016) discusses how different methods (Bayesian vs. computational methods) can contribute to testing the plausibility of arguments. This could be also relevant for analytical moral argumentation, even if the author himself does not discuss this option explicitly. Lara and Deckers (2020) provide a theoretical approach to the use of AI as an auxiliary (supportive) system for 'enhancing' human morality. By a Socratic technique, the machine helps the human agent to learn to reason ethically, but the aim is not to delegate decisions to the technical system or to train a system to be compatible with particular values (*p*. 282). Volkman and Gabriels (2023) build on this idea of 'AI mentors' but stress the need for a 'total' socio-technical system "to operate through a diversity of perspectives challenging one another" (*p*. 10). Their general idea of support that strives for more complexity and considers many perspectives shows analogies with our still more specific approach. Our approach refers to a specific field of application where we see particular challenges, as described in the following. We focus here on how to improve the process of ethics-by-design by considering the diversity and uncertainty of moral perspectives during the process. This processual focus is consistent with our deliberative participatory ethics background (Schicktanz et al., 2012). We do not claim that it automatically provides better moral outcomes.

A specific field of human-AI interaction in which the human agents involved differ according to 1) their role (e.g., professional vs. informal caregiver), 2) their values regarding care and assistive technology (e.g., privacy over safety), and 3) their cognitive and emotional capacities is tied to technologies for monitoring and assistance of people with physical and mental impairments, e.g., persons with dementia. These technologies are increasingly equipped with different types of AI and therefore also fit under the term IAT (Ienca et al., 2017). As a review by Löbe and AboJabel (2022) revealed, assessments of risks, benefits, and empowerment for persons with dementia often are undertaken when a prototype is introduced in care settings experimentally to test usability, safety, or social acceptance. Such testing can be understood as *in situ* simulation if the setting is a natural setting or as *in vitro* simulation[4] if conducted in a laboratory that mimics smart homes or care units. *In silico*, noted below, are computational simulations of such settings.

As dementia poses particularly ethical challenges to the use of AI-based monitoring and assistive systems due to limits regarding "classical" informed consent, the possibility of changing values and preferences without clear verbal expression, and the extremely high burden on caregivers, the assessments in this field of application promise to provide highly sensitive insights in fundamental problems regarding the development and use of new technologies in eldercare. In a next step, abstracting the results from dementia care to other, particular sensitive fields of care giving, the approach can be also very fruitful. However, here, dementia is for various reasons (see Section 3.1) a reasonable starting point.

In a previous expert interview study (Welsch, 2022a; Welsch, 2022b; Welsch and Schicktanz, 2022, Abojabal et al., under review) we found that the interviewed experts stressed the fact that providing clear definitions of AI or IAT is difficult. Nevertheless, many interviewees gave specific examples of IATs: reminder systems, orientation systems, smart home applications, and robots. Advanced AI features like machine learning or deep learning is not necessarily a constitutive part of this; existing IAT makes use of traditional algorithms more often. The users and purposes of such IAT have been characterized as quite complex, as these IATs include a wide variety of digital applications which contribute to improving the self-determination, the mobility, the social participation, and, in sum, the quality of life of users. Hence, IAT users are not one homogenous group, but include different, interacting groups–often characterized by having different experiences, values, or preferences–such as people in need of care, family caregivers, other relatives, and professional caregivers. This is an important point to consider for an ethics-by-design approach, as different users may be differently affected and have different moral intuitions about the way IATs should operate. Furthermore, it becomes clear that such technologies can also have multiple goals: self-determination, mobility, quality of life, quality of care, safety, or social participation. This wide range of goals will likely create conflicts during the design phase and in actual use (cf. Schweda et al., 2019; Welsch and Schicktanz, 2022).

While ethics-by-design approaches and, in particular, the participation of future users is often seen as important, however, neurodegenerative diseases—common in old age—pose a major challenge for participatory design approaches, e.g., if people cannot communicate, as in later stages of dementia. This problem is exacerbated by short project durations which prevent the investments of time needed for participation. This points to another serious problem of technology assessment in practice: new technologies are developed, but time and money limitations cut short ethical reflection about their implementation. This can be one motivation for demanding standardized ethical evaluation checklists in technology development. However, the implementation of ethical evaluation checklists and their thorough application appears to be a difficult problem given developers' limited time and the complexity

4 See Chandrasekharan et al., 2013 for the differentiation of in-vitro and *in silico* simulations and thought experiments.

of implementation conditions which involve multiple agents with potentially different goals and communication capacities.

Overall, AI is generally thought of as being integral to IAT products, that is, as a feature of the device or system designed with specific end-users in mind. At the same time, a number of practical desiderates and limitations of a classical ethics-by-design approach have become clear. Here we propose using AI as a tool for solving these challenges by presenting an AI-assisted procedural ethics. This problem has also been addressed by Aliman and Kester (2019) who propose to use various socio-technological feedback loops, e.g., by preemptive simulations, to ethically enhance AI technologies.

3 Methodology: model conceptualization

3.1 Premises regarding the need of AI-assisted ethics for supporting IAT development

In our understanding, ethical reflection about modern technologies, including AI, entails taking the following steps: recognition of problems (not only dilemmas); consideration of relevant facts; knowledge of various ethical approaches, principles, and theories to test for alternative conclusions; testing for consistency with accepted norms (does the application of this rule violate uncontroversial norms?); testing for adequacy (can abstract rules be applied to concrete situations without contorting them?); justification of specific decisions (an aspect of explainability), and finally, societal legitimacy of the whole reflective procedure. Such a process can be called "complex" in that it cannot be replaced by a fixed set of values. Most of the above-mentioned approaches start from *a priori* moral intuitions and theoretical generalizations (such as "values").

In the cases where ethical considerations are applied prior to or during the development of a technology, they have to rely on principles that may be too general for concrete design decisions (a limitation of the top-down models noted above). In order to become more relevant for a concrete design decision, ethical issues must rely on analogies from previous situations which are extrapolated to the new situation. This extrapolation is prone to error, but not all errors are evident before product implementation. Obviously, it would be desirable to fill the gap between too general and too specific (but extrapolated) recommendations for ethical design to better adjust to the needs and goals of users, especially when they are vulnerable as, for example, persons with dementia.

Adapting IAT systems to complex settings–characterized by multiple agents with different goals, varying moral intuitions, and different cognitive states and communication skills–during the design phase requires a different approach.[5] The situation could be improved if human ethical reflection would accompany the design process so that experiments with different designs could be conducted to detect practical moral problems and potential value conflicts. *In situ* experimentation, however, raises other problems. It can be unethical to expose vulnerable people, e.g., those with dementia, to new, prototypical technology. For example, the COACH prompting system intended to assist older adults with dementia with handwashing served only to prompt fear and anger in some cases (Mihailidis et al., 2018). Also the review by Alkadri and Jutai, 2016 concludes that many of such technologies for this target group is weak regarding safety and efficacy (Alkadri and Jutai, 2016). Furthermore, the costs of experimentation can easily exceed available resources. In our field of study, i.e., technology-assisted dementia care, another important challenge needs to be considered: communication between human and AI, now often seen as a solution in which the machine "explains" to humans the criteria used for a decision, is not feasible. In contrast to the scene in Dark Star discussed above, persons with dementia have very varying and limited capacities for effectively communicating with a machine. Nor is this group able to give detailed comments to designers or scholars,[6] hence interactive approaches such as participatory design are limited.

These problems lead us to follow the idea of re-thinking AI as an integral tool of the ethical design process, not just a product of technology. Thus we propose to use *in silico* simulation, which is a computational simulation of the technology in its environment as a proxy for *in situ* experimentation. Ideally, these simulations should encompass multiple human agents, a representation of their goals, an individualized model of their internal decision making (from deterministic to stochastic models), and their environment including the device or procedure under development. Simulations can be run repeatedly at little cost and without unethical exploitation of people. The simulations can serve to assess the effects of a product on agents in a setting while varying inputs. Hence, they would allow reflection on the model-building process (Chandrasekharan et al., 2013:242). Other forms, such as *in situ* experiments or thought experiments, focus on the outcome with questions of ethical acceptability, inefficiency, and safety.

As we suggest using simulation for gaining insight and somewhat oppose it to experiments, it is necessary to briefly reflect on the epistemic advantage of using simulation in our setting. There is a substantial debate on this issue, as there are scholars who significantly question the epistemic benefit of simulations in comparison to experiments and other that take the opposite position (see for instance the positions taken and the sources reviewed in Peck, 2004; Parke, 2014; Di Paolo et al., 2000). The relation between simulation and experiment in general is subject to a multi-faceted discussion (see Winsberg, 2022 for an overview). A simulation of a real-world phenomenon based on a mathematical model of this phenomenon may be considered inferior for two reasons (a) the mathematical model may be deficient or (b) the simulation algorithm may require simplifications (such as discretization) that limit precision, up to

5 That this is a complex situation for which the persons involved require training has also been proposed by projects that try to develop simulations of patients with Alzheimer dementia for training facility staff, e.g., "Virtual Patient Simulation Tool for Training Health and Social Care Staff Working with People with Alzheimer's Disease or Related dementia–VIRTUALZ" https://anr.fr/Project-ANR-17-CE19-0028.

6 Such as children, persons with dementia, persons with severe cognitive impairments, or persons with very limited communication skills.

the point of unlimited divergence between simulation results and model content[7]. Considering the first issue, we think that the discussion reflected above does not pertain to the use of simulation we consider. With the simulation model we propose here we do not strive to test existing theories or develop new theories of real-world phenomena. Rather, we suggest to use simulation for analyzing the implications and stepwise construction and explication of a normative model. The normative model we consider consists of ethical value dimensions, the set of events that are being considered relevant with respect to these ethical value dimensions, and mathematical operationalizations of how events are to be quantified (as scores) with respect to values. This mathematical description is the model that is the object of investigation. The method for experimenting with mathematical models indeed is the simulation. This reflects the definition of the term "simulation" already given in Korn and Wait (1978) that a simulation is an experiment performed on a model.

This provides additional clarification to our position as far as this paper is concerned: our main claim is not, that simulation helps to faithfully analyze the real-world effect of an IAT on a set of ethical values. We rather do propose that a simulation is helpful for the iterative development of an ethical value model for IAT design in the first place, for analyzing its implications, and for gaining insight into the consistency or even existence of a value model; we will come back to this in the discussion in Section 4. In addition, we think it is important that the need to provide a mathematical description of the ethical value orientation forces assumptions to be made explicit and thus made accessible to critical review–this is a property that thought experiments do not necessarily have.

Eventually, we also want to develop IAT that provide optimal assistive strategy with respect to a given mathematical model of the values. If one assumes such a value model to exist, this then is conceptually a surprisingly well-defined task, as it can be framed as a standard engineering-level optimization problem. In this paper, we suggest that both tasks can be solved in the same framework. However, as we will see in the example discussed below (Section 3.2.3), strategy optimization may be more sensitive to simulation validity: in strategy optimization, the distribution of events in the state space must adequately reflect the real world in order to correctly identify the optimum. This is the topic of issue (b) identified above. We are confident that the study we discuss below does not fulfill this stronger requirement. However, considering the success of simulation in much more complex situations (see, e.g., Bicher et al., 2021), we are confident that it is possible to build models of adequate validity.

In the following, we give an example of how an AI-assisted simulation can work. The simulation is situated in the field of IAT for dementia care and is a system that guides persons with dementia who have lost their orientation inside a care facility. It illustrates the complexity of the situations that should be considered and what kind of assessment loops are conceivable (see Info Box 1). This example is a conceptual proposition that can be adopted to other settings. We do not claim that the current model has the optimal structure, parameters, or even sampling strategy.

3.2 A concept for an AI-assisted simulation

Our AI-assisted ethics simulation (Figure 1) comprises the several elements, explained below in turn.

3.2.1 Multiple agents

A multi-agent simulation environment provides a simulated world where simulated agents can interact. "Simulation" means that the state of the world and the state of the agents in this world can be represented by a set of variables in a programming language such that the values of these variables (referred to here as the variable "score"[8]) represent the state of the simulated world at any given time. There might, for instance, be a variable called "location" that contains two scores that indicate the location of an agent in a two-dimensional simulation world. The simulation proceeds in steps, where at each step the set of variables is manipulated according to the rules that define the temporal evolution of this simulated world. There might for instance be a "move" rule that an agent whose "destination" variable contains a position that is not equal to the "location" variable will update its location by an amount of "step length" in direction of "destination." Eventually, these rules are represented as pieces of program code.

The interaction of agents is modeled by rules that depend on (and change) the variables representing the state of two or more agents. In general, simulation environments allow the definition of stochastic rules, whereby the outcome depends on a sampling of some random process. For instance, the step length used in a specific application of the "move" rule may be given by sampling from a normal distribution defined by a mean step length and a certain standard deviation. Specifically, in simulations where agents represent humans, such stochastic rules are important for simulating behavior non-deterministically.

A simulation run is produced by initializing the state variables (e.g., location of the simulated patient and current disorientation state, locations of simulated caregivers, etc.) with pre-defined scores (such as the location coordinates) and then by stepwise advancing simulation until the simulation state fulfills a specified termination condition (such as having reached a certain simulation time point or reaching a certain simulation state). If the simulation uses stochastic rules, different "runs" of the simulation may result from the same initial conditions. Based on many runs, it then becomes possible to analyze the statistical properties of state variables in the simulation and their temporal development by combining the records in the run protocols. For instance, one could estimate the expected number of steps required to reach a given destination from a given starting point by averaging the step counts obtained from multiple run

7 Consider the 'Attofox'- problem (Mollison, 1991); but note that this is an illustration of the *opposite* situation: the discretization is more realistic than the continuous model.

8 Note that the term "value" has in our simulation setting two meanings: "variable value" and "ethical value". By "variable value" we mean the quantity stored in a variable of the simulation model. By ethical value we mean normative concepts that have a clear moral connotations and serve for moral orientation, e.g., such as autonomy, freedom, safety, or wellbeing. To avoid confusion, we use the term "score" for variable value although this is not common in simulation modeling.

Info Box 1

Problem statement: IAT system to guide persons with dementia who have lost their orientation inside a care facility.

When residents wander and lose their orientation, it can be a challenge for everyone living and working in care facilities. Hence, an IAT system might help to actively guide patients through buildings. It might lock doors depending on the perceived cognitive state of the patients and on an assessment of safety and privacy. It might also call for human assistance (Landau and Werner, 2012; Ray et al., 2019; Bayat and Mihailidis, 2021; Lancioni et al., 2021).

In nursing homes, residents with limited orientation, for instance due to cognitive decline, often experience a reduced ability to manage their activities autonomously and safely. One obvious problem is getting lost in the nursing home on the way to a destination. Indeed, a substantial amount of care-giver attendance is required for providing guidance to disoriented residents. As a possible IAT for supporting autonomy and safety, one could imagine a "smart bracelet" that detects disorientation, provides orientation cues as appropriate, and calls a caregiver in case the problem persists. Such a system may increase autonomy of residents. It may decrease the amount of caregiver attendance to routine activities and thus free up caregiver resources for socially more salient activities. The benefit of such a system depends on its reliability in detecting an instance where help is required and on the effectiveness of its orientation cues. At the same time, such a system affects different stakeholder values: autonomy and safety for the resident, workload for caregivers, workforce efficiency, nursing quality, and safety regulations for the nursing home operator. It seems reasonable to assume that these values interact with each other. Some may reinforce each other; others may contradict each other. Even this situation can be considered complex, as we have seen in our own empirical research with affected persons. Patients and professionals might differ regarding the criteria of acceptance of such technological guidance (Buhr and Schweda, in prep.; Köhler et al., 2022). Values such as autonomy (of the person with dementia), privacy (of the user but also of other residents), safety, wellbeing, and costs (e.g., professional time) are balanced or prioritized differently across different stakeholders, as an empirical ethics study revealed (Buhr et al., under review; Welsch and Schicktanz, 2022). For example, we identified a group of persons with dementia (calling this type "individual self-determination") for whom disorientation technology should provide directions and guidance but should not inform third persons nor restrict the person's range of mobility. Another type of patient ("relational autonomy") accepts any technology that prevents them from wandering or getting disorientated with the goal of relieving caregivers' burden. They would also consent to having others be tracked or third persons alarmed in cases of disorientation. Thus, we see no empirical justification here for a "one value-profile fits all" approach. Further, the advantages or disadvantages of a technology that gives priority to different values must also be assessed with regard to "realistic" outcomes, potential side-effects, and how free they actually are to select between multiple values (e.g., in light of legal restrictions including liability issues).

protocols. The interesting aspect here is that the quantity "expected step count" is not a pre-defined parameter of the simulation but rather a quantity that arises from analysis of multiple simulation runs.

The following thought experiment will help in understanding the usefulness of simulation-based analyses. Assume you are building a new eldercare facility. In the planning process, it would clearly be of interest to see how a quantity such as the expected step count changes in response to modifications of the floor plan or other design elements. The above-described simulation could measure the specific benefit of such changes in terms of any given desirable outcome, such as reducing overall walking times, and can thus provide a quantitative rationale for choosing between corresponding design options in the real world. The use of simulation-based techniques is standard in analyzing the effects of design decisions and exploring what-if-scenarios in a wide range of application domains (such as economic decision making or the analysis of climate change).

Now, focusing on IAT again, we can use this technique also in the specific situation in which the quantities derived from simulations represent the degree of an IAT's compliance with or violation of a set of ethical values. Concerning the aspect of investigating ethics in IAT, this is an interesting shift in perspective from considering how to embed ethical values into IATs to considering how an IAT's actions reflect such values in practice. We call the computation of numerical scores that represent compliance with a set of ethical values ethical compliance quantification (ECQ, see below). As we will discuss below, such an approach is not only interesting because it might provide relevant information for the ethical assessment of an IAT. It also requires all assumptions to be made explicit in order to render them computable and is therefore also an interesting mechanism to discuss and investigate the design and effect of value structures in specific use cases.

Figure 1 identifies the central components involved in this simulation-based approach and their interplay. The objective of the simulation system component is to simulate the interaction between human stakeholders and IAT in a given environment (such as the interaction between residents, nurses, and a smart-watch-based orientation IAT during nursing home routines, as outlined in our use case below). This means that it is necessary to provide computational rules and stochastic processes that define–in non-deterministic fashion–stakeholder[9] behavior (stakeholder model)[10]. In the process of this definition, it might become necessary to quantify the mental states in stakeholder models–such as the state of a patient's sense of orientation, their likelihood of losing their way altogether, or preferences for certain forms of interaction.[11] It is also necessary to describe the IAT behavior (which is usually comparatively easy because the IAT implementation itself provides the blueprint), as well as the IATs sensor characteristics that define how well it is able to observe the current situation (IAT model). Depending on the application setting, the IAT's sensor reliability may be crucial for being able to make right decisions. A central component in defining IAT behavior is the IAT's "policy," i.e., a set of rules that define how the IAT will chose what assistive action in which situation; it represents the IATs decision-making component. From the viewpoint of the IAT designer, the objective is to define a policy that–within the technical limits of the system environment–achieves optimal results. Such "optimal" results should also be ethically compliant. How to reach this will be discussed in the next step.

9 Stakeholder means here all the people whose concerns should be considered in system design.

10 Any kind of knowledge on stakeholder behavior is a reasonable source for model building, empirically or theoretically based. This information then must be transformed into an algorithmic structure a machine can execute. This is the stakeholder model.

11 It should be noted that there exist several cognitive architectures–such as ACT-R, Psi, or SOAR–that provide building blocks for creating a computational model of mental states (Though, the use of such an architecture is not a necessity for setting up a simulation.)

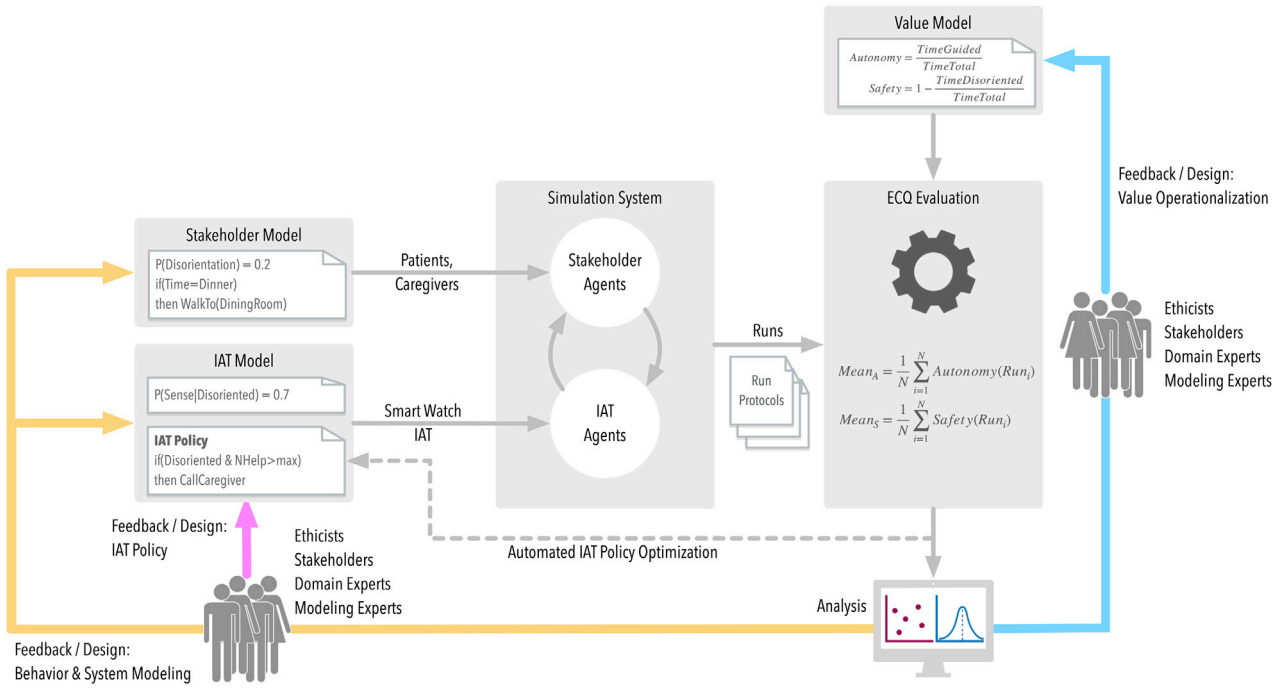


FIGURE 1
Concept of an AI-Assisted Simulation includes Ethical Compliance Quantification (ECQ) and two feedback loops (blue and yellow).

3.2.2 Ethical compliance quantification (ECQ)

The objective of the ECQ evaluation is to provide a quantitative statement on IAT adherence to a score model, based on a set of run protocols generated by the simulation. From an ethicist's viewpoint, the value model is the crucial component, as it provides the translation between the sequence of events in the simulation runs, especially the IAT actions in specific situations, and their ethical assessment. Let us illustrate what defining in such a value model means. Let us consider the care facility floor planning as discussed above. The first step in defining a value model is to identify its values or, better, its value dimensions. Since we will provide scores (numbers) for values, a set of scores–one for each value–defines a point in a space where each dimension corresponds to a value. A very simple ethical value system might ask for "efficiency" and "fairness." The next step in defining a value model is to provide formulae that instruct how to compute a quantitative score as data for the value dimensions of "efficiency" and "fairness" from a simulation run. In our thought experiment world, where stakeholders move between locations, efficiency might for instance be given by the ratio of straight-line distance to distance travelled, while fairness might be given by the quotient of the efficiency scores for different stakeholders (the value "1" representing optimal fairness when all stakeholders experience equal "efficiency"). Then an ECQ setup can be used to compare different floor plans with respect to their rating on the different value dimensions. Even this very simple thought experiment illustrates the core challenge in defining a value model: providing a model that adequately reflects how values are connected to the real world. For instance, consider the–rather trivial–example definition of fairness. One might rightfully wonder, if it is really fair to compare just efficiency and ignore the physical fitness of stakeholders (e.g., the fitter one is, the longer one can walk). So, stakeholders might rightfully call for a correction factor for the fairness computation that reflects physical fitness.

This simple example illustrates the multilateral nature and the value-sensitive design process required for defining a value model, because it makes value judgements explicit. And by this, it exposes the degrees of freedom that are available in designing the mapping from event sequences to value ratings. Note that simulation-based ECQ also allows the assignment numbers to qualitative value statements: for instance, by counting how often a certain qualitative requirement is observed or violated in a number of simulation runs.

Note that the ECQ-concept provides something impossible in the real world: to evaluate the quantification across different design alternatives for all of the involved models. By varying the IAT policy, it becomes possible to assess the impact of different design alternatives on the compliance to values (policy feedback), possibly with the objective of arriving at an optimal IAT policy. Varying the value model allows assessing the plausibility of the resulting value quantification and thus the plausibility of the value model itself (value operationalization feedback). Finally, by varying stakeholder and IAT model, the sensitivity of the ECQ results to the ecological validity of the simulation model can be assessed.

3.2.3 An example

SimDem (Shaukat et al., 2021) is a simulation system we developed to analyze a smart-watch based IAT in a nursing home for dementia patients. A "smart watch" supports residents by detecting deviations from routes and then prompts the wearer

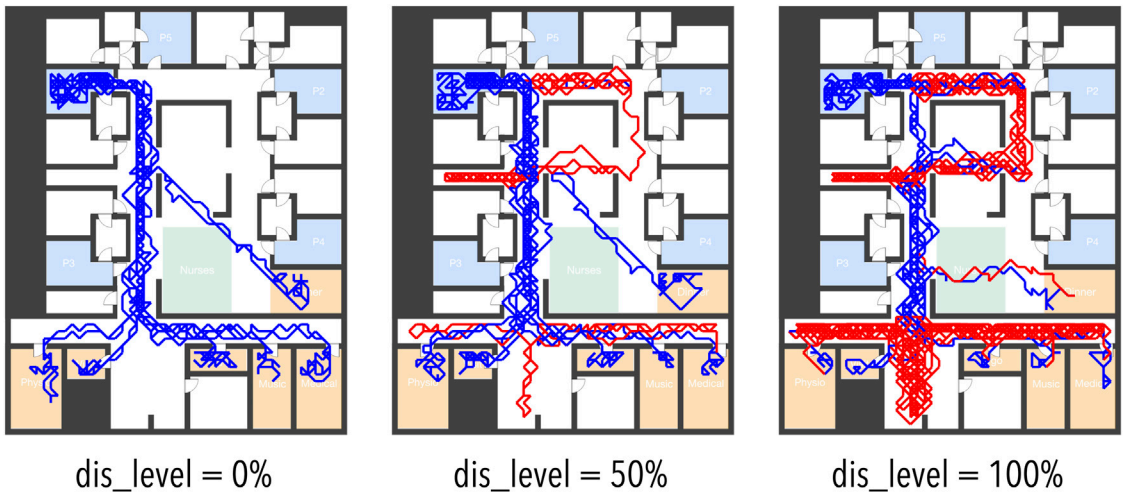


FIGURE 2
Visualization of Trajectories for Simulated Patient Agents with Various Levels of Disorientation Probability (dis_level). Blue = patient state oriented; Red = patient state disoriented.

about which direction to go to reach the destination. One very basic design issue now is the question of how many guidance interventions should trigger the assumption in the system that the wearer is permanently disorientated and thus alert a caregiver. On the basis of previous expert interviews, our own reasoning, and literature research, one might label such values as "safety" and "fairness."

When performing ECQ, the very first step is creating the simulation model. In this case, it is a 2D virtual nursing home (the floor plan based on a real nursing home), where the way-finding behavior for the patient simulation (see Figure 2) has been assessed via observations of real subjects[12].

The simulated IAT–the simulated smart watch–has a certain probability for detecting disorientation and there is a certain probability that a smart watch intervention will help the supported person regain orientation (both these probabilities are design parameters of the simulation model). Based on this setup, it is then possible to perform multiple simulation runs and analyse the quantitative effect of different assistive strategies on values of interest. In figure three, we show the aggregated results from 1809 runs, using different value models. We use this figure to discuss the crucial aspect of value model definition. Concerning the value model, it is first of interest to operationalize "safety." It turns out that there are multiple ways to do this. One might consider the relative amount of time in disorientation as "unsafe" time. This approach produces–as a function of the intervention policy–the reddish colored box plots in Figure 3, labelled "Safety (Original)"

Figure 3 shows that this operationalization is not plausible. The plot shows that, using this operationalization, the resulting score for the strategy of immediately calling a nurse (Nhelp = 0) indicates a higher non-compliance (i.e., a longer time in unsafe state) than the score for the strategy of waiting for five failed smart watch interventions (Nhelp = 5). But, obviously, the more failed interventions we wait for, the longer the disoriented patient will wander unguided. Therefore, this operationalization clearly results in score values that disagree with common sense. The implausibility of this value operationalization design is obvious once the plot provides a visualization of the outcome: as soon as a nurse is accompanying a patient, the situation should be considered as safe by the value operationalization, independent of the patient's disorientation state. Note how the ECQ approach allows discovery of such mistakes in value operationalization through visualizing the value scores across different strategies, as shown in this example.

Providing a more plausible value operationalization now is straightforward: as suggested above, we only consider the time during which a patient is disoriented while not guided by a nurse as "unsafe" time. Using this improved operationalization of "safety" we now see that indeed, immediately calling a nurse is safer than waiting for multiple interventions (see Figure 3, purple box plot, labelled as "Safety (Refinef)"). Note that this plot also reveals that having no IAT at all ("Nurse Only") is the least safe strategy (aside from leaving the patient completely unattended, "No Help") Note that this plot also reveals that having no IAT at all ("Nurse Only") is the least safe strategy (aside from leaving the patient completely unattended, "No Help"). The reason for this is that without a smart watch detecting disorientation, nurses have to actively discover disoriented patients. In this simulation setting, the smart watch therefore always increases safety.

Using the improved operationalization of "Safety," it is now interesting to see how this value is affected by IAT policy in comparison to the "Fairness" value, which reflects the relative amount of time available for forms of caregiving other than route guidance. The rationale behind this is that the more time a caregiver is occupied with route guidance, the less time is available for other, possibly more important tasks, such as social interactions. This

12 The probability of selecting a wrong turn is based on data from a study on indoor wayfinding of the University Medicine Rostock. Participants were 8 subjects diagnosed with light to medium dementia (Male/Female = 4/4; Age M = 73.4, SD = 6.3; MMSE M = 22.5, SD = 3.4). Study protocol approved by ethics commitee of University Rostock, approval number A2012-0083.

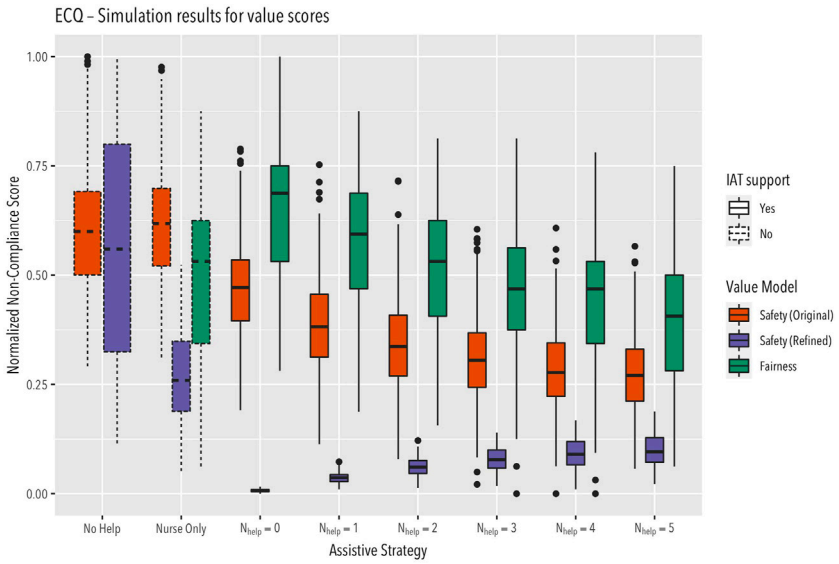


FIGURE 3
Value scores computed different value models across 1809 simulation runs (see text for details).

information is provided by comparing the box plots for the values "Safety (Refined)" and "Fairness" in Figure 3. We see that there is an obvious value conflict, as decreasing Fairness violations (by waiting longer before calling a nurse) leads to increasing Safety violations. Figure 3 also shows that the safety gain provided by the smart watch IAT in general comes at the price of increased workload, as caregivers are now proactively called for guidance as soon as the IAT gives up on interventions.

Obviously, given the simplicity of the simulation model and its operationalization, this example is of limited significance for the practical design of an orientation-support IAT based on smart watch devices. However, the example does clearly show the potential of ECQ as a means to provide insight into the ethical assessment of IAT, which is the point of interest here. We see that ECQ allows visualization of value trade-offs, and the potential non-linear dependency of score functions on policies. It also shows that ECQ helps in operationalizing values with respect to "real world events" in such a way that the operationalization overlaps with commonly accepted moral precepts. It also shows that human intuition is not guaranteed to provide a plausible operationalization (as illustrated by the first version of the safety operationalization).

4 Discussion

Simulation-based ECQ is a method for exploring the ethics design space, for developing "ethics awareness" in designers, and for informing ethicists about not only outcomes of different scenarios, but how different variables influence the process. Furthermore, as in our case, it allows the simulation or anticipation of complex ethical trade-offs, not only purely hypothetically or very generally (as thought experiments), but as visualized trade-offs regarding human-AI and human-human interaction that cannot be explained or rationalized by the persons involved. In the following discussion, we want to focus on three main challenges.

1) How qualitative values can (or must be) operationalized for such computational simulations and what this requires.
2) In which contexts and for what purposes the advantages of such an AI-assisted simulation outweigh their disadvantages and limits.
3) Why AI-assisted ethics simulations can be compared to thought experiments but provide innovative epistemic dimensions for ethical reasoning.

First, the methodology discussed in this paper is not about a specific value–such as autonomy or safety, but on how to improve the process of ethical reflection for IAT development by considering the diversity of values. Aliman and Kester 2019 argue in favor of a consequentialist approach that predicts the overall utility of a future outcome for a given population. While they thus make a very general argument for AI-assisted simulation with regard to the value of utility, our methodology rather proposes a strategy to understand the impact of an IAT regarding different moral values that can be operationalized. In this sense, our proposal is agnostic regarding the specific values considered during design, but it is not agnostic with regard to the requirement of a participatory and pluralistic approach. Hereby, our methodology is aware of the central challenge of value operationalization. Unless a value is operationalized, it cannot be analyzed by ECQ. While this may be seen as a drawback of the method, we see it rather as an advantage. ECQ poses a *challenge* to value experts to operationalize their value concepts, because ECQ provides the *opportunity* to make use of such an operationalization. A claim that a value cannot be quantified can now be challenged by providing an operationalization, counter-challenging the opponent to show where it violates the value system. In a similar vein, one of the core benefits of using ECQ will be to expose situations where an operationalization indeed cannot be found–or rather cannot be agreed upon. By forcing stakeholders to give an explicit semantics to their value concepts, ECQ exposes conflicts that are indeed fully independent of the question of

"machine ethics," but rather are caused by our own inconsistent or ambiguous opinions concerning moral behavior (Note that behavioral economics has shown that even single persons may make contradictory assessments of situations, depending on whether a situation is experienced or remembered, see also Aliman and Kester, 2019 for this problem.)

We think that the 'values' relevant for machine ethics are far from being sufficiently defined and hinge on theoretical backgrounds, often of conflicting philosophical stances (see also Volkman and Gabriels, 2023). Consider the confusion between "autonomy" as moral self-determination, reflective self-governance or freedom of choice and "automation" understood as "absence of human control", but often seen as model for "autonomous decision of a machine". For example, the popular "levels of automation" model discussed in Sheridan (1980) provides a quantitative theory of "autonomous decisions". This model identifies eight automation levels, level 1 being no automation, level 8 removing any human involvement. While in a simulation, the levels of automation can be tested and quantified, their moral assessment–which level of interaction are better or morally more acceptable–cannot be quantified or answered. ECQ provides a methodology to experiment with different operationalizations of values to analyze which of them coincide with intuitive ethical judgement. The stakeholder feedback-loop external to the simulation furthermore permits a critical reflection on the consequences of the selection of values and operationalizations and allows for radical revision.

Second, another objection concerning the simulation approach might be: Why not simply ask the user for the level of support she would like? This objection is also highlighted by Aliman and Kester (2019) who discuss the possibility of predicting the utility of an AI-based device by the potential users. As they aptly remark, "predicted utility is subject to diverse considerable cognitive biases and often crucially differs from instant utility." (ibid.: 28) It should be emphasized that this applies even more to persons with dementia whose value preferences can change or become unpredictable with progressing dementia. With respect to people with dementia, another obvious reply is that some of them will not be able to express a well-considered preference. But, on a more general level, this is an aspect that holds for all stakeholders. It is in general difficult to assess the consequences of a rather abstract decision ("How many times should the smart watch provide navigation hints before calling a caregiver?") with respect to the impact on the personal experience. Moreover, empirical or participatory approaches that, for example, interview stakeholders also have limitations. Sample sizes are often small, the situations that can be morally assessed are anecdotal, the expectations of future technologies are biased by the experience with current technologies, the experiences and values might be biased by the individual perspectives, they provide only limited and biased reflections of reality, more complex technological features that are opaque to the individual are not considered, and the information gained is static and again requires thought experiments to consider novel what-if-scenarios. All these restrictions limit generalizations of ethical design. This is because sufficient experience to judge the decision impact for a novel technology does not exist in the rule. The simulation approach allows stakeholders (e.g., ethicists, engineers, healthcare providers, patient advocates) to see what her decisions would mean in "practice." This makes it particularly helpful in the context of new technologies that have not been implemented yet. Of course, we do not suggest *not* to ask the stakeholders, but rather to provide sufficient information before asking and hence, to have a more informed and reflected discussion about potential outcomes and ethical trade-offs. In this sense, the simulation approach does not aim to surpass or replace but to complement participatory approaches.

Third, we started above with the role of thought experiments and their importance as a tool for reflecting on new technologies (or new ideas in general). However, thought experiments have their limitations: They are fictional, and often neglect physical, biological, or social conditions since they are usually primarily aimed to test for logical implications and conceptual premises and therefore tend to operationalize critical variables categorically even when they are continuous. This can become problematic when ethical implications for single agents are analyzed as categorization can introduce errors and contradictory results when only few cases and few variables are considered. An ethical technology assessment, on the other hand, is understood as the exploration and evaluation of more or less likely (or plausible) future scenarios. Even if possible future developments are anticipated under uncertainty or ignorance, relevant and reliable physical, biological, and social knowledge must be considered. The simulation approach can support this anticipation beyond a thought experiment by systematically running through a whole range of possible baseline conditions and their respective outcomes. It can introduce statistical variations and be re-run multiple times to produce a population of outcomes on the variables of interest. In principle, well developed methods for model checking and model evaluation can be applied to better understand functional relations between components and variables in the simulations. However, similar to thought experiments, simulations require that variables are adequately operationalized. In addition, other important variables of the virtual world in which the simulation unfolds need to be sufficiently realistically implemented. In our example, physical variables reflecting the state of the simulation, like space and time, must be part of the simulations and appropriately implemented. Even such "technical" variables can have an influence on the outcome of the simulation and can produce biases if implemented inappropriately. For example, modelling a discrete quantity with a continuous number can lead to meaningless results of an "in between state" on a quantity that is categorical of can only take discrete values. The final ethical evaluation–i.e., whether respective developments or at least individual consequences are considered desirable, undesirable, or even unacceptable from a moral point of view–must, however, be made through human reflection by the observers of such simulation. Therefore, the use of AI in this context does not mean a replacement of human ethical reflection through ethical machines, or that machines can make moral decisions. Instead, our proposed model can be subsumed as a form of "human-centered AI" (Shneiderman, 2021) in the field of ethics which strives to support humans in reasoning about complex systems by means of computational simulations (again, just like in the case of the world climate). In the ethical context, also usually more than one stakeholder is involved, and the values of different stakeholders can be in conflict (e.g., of caretakers and patients). In that case, a compromise needs to be found. Empirically analyzing the effects of different compromises for IAT policy in practice is obviously not possible and would also be morally problematic. Instead, a simulation as ECQ can systematically explore different alternatives without intervening with the actual practice. In this sense, ethical reasoning can directly benefit from the simulation system and our approach could be considered as a form of

"intelligent augmentation" of ethical reasoning. Indeed, both have in common that they stress two aspects (Shneiderman, 2021, 9). First, their design method builds on user experience, stakeholder engagement, and iterative refinement. Second, they are designed as a "supertool" (Shneiderman) to amplify, augment, and empower human performance, but emphasize human control. An additional challenge to such simulation as we have proposed is the significance of space and time for the results. Therefore, we emphasize that we do not advocate to use simulation as full replacement for real world experiments. Rather, our objective is to help stakeholders do develop insight into the conceptual validity of their models of how to quantify the effect of assistive system actions on ethical values. In our example, we have modeled these two dimensions as a continuum, i.e., moral evaluations and of the simulated users with dementia were stable over time. This is not to be expected for the reality: Real users' values and preferences will change over time due to the progress of dementia, to the experiences made with IAT or other endogene and exogene factors. Hence, an AI-based simulation cannot eliminate the need for an ex post assessment via experiments and corresponding individualization. Nevertheless, it contributes to an ex ante ethical alignment of the new technologies for vulnerable groups, e.g., persons with dementia.

5 Conclusion and outlook

AI-assisted simulations can address shortcomings of the current gold standard of empirically informed ethical reasoning, as well as of traditional approaches such as thought experiments and forecasting methods. They could help in the exploration of numerous complex what-if scenarios with great flexibility and provide objective observations that can be visualized and analyzed processually. The process of visualization seems especially relevant as it helps to manifest trade-offs and observations.

In particular, our contribution considers how empirical data about the scope of stakeholders' value preferences and potential ways of behavior can inform a "supertool" to permute the range of ethically relevant baseline parameters and thus simulate different possible outcomes. In this vein, ethically motivated empirical research and AI-assisted simulation strategies are combined and complement each other. In this sense, what we propose here is neither traditional ethics of technology nor machine ethics, but AI-assisted ethics as a new, innovative methodology for empirically informed ethical reflection. However, as an interdisciplinary working group, we also realized that time for collaborative learning is needed to achieve a productive combination of theoretical and methodological perspectives. Of course, in our case, the object of such ethical reflection is also AI-technology. While this does not necessarily have to be the only conceivable use case for AI-assisted ethics, the approach proves to be particularly suited to this still young field of technology development with its comparatively low degree of practical implementation and actual empirical experience. In future the AI model simulations, the inner states (e.g., values, emotions) and behavior of the simulated agents could be included, tracked, and related via AI methods to the interventions (e.g., Francilett et al., 2020). Such a prediction model could produce information about the combined effects of the intervention on various inner states and make this information available for ethical analysis. The results of simulations with such models could provide data similar to results generated by empirical interviews for use in ethical reasoning. Agents may be simulated using symbolic or sub-symbolic AI techniques. Both have a tradition in cognitive psychology and gaming. In the end, the approach can lead to better-informed ethical reasoning by providing data on how humans are affected by a AI-based system and may help to identify critical factors that lead to problematic situations and support the investigation of ways to mitigate them. It can do this in complex, realistic situations with multiple actors and technical components interacting with each other. This opens new perspectives for the systematic ethical reflection of technological futures in the middle ground between dystopia and utopia.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

All authors have contributed to an early stage of conceptual design of this article during a series of internal workshops. SiS and TK have drafted and structured the first manuscript version, together with JW who drafted main parts of subchapter 2. MS has helped in drafting subchapter 2, too, and made substantial contributions for the discussion part. JR has co-edited subchapter 1, 2, and 3. AH has helped to design subchapter 4 and also to conceptualize Figure 1. All authors contributed to the article and approved the submitted version.

Funding

The current study was carried out within the context of the ongoing BMBF research project "Ethical and Social Issues of Co-Intelligent Monitoring and Assistive Technologies in Dementia Care (EIDEC)." This work is funded by the Federal Ministry of Education and Research (BMBF), funding number: 01GP1901 (January 2020–June 2023). SiS received a Fellowship from the Hanse Institute of Advance Studies, Delmenhorst, Germany which allowed her to concentrate on this interdisciplinary paper.

Acknowledgments

We are grateful to the experts who were willing to share their opinions and time with us. Especially, we thank Prof. Stefan Teipel, Rostock, for his inspiring remarks on the manuscript, Scott Gissendanner for help in language editing and reviewers for their critical-constructive remarks.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abojabal, H., Welsch, J., and Schickanz, S. (under review). Cross-cultural perspectives on intelligent assistive technology in dementia care: Comparing Israeli and German experts' attitudes

Aliman, N., and Kester, L. J. H. M. (2019). *Transformative AI governance and AI-empowered ethical enhancement through preemptive simulations*. Delphi 1/2019. doi:10.21552/DELPHI/2019/1/6

Alkadri, J., and Jutai, J. (2016). Cognitive impairment and assistive devices: Outcomes and adverse effects. *J. Rehabil. Assist. Technol. Eng.* 3, 2055668316668146. doi:10.1177/2055668316668146

Amann, J., Blasimme, A., Vayena, E., Frey, D., and Madai, V. I.Precise4Q consortium (2020). Explainability for artificial intelligence in healthcare. A multidisciplinary perspective. *BMC Med. Infrom. Decis. Mak.* 20 (1), 310. doi:10.1186/s12911-020-01332-6

Anderson, M., and Anderson, S. L. (2007). Machine ethics. Creating an ethical intelligent agent. *Ai Mag.* 28 (4), 15–26. doi:10.1609/aimag.v28i4.2065

Anderson, M., Anderson, S. L., and Armen, C. (2004). *Towards machine ethics*. Conference paper. Available at: https://www.aaai.org/Papers/Workshops/2004/WS-04-02/WS04-02-008.pdf (Accessed August 31, 2022).

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., et al. (2018). The moral machine experiment. *Nature* 563, 59–64. doi:10.1038/s41586-018-0637-6

Bayat, S., and Mihailidis, A. (2021). Outdoor life in dementia. How predictable are people with dementia in their mobility? *Alzheimers. Dement. (Amst.).* 13 (1), e12187. doi:10.1002/dad2.12187

Bicher, M., Rippinger, C., Urach, Ch., Brunmeir, D., Siebert, U., and Popper, N. (2021). Evaluation of contact-tracing policies against the spread of SARS-CoV-2 in Austria: An agent-based simulation. *Med. Decis. Mak.* 41 (8), 1017–1032. doi:10.1177/0272989X211013306

Border, S. P., and Sarder, P. (2022). From what to Why, the growing need for a focus shift toward explainability of AI in digital pathology. *Front. Physiol.* 12, 821217. doi:10.3389/fphys.2021.821217

Brundage, M. (2014). Limitations and risks of machine ethics. *J. Exp. Theor. Artif. Intell.* 26 (3), 355–372. doi:10.1080/0952813X.2014.895108

Buhr, E., and Schweda, M. (in prep). *Moral issues of assistive technologies in dementia care: An ethical analysis of views and attitudes of affected people in Germany*.

Buhr, E., Welsch, J., and Shaukat, M. S. (under review). *Value preference profiles. A source for ethics by design in technology-assisted dementia care*.

Chandrasekharan, A., Nersessian, N. J., and Subramanian, V. (2013). "Computational modeling: Is this the end of thought experimenting in science?," in *Thought experiments in philosophy, science and the arts*. Editors J. Brown, M. Frappier, and L. Meynell (London: Routledge), 239–260.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Sci. Eng. Ethics.* 26 (4), 2051–2068. doi:10.1007/s11948-019-00146-8

Currie, G., Hawk, K. E., and Rohren, E. M. (2020). Ethical principles for the application of artificial intelligence (AI) in nuclear medicine. *Eur. J. Nuc. Med. Mol. Imaging.* 47, 748–752. doi:10.1007/s00259-020-04678-1

De Freitas, J., Cenis, A., Smith, B. W., and Frazzolio, E. (2021). From driverless dilemmas to more practical commonsense tests for automated vehicles. *Proc. Natl. Acad. Sci. U S A* 118 (11), e2010202118. doi:10.1073/pnas.2010202118

Di Paolo, E. A., Noble, J., and Bullock, S. (2000). *Simulation models as opaque thought experiments*. Cambridge, MA: MIT Press, 497–506.

Endter, C. (2021). *Assistiert Altern. Die Entwicklung digitaler Technologien für und mit älteren Menschen*. Wiesbaden: Springer VS.

Francillette, Y., Boucher, E., Bier, N., Lussier, M., Bouchard, K., Belchior, P., et al. (2020). Modeling the behavior of persons with mild cognitive impairment or Alzheimer's for intelligent environment simulation. *User Model User-Adapt. Interact.* 30 (5), 895–947. doi:10.1007/s11257-020-09266-4

Friedmann, B., and Kahn, P. H., Jr. (2007). "Human values, ethics, and design," in *The human-computer interaction handbook*. Editors J. A. Jacko and A. Sears (Boca Raton, FL: CRC Press), 1267–1292.

Goodall, N. J. (2019). More than trolleys: Plausible, ethically ambiguous scenarios likely to Be encountered by automated vehicles. *Transf. Interdiscipl. J. Mobili. Stud.* 9 (2), 45–58. doi:10.3167/TRANS.2019.090204

High Level Expert Group on Artificial Intelligence (2019). *Ecthics guidelines for trustworthy AI*. Available at: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf (Accessed August 31, 2022).

Hofmann, B. (2013). Ethical challenges with welfare technology. A review of the literature. *Sci. Eng. Ethics.* 19, 389–406. doi:10.1007/s11948-011-9348-1

Ienca, M., Jotterand, F., Elger, B., Caon, M., Pappagallo, A., Kressig, R. W., et al. (2017). Intelligent assistive technology for Alzheimer's disease and other dementias. A systematic review. *J. Alzheimer. Dis.* 56 (4), 1301–1340. doi:10.3233/jad-161037

Ienca, M., Wangmo, T., Jotterand, F., Kressig, R. W., and Elger, B. (2018). Ethical design of intelligent assistive technologies for dementia: A descriptive review. *A Descr. Rev. Sci. Eng. Ethics.* 24, 1035–1055. doi:10.1007/s11948-017-9976-1

Köhler, S., Görß, D., Kowe, A., and Teipel, S. (2022). Matching values to technology: A value sensitive design approach to identify values and use cases of an assistive system for people with dementia in institutional care. *Ethics. Inf. Technol.* 24, 27. doi:10.1007/s10676-022-09656-9

Korn, G. A., and Wait, J. V. (1978) *Digital continous-systems simulation*, Prentice-Hall, Englewood Cliffs, N.J.

Kunze, C., and König, P. (2017). "Systematisierung technischer Unterstützungssysteme in den Bereichen Pflege, Teilhabeunterstützung und aktives Leben im Alter," in *Umgebungsunterstützes leben. Beiträge zum usability day XV*. Editors I. Hämmerle and G. Kempter (Lengerich: Pabst), 15–22.

Lancioni, G. E., Desideri, L., Singh, N., O'Reilly, M. F., Sigafoos, J., De Caro, M. F., et al. (2021). Use of technology to sustain mobility in older people with cognitive impairment and dementia. A scoping review. *Disabil. Rehabil. Assist. Technol.* 222, 1–15. doi:10.1080/17483107.2021.1900935

Landau, R., and Werner, S. (2012). Ethical aspects of using GPS for tracking people with dementia. Recommendations for practice. *Int. Psychogeriatr.* 24 (3), 358–366. doi:10.1017/S1041610211001888

Lara, F., and Deckers, J. (2020). Artificial intelligence as a socratic assistant for moral enhancement. *Neuroethics* 13, 275–287. doi:10.1007/s12152-019-09401-y

Löbe, C., and AboJabal, H. (2022). The role of intelligent assistive technology for empowering people with dementia. A scoping review. *Arch. Geronto.* Epub. doi:10.1016/j.archger.2022.104699

Manzescheke, A., Weber, K., Rother, E., and Fangerau, H. (2013). *Ethische Fragen im Bereich Altersgerechter Assistenzsysteme. Ergebnisse der Studie*. VDI/VDE Innovation + Technik.

Markus, A. F., Kors, J. A., and Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care. A comprehensive survey of the terminology, design choices, and evaluation strategies. *J. Biomed. Inf.* 113, 103655. doi:10.1016/j.jbi.2020.103655

McLennan, S., Fiske, A., Tigard, D., Müller, R., Haddadin, S., and Buyx, A. (2022). Embedded ethics. A proposal for integrating ethics into the development of medical AI. *BMC Med. Ethics.* 23, 6. doi:10.1186/s12910-022-00746-3

Mertz, M., Inthorn, J., Renz, G., Rothenberger, L. G., Salloch, S., Schildmann, J., et al. (2014). Research across the disciplines: A road map for quality criteria in empirical ethics research. *BMC Med. Ethics.* 15, 17. doi:10.1186/1472-6939-15-17

Mihailidis, A., Boger, J. N., Craig, T., and Hoey, J. (2008). The COACH prompting system to assist older adults with dementia through handwashing. An efficacy study. *BMC Geriatr.* 8, 28. doi:10.1186/1471-2318-8-28

Misselhorn, C. (2021). Artificial systems with moral capacities? A research design and its implementation in a geriatric care system. *Artif. Intell.* 278, 103179. doi:10.1016/j.artint.2019.103179

Mollison, D. (1991). Dependence of epidemic and population velocities on basic parameters. *Math. Biosci.* 107 (2), 255–287. doi:10.1016/0025-5564(91)90009-8

Nallur, V. (2020). Landscape of machine implemented ethics. *Sci. Eng. Ethics.* 26, 2381–2399. doi:10.1007/s11948-020-00236-y

Novitzky, P., Smeaton, A. F., Chen, C., Irving, K., Jacquemard, T., O'Brolcháin, F., et al. (2015). A review of contemporary work on the ethics of ambient assisted living technologies for people with dementia. *Sci. Eng. Ethics.* 21, 707–765. doi:10.1007/s11948-014-9552-x

Parke, E. C. (2014). Experiments, simulations, and epistemic privilege. *Philos. Sci.* 81 (4), 516–536. doi:10.1086/677956

Peck, St. L. (2004). Simulation as experiment: A philosophical reassessment for biological modeling. *Trends Ecol. Evol.* 19 (10), 530–534. doi:10.1016/j.tree.2004.07.019

Ray, P. P., Dash, D., and De, D. (2019). A systematic review and implementation of IoT-based pervasive sensor-enabled tracking system for dementia patients. *J. Med. Syst.* 43, 287. doi:10.1007/s10916-019-1417-z

Schicktanz, S., and Schweda, M. (2021). Aging 4.0? Rethinking the ethical framing of technology-assisted eldercare. *Hist. Philos. Life Sci.* 43, 93. doi:10.1007/s40656-021-00447-x

Schicktanz, S., Schweda, M., and Wynne, B. (2012). The ethics of public understanding of ethics -Why and how bioethics expertise should include public and patients voices. *Med. Health Care Philos.* 15 (2), 129–139. doi:10.1007/s11019-011-9321-4

Schweda, M., Kirste, T., Hein, A., Teipel, A., and Schicktanz, S. (2019). The emergence of co-intelligent monitoring and assistive technologies in dementia care - an outline of technological trends and ethical aspects. *Bioethica Forum* 12 (1/2), 29–37.

Shaukat, M. S., Hiller, B. J., Bader, S., and Kirste, T. (2021). SimDem A multi-agent simulation environment to model persons with dementia and their assistance. *arXiv*. doi:10.48550/arXiv.2107.05346

Sheridan, T. B. (1980). Computer control and human alienation. *Technol. Rev.* 83 (1), 60–73.

Shneiderman, B. (2021). *Human centered AI*. Oxford: University Press.

Spiekermann, S. (2016). *Ethical IT innovation. A value-based system design approach*. Boca Raton, FL: CRC Press.

Spitzenverband der Gesetzlichen Krankenversicherungen in Deutschland (2019). *Digitalisierung und Pflegebedürftigkeit. Nutzen und Potenziale von Assistenztechnologien*. Hürth.

Umbrello, S., and van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI Ethics* 1, 283–296. doi:10.1007/s43681-021-00038-3

van Wynsberghe, A. (2013). Designing robots for care. Care centered value-sensitive design. *Sci. Eng. Ethics.* 19 (2), 407–433. doi:10.1007/s11948-011-9343-6

Vandemeulebroucke, T., Dierckx de Casterlé, B., and Gastmans, C. (2018). The use of care robots in aged care. A systematic review of argument-based ethics literature. *Arch. Gerontol. Geriatr.* 74, 15–25. doi:10.1016/j.archger.2017.08.014

Volkman, R., and Gabriels, K., (2023): AI moral enhancement. Upgrading the socio-technical system of moral engagement. *Sci. Eng. Ethics* 29, 11. doi:10.1007/s11948-023-00428-2

Wallach, W., Franklin, S., and Allen, C. (2009). A conceptual and computational model of moral decision making in human and artificial agents. *Top. Cogn. Sci.* 2, 454–485. doi:10.1111/j.1756-8765.2010.01095.x

Walsh, A. (2011). A moderate defence of the use of thought experiments in applied ethics. *Prac* 14, 467–481. doi:10.1007/s10677-010-9254-7

Walton, D. (2016). Some artificial intelligence tools for Argument evaluation: An introduction. *Argumentation* 30, 317–340. doi:10.1007/s10503-015-9387-x

Weber, K. (2021). "Altersgerechte Assistenzsysteme: Ein Überblick," in *Gute Technik für ein gutes Leben im Alter? Akzeptanz, Chancen und Herausforderungen altersgerechter Assistenzsysteme*. Editors K. Weber, S. Haug, D. Frommeld, and U. Scorna (Bielefeld: transcript Verlag), 27–62. doi:10.14361/9783839454695-002

Welsch, J., and Schicktanz, S. (2022). "Developing value preference profiles as tool for value-oriented technology design for living with dementia? Practical and methodological considerations," in Presentation given at the World Congress of Bioethics 2022 (Switzerland: University of Basel). Available at: https://organizers-congress.org/frontend/index.php#.

Welsch, J. (2022a). "Empowerment and Technology. An ethical-empirical exploration of technology-assisted dementia care," in Poster presentation, International Symposium The Future of Assistive Technologies in Dementia Care (Delmenhorst).

Welsch, J. (2022b). "Digital-assisted care in the digital desert? Experts' perspectives on the preconditions of modern monitoring and assistive technology in dementia care," in Presentation given at the Jahrestagung der Akademie für Ethik in der Medizin e.V.

Winsberg, E. (2022). "Computer simulations in science," in *The stanford encyclopedia of philosophy (winter 2022 edition)*. Editors E. N. Zalta and U. Nodelman URL: https://plato.stanford.edu/archives/win2022/entries/simulations-science/.

WHO - World Health Organization (2018). *Assistive technology*. Key facts. Available at: https://www.who.int/news-room/fact-sheets/detail/assistive-technology (Accessed August 31, 2022).

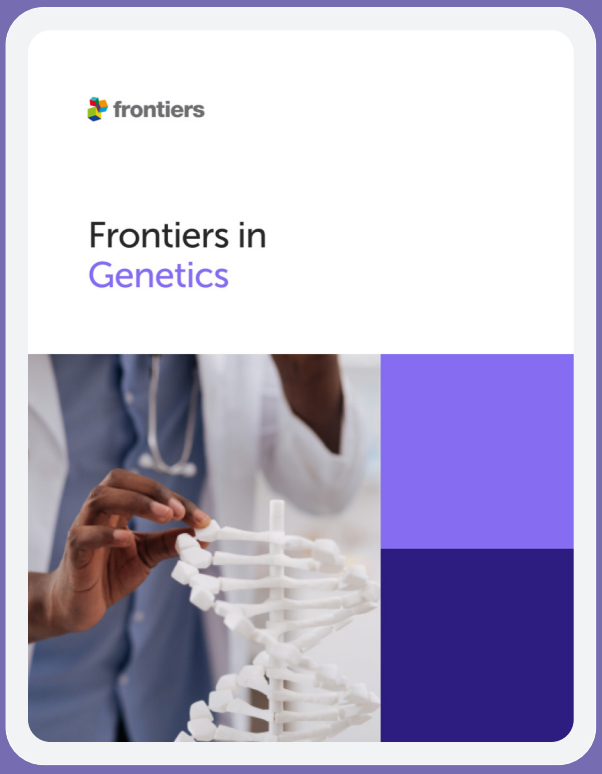
Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact