Original Paper

Finding Consensus on Trust in AI in Health Care: Recommendations From a Panel of International Experts

Georg Starke[1,2], MD, PhD; Felix Gille[3,4], PhD; Alberto Termine[1,2,5], PhD; Yves Saint James Aquino[6], MD, PhD; Ricardo Chavarriaga[7], PhD; Andrea Ferrario[8], PhD; Janna Hastings[4,9], PhD; Karin Jongsma[10], PhD; Philipp Kellmeyer[11,12], MD; Bogdan Kulynych[13], PhD; Emily Postan[14], PhD; Elise Racine[15,16], MPA, MSc; Derya Sahin[17], MD; Paulina Tomaszewska[18], MSc; Karina Vold[19,20], PhD; Jamie Webb[21], PhD; Alessandro Facchini[5], PhD; Marcello Ienca[1,2], PhD

[1]Institute for History and Ethics of Medicine, Technical University of Munich, Munich, Germany

[2]College of Humanities, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[3]Digital Society Initiative, University of Zurich, Zurich, Switzerland

[4]Institute for Implementation Science in Health Care, Faculty of Medicine, University of Zurich, Zurich, Switzerland

[5]Dalle Molle Institute for Artificial Intelligence (IDSIA), The University of Applied Sciences and Arts of Southern Switzerland (SUPSI), Lugano, Switzerland

[6]Australian Centre for Health Engagement, Evidence and Values, University of Wollongong, Wollongong, Australia

[7]Centre for Artificial Intelligence, Zurich University of Applied Sciences (ZHAW), Zurich, Switzerland

[8]Institute of Biomedical Ethics and History of Medicine, University of Zurich, Zurich, Switzerland

[9]School of Medicine, University of St. Gallen, St. Gallen, Switzerland

[10]Bioethics & Health Humanities, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

[11]Data and Web Science Group, School of Business Informatics and Mathematics, University of Mannheim, Mannheim, Germany

[12]Department of Neurosurgery, University of Freiburg - Medical Center, Freiburg im Breisgau, Germany

[13]Lausanne University Hospital (CHUV), Lausanne, Switzerland

[14]Edinburgh Law School, University of Edinburgh, Edinburgh, United Kingdom

[15]The Ethox Centre and Wellcome Centre for Ethics and Humanities, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom

[16]The Institute for Ethics in AI, Faculty of Philosophy, University of Oxford, Oxford, United Kingdom

[17]Development Economics (DEC), World Bank Group, Washington, DC, United States

[18]Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland

[19]Institute for the History and Philosophy of Science and Technology, University of Toronto, Toronto, ON, Canada

[20]Schwartz Reisman Institute for Technology and Society, University of Toronto, Toronto, ON, Canada

[21]The Centre for Technomoral Futures, University of Edinburgh, Edinburgh, United Kingdom

Corresponding Author:
Georg Starke, MD, PhD
Institute for History and Ethics of Medicine
Technical University of Munich
Ismaninger Str. 22
Munich, 81675
Germany
Phone: 49 8941404041
Email: georg.starke@tum.de

Abstract

Background: The integration of artificial intelligence (AI) into health care has become a crucial element in the digital transformation of health systems worldwide. Despite the potential benefits across diverse medical domains, a significant barrier to the successful adoption of AI systems in health care applications remains the prevailing low user trust in these technologies. Crucially, this challenge is exacerbated by the lack of consensus among experts from different disciplines on the definition of trust in AI within the health care sector.

Objective: We aimed to provide the first consensus-based analysis of trust in AI in health care based on an interdisciplinary panel of experts from different domains. Our findings can be used to address the problem of defining trust in AI in health care applications, fostering the discussion of concrete real-world health care scenarios in which humans interact with AI systems explicitly.

Methods: We used a combination of framework analysis and a 3-step consensus process involving 18 international experts from the fields of computer science, medicine, philosophy of technology, ethics, and social sciences. Our process consisted of a synchronous phase during an expert workshop where we discussed the notion of trust in AI in health care applications, defined an initial framework of important elements of trust to guide our analysis, and agreed on 5 case studies. This was followed by a 2-step iterative, asynchronous process in which the authors further developed, discussed, and refined notions of trust with respect to these specific cases.

Results: Our consensus process identified key contextual factors of trust, namely, an AI system's environment, the actors involved, and framing factors, and analyzed causes and effects of trust in AI in health care. Our findings revealed that certain factors were applicable across all discussed cases yet also pointed to the need for a fine-grained, multidisciplinary analysis bridging human-centered and technology-centered approaches. While regulatory boundaries and technological design features are critical to successful AI implementation in health care, ultimately, communication and positive lived experiences with AI systems will be at the forefront of user trust. Our expert consensus allowed us to formulate concrete recommendations for future research on trust in AI in health care applications.

Conclusions: This paper advocates for a more refined and nuanced conceptual understanding of trust in the context of AI in health care. By synthesizing insights into commonalities and differences among specific case studies, this paper establishes a foundational basis for future debates and discussions on trusting AI in health care.

(J Med Internet Res 2025;27:e56306) doi: 10.2196/56306

KEYWORDS

expert consensus; trust; artificial intelligence; clinical decision support; assistive technologies; public health surveillance; framework analysis

Introduction

Background

The integration of artificial intelligence (AI) systems into health care is one of the most widely anticipated transformations of health systems worldwide [1]. AI promises improved diagnostics [2,3], optimized treatment strategies [4], and early identification of at-risk patients [5]. Prominent examples also include AI for assistive technologies offered to patients directly [6,7] and for informing public health decision-making beyond the individual [8]. More recently, large language model–based applications promise to revolutionize health care, with applications spanning from clinical research and processes to physician-patient relations [9]. Despite this range of potentially beneficial applications, the broader adoption of AI systems in health care has been struggling due to many inhibiting factors. Problems arise at the level of development, with data bottlenecks impeding the training of machine learning models or a lack of user-centered and value-sensitive design procedures affecting AI acceptability [10,11], and stretch to the level of practical implementation of AI systems [12-14]. Questionable improvements in real-life health care settings [15], a lack of regulatory frameworks [16], and unresolved questions of reimbursement [17] further complicate adoption of AI in health care.

In this paper, we focus on a central inhibitor of successful AI adoption in health care, namely, the low levels of user trust in AI systems [18]. Understanding and fostering trust in AI remains challenging, not only practically but also conceptually. Interpersonal trust constitutes a complex and contested construct in philosophy and social sciences [19-24]. Moreover, accounts diverge on what constitutes being worthy of trust, namely, on the definition of *trustworthiness* [19,20]. Unsurprisingly, there is also no consensus across disciplines such as computer science, philosophy of technology, and social sciences on the definition of trust in AI and the capabilities that a trustworthy AI should maintain [25,26]. As a recent publication assembling existing empirical literature on trust in AI and clinical decision support put it, "Different groups [of researchers], whilst seemingly agreeing in principle that 'trust' and 'trustworthiness' are important, can in fact be referring to very different concepts and talking past one another" [27].

The combination of fragmented conceptual research, practical concerns, and implementation difficulties inhibits fostering warranted trust in AI (ie, trust caused by the trustworthiness of the AI) [25,28], which will be increasingly crucial for delivering health care [29]. As recent empirical work using path analysis has highlighted, trust in AI-based technologies seems to have a significant effect on users' intention to use such systems [30]. However, to build trust in—possibly trustworthy—AI systems in health care applications, we first need a conceptual understanding of trust in AI within the health care sector. This goal requires considering the diverse perspectives, expectations, and limitations that various research disciplines bring to the table in the context of human-AI interactions. Without conceptual clarity, attempts to foster the multifaceted concept of trust in AI—whatever it may *precisely* mean—run the risk of being inefficient or even detrimental, leaving users, including data scientists, physicians, and patients, vulnerable to placing trust in systems that may not warrant it or refraining from it in

situations in which grounds for trusting actually exist, an occurrence of what we might call "unwarranted distrust" [31,32].

Objectives

Bringing together international and multidisciplinary perspectives on the topic of trust in AI, this paper aims to provide common ground for defining trust in AI in health care as a reference point for future debates. To this end, we developed the first consensus statement on trust in AI in health care based on input from international experts on the topic drawing on iterative synchronous and asynchronous discussions of realistic case studies. Our results highlight the need for a more refined and nuanced understanding of trust in the context of AI in health care if the concept is to guide AI design and adoption processes and inform national and international governance.

Methods

Overview

Consensus statements from meetings of experts working on a particular topic constitute a common scientific approach across the fields of medicine [33,34] and ethics [35,36]. While they can take many different forms [37], their shared goal is to identify agreement among people working in a field and assemble multiple perspectives with peer-informed legitimacy [38]. The reporting of our findings follows the Accurate Consensus Reporting Document guidelines [37] and considered applicable elements of guidelines for reporting qualitative research [39,40].

Recruitment

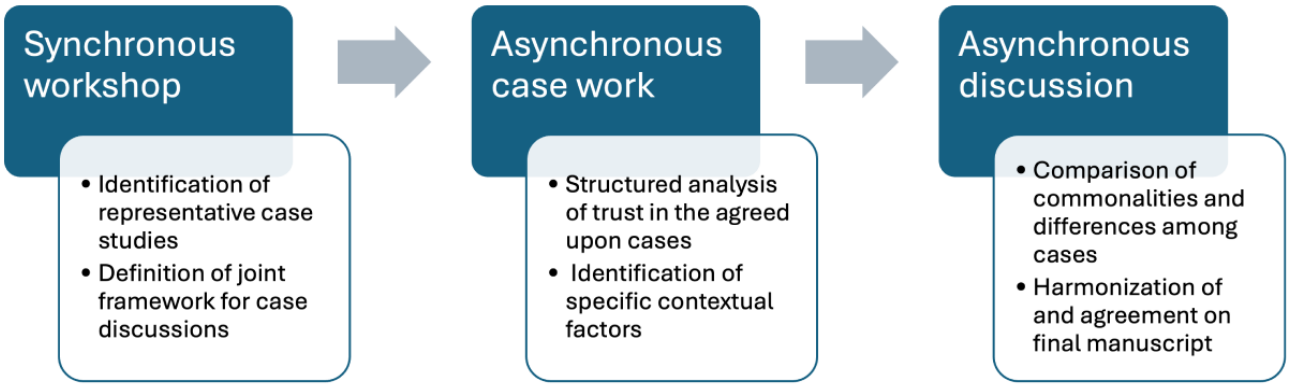
Recruitment for the consensus process took place as part of a 3-day workshop held at the École Polytechnique Fédérale de Lausanne in September 2022 in Lausanne, Switzerland. Authors GS, FG, AT, A Facchini, and MI, who organized the workshop, invited some participants directly based on their contribution to the field and to accommodate different types of expertise and backgrounds. They selected further participants from submissions to a call for abstracts for the workshop. Unanimous agreement concerning participants was achieved among conference organizers GS, FG, AT, A Facchini, and MI.

Consensus Process

Overview

The consensus process comprised 3 sequential steps involving both synchronous (step 1) and asynchronous (steps 2 and 3) activities. We describe them in the following sections. Figure 1 shows the steps of the consensus process and provides an overview of the intended outputs of each step.

Figure 1. The 3-step consensus process and envisioned results.



Synchronous Phase: Selection of Case Studies and Framework

We conducted a workshop involving an international group of researchers from various disciplines. To arrive at a consensus regarding trust in AI in health care, we discussed the topic from complementary angles in a series of group activities that iteratively involved talks on specific aspects of trust, moderated group discussions, plenary sessions, and interactive panels. The moderated group discussions were spread over 3 days. They covered the topics of (1) conceptualizing trust in the context of human-AI interactions, (2) conceptualizing trust in AI in the context of health care specifically, (3) requirements for trustworthy AI in health care, and (4) recommendations for implementing requirements for trustworthy AI in health care. To achieve consensus on the case studies, all participants reviewed them. Only case studies assessed as representative yet distinct from other cases by all participants were promoted to the next step.

To guide our analysis, we made use of a conceptual framework. The framework was proposed by FG, critically discussed, and agreed upon by all authors. A conceptual framework represents a phenomenon in the form of a network with interlinked themes that describe how the phenomenon works [41]. Conceptual framework analysis provides an established qualitative method to guide the identification and organization of key aspects and relationships in contexts that draw on multidisciplinary bodies of knowledge [41]. We used framework analysis to provide structure and coherence of discussions for the asynchronous steps, systematically evaluating various dimensions of trust in AI in health care.

Asynchronous Step: Developing Case Studies on Trust in AI in Health Care

To understand how trust unfolds in the context of the agreed upon 5 cases, in the second step, smaller groups of authors expanded on the individual cases agreed upon in step 1. All contributors involved in the writing are listed as authors.

Participants in this step were asked to explicate trust-relevant factors of the cases in accordance with the agreed upon conceptual framework. Responsibilities for coordinating the writing of individual case studies were distributed among the authors, with 1 person taking responsibility for leading each case study. KJ led the consensus for the first case study, JH led the consensus for the second case study, YSJA led the consensus for the third case study, GS led the consensus for the fourth case study, and JW led the consensus for the fifth case study. The results were compiled and shared by GS, FG, and MI.

Asynchronous Step: Comparison and Discussion of Case Studies

Having combined the individual case studies, we compared and discussed relevant aspects to synthesize our findings. We discussed our findings critically in light of existing literature and asked participants again to provide written feedback on the synthesized findings. This process was coordinated by GS, who wrote a first draft of the manuscript, which was then refined by FG, A Facchini, A Ferrario, and MI. All authors provided their feedback on the manuscript, which was incorporated and harmonized by GS and shared with all authors for approval.

Ethical Considerations

In accordance with the existing regulatory framework in Switzerland, such as the Human Research Act and the local rules at the École Polytechnique Fédérale de Lausanne as hosting institution, no ethics approval was required for this collaborative research project involving only consenting experts in the field.

Results

Overview

The results of our study unveil the complexities of understanding trust in AI within the health care domain. Our findings represent consensus among international experts guided by conceptual framework analysis. The following sections provide the characteristics of the involved experts and report the findings from the individual synchronous and asynchronous steps of our consensus process.

Recruitment

The characteristics of the 18 participants in the consensus process are described in Table 1.

Table 1. Participant characteristics—multiple entries possible for location and background (N=18).

Characteristics	Participants, n (%)
Gender	
Women	7 (39)
Men	11 (61)
Nonbinary	0 (0)
Prefer not to say	0 (0)
Geographic location	
Europe	16 (89)
North America	2 (11)
Oceania	1 (6)
Background	
Computer science	6 (33)
Ethics	5 (28)
Medicine	4 (22)
Philosophy	8 (44)
Public health	2 (11)
Social science	3 (17)

Consensus on Trust and Trustworthiness

Synchronous Phase: Selection of Case Studies and Framework

A total of 5 case studies emerged from the synchronous group discussion and were identified as paradigmatic examples of the various applications of AI in health care, namely, diagnosis, clinical risk assessment, public health surveillance, assistive technologies, and health care resource allocation. The involved experts agreed that these cases may not provide an exhaustive sample of all potential AI applications in health care yet allow for a meaningful representation and comparison of trust in AI across largely different health care settings.

To guide the asynchronous analysis of the individual cases, all involved experts identified and defined key components of a conceptual framework for trust in AI in health care. We developed this framework with the backdrop of existing conceptual work on trust in data use in health care, preliminary discussions of the shortcomings of present conceptual work describing what trust in AI is, and existing guidance on conceptual framework development for research and scale development in medicine [29,31,42-49].

Following existing conceptualizations of trust in the context of data use in health care [47], themes were grouped into two main areas: (1) the context of trust and (2) causes and effects of trusting. We considered trust as a context-specific, relational construct between 2 actors that is shaped by the environment in which it develops over time [50]. Therefore, our framework implemented the context specificity of trust by considering (1) the environment within which an AI system is deployed (ie, the health care application in which the AI operates), (2) the actors involved in the specific trust relationship, and (3) the factors that frame the trusting relationship. Here, framing factors influence "the process by which people develop a particular conceptualization of an issue or reorient their thinking about an issue" [51]. Examples of framing factors are historical context, cultural aspects, norms and values, fears, public sentiment, overarching belief systems, and religious beliefs. In addition to context, our framework includes factors that support or inhibit trust as well as the effects of trusting AI systems. Causally important themes for trusting relationships typically relate to factors that make AI trustworthy or untrustworthy in the eyes of those placing trust. Examples are the reliability and accuracy of the AI and, arguably, the level of interpretability of its outcomes [28,52]. In summary, our framework comprised 5 themes: environment, actors, frames, causes, and effects (Textbox 1).

Textbox 1. Conceptual framework for warranted trust in artificial intelligence (AI) in health care.

Context of trust in AI in health care

- Environment: What is the setting for which the AI system is intended?
- Actors: Who are the involved actors in the trust relationship?
- Frame: What frames trust in relation to AI?

Cause and effect of trust in AI in health care

- Cause: What makes AI trustworthy?
- Effect: What is the effect of trust in relation to AI?

Asynchronous Step: Developing Case Studies on Trust in AI in Health Care

Drawing on our framework and previous research, smaller groups of authors examined particular aspects of AI used for diagnostic purposes, clinical risk assessment, public health surveillance, assistive technologies, and health care resource allocation. The results of the individual groups led by KJ, JH, YSJA, GS, and JW, respectively, are reported in the following sections.

Diagnostic AI in the Clinic

Machine learning and, more specifically, deep learning have proven to be particularly suitable for computer vision, especially image processing and pattern recognition. Important preconditions to apply these techniques in medicine, such as suitable infrastructure or availability and storage of digital images, are also being increasingly met, at least in high-income countries. This may explain why most current and proposed AI systems in medicine are used to aid image-based diagnostics in fields such as radiology, ophthalmology, and pathology [53-55]. As a large part of the workload of, for instance, radiologists is to interpret medical images [56], diagnostic AI systems could increase this capacity; support or take over the tasks; and, thereby, change the organization of image-driven diagnostics [57].

Some studies have indicated that, under specific circumstances, AI systems achieve accuracy that is at least equal to that of expert radiologists and pathologists or even outperform them when detecting, classifying, and segmenting tumors in ultrasonography, x-ray imaging, magnetic resonance imaging scans, and digitized microscopy slides [58,59]. These findings have raised massive enthusiasm and fueled the motivation to develop and use algorithms in image-driven diagnostics. As a particular example, we will focus on a recently approved AI system for the diagnostic analysis of chest x-rays [60].

The system aids diagnosis by analyzing chest x-rays, the most frequent radiological examination worldwide. It does so by classifying the scans in nonpathological and potentially pathological cases. For the first group, the program provides a fully automated report, removing the necessity for any further follow-up on the image by radiologists. The scans of the second group are forwarded to trained physicians for further radiological analysis. A test of the software on approximately 10,000 chest x-rays from Finland showed a very high sensitivity (99.8%) with a low specificity (36.4%), yielding a very low probability of missing any critical findings [61].

However, the use of diagnostic AI is associated with technological obstacles as well as epistemic and ethical challenges [62]. Although the potential of self-supervised learning to generate expertlike annotations has been demonstrated, an insufficient number of expertly annotated diverse images is currently often a limiting factor in training AI models on a technical level [63]. Large datasets, for instance, in pathology, can also pose problems if downsampling leads to crucial information being lost, demanding different kinds of representation [64]. Beyond such technical difficulties, there are also more fundamental obstacles, including the fact that diagnostic tasks are not usually choices between 2 distinct outcomes [65]. Another concern is the frequent lack of explainability of medical AI, which limits physicians' ability to recheck the AI's output [66]. It has been argued that this raises particular challenges as decisions based on black-box AI systems are potentially not fully interpretable for the human physician. Moreover, when offering explanations, it is essential

to consider the epistemic requirements of users (eg, radiologists and patients). Explanations that lack a clear connection to clinical knowledge despite being technically accurate fall short in delivering the desired support for clinical practice. Similarly, patients who wish to understand the basis of their diagnosis may find explanations that diverge from their own illness experiences unsatisfactorily [67,68]. That said, there is empirical evidence that a lack of explainability is not necessarily perceived as a problem by physicians if a diagnostic system has been properly validated and other detailed information about the AI system is available [69,70].

The case discussed here takes place in a clinical environment of image-driven diagnostics. It directly involves medical professionals and the AI system, whereas patients do not interact with the system itself; in addition, the developing company and regulatory bodies are indirectly involved. Framing factors are discourses about job security in radiology and the danger of physicians being replaced by AI [71]. Trust leads to acceptance of the system by physicians, potentially at the cost of deskilling, and is arguably fostered by the accuracy of the diagnostic program in question [72].

Predictive AI for Clinical Risk Assessment

AI algorithms can be used to predict individual patient trajectories. Examples include the prediction of COVID-19 severity [73], the prediction of delirium [74], prognostic models of respiratory diseases [75], or systems predicting future lung cancer risk [76]. In addition, predictive models are crucial for determining individualized treatment plans, often particularly relevant in oncology [77]. The difference from the diagnostic use case is that the prognostic case concerns the future (ie, informs decision-making under situations characterized by an inherently larger uncertainty). The prognostic prediction informs choices of action, treatments offered, support, and resource allocation and, therefore, can, in turn, influence the very outcome that has been predicted [78].

As a specific case study, let us consider the prediction of circulatory failure in intensive care units (ICUs) [5]. In ICUs, a great number of machines monitor the state of patients and need to be observed constantly as patient conditions may worsen rapidly. The existing technologies for life support and vital sign monitoring provide information and alerts that clinicians need to oversee. However, the high rates of alerts, including false positives, may cause alert fatigue for practicing clinicians, impairing optimal care. In this context, systems that predict severe deterioration more accurately can obviate alert fatigue and potentially lead to better patient outcomes. The system reported in the study by Hyland et al [5] integrates information from multiple organ function–monitoring systems to alert clinicians to potential circulatory failure 8 hours in advance. Note that the information available to the system is necessarily incomplete, and it is impossible to predict future circulatory failure with 100% accuracy using any algorithm. However, the system in question successfully predicted 90% of cases of circulatory failure in the test set and 82% of them >2 hours before the event [5], so it would likely prove useful in a clinical setting.

The system described here can only be deployed in the environment of hospitals or clinics. Physicians, nurses, and potentially caregivers may interact directly with the AI system, and the social circle of a patient could also potentially receive information derived by the AI system, whereas the patients admitted to an ICU will likely not interact with the system itself. Critical contextual factors are the system's use in stressful situations with the risk of severe consequences, sometimes under time pressure, and the need to synthesize a multitude of complex information. Trust in such a situation can be built by overall improved outcomes in settings where the algorithm is used, ideally by investigating it through a randomized controlled trial. In addition, there should be reasons to believe that the AI system is not biased or will not lead to harmful consequences for any subgroups of the population and that it produces reliable individual decisions that are robust to design choices [79,80] or randomness [79,81] in the AI pipelines. Trust built on these premises will promote AI's acceptance and use. However, as with other medical interventions aiming to change the future, such as screening programs [82,83], trust could also lead to worse outcomes if the system does not actually deserve it (ie, if it is not *trustworthy*).

AI for Public Health Surveillance

Public health surveillance involves the identification of signs of population-level health anomalies and potential disease outbreaks from a heterogeneous collection of data sources [8]. With its intrinsically data-driven nature, public health surveillance has increasingly become a domain for AI applications as AI provides a range of novel methods for data collection and data analysis in large and varied samples [8]. A comprehensive survey of the literature showed that the COVID-19 pandemic highlighted the potential of AI systems in improving surveillance of infectious disease outbreaks [84].

Consider the example of EPIWATCH among the numerous AI-based surveillance systems that provide early signs of disease outbreaks [85]. Developed at the University of New South Wales in Australia, EPIWATCH is an open access web-based tool that offers an interactive dashboard with a sortable, searchable, and filterable global map based on 30 days of data [85]. The tool has been successfully used to model and identify patterns of disease outbreaks, providing crucial information on risk factors and geographic distribution.

For instance, using publicly available data, EPIWATCH has been used to trace global Zika virus outbreaks, tracing transmission modes, affected countries, and complications such as microcephaly [86]. Similarly, a different group of authors used EPIWATCH to model the global epidemiology of hepatitis A, identifying the United States and Europe as major centers of outbreaks, and provided quantitative global data on the most common risk factors, which seem to be homelessness and foodborne outbreaks [87]. Such data can be helpful to inform policy making at both a global and local level.

Let us summarize the contextual factors of this case. Contrary to the previous examples, AI systems for public health are deployed in nonclinical settings. This web-based example is potentially accessible by anyone with a computer or mobile phone with an internet connection interested in data analysis to

predict disease outbreaks. Therefore, relevant actors include the developers as well as a broad spectrum of potential users, ranging from public health practitioners and policy makers to the public. Important framing factors for trust are the severity of the modeled disease as well as the stage of the epidemic or pandemic, impacting the relevance and acceptability of the tool. As the COVID-19 pandemic has demonstrated, antiscience sentiments and conspiracy theories contribute to how public health tools are viewed. Finally, usability aspects further influence whether information will be taken up by practitioners. Trust in the application will likely lead to increased uptake among decision makers. The trustworthiness of the platform developers and endorsement by authorities, such as public health experts and policy makers, will further determine whether the wider public will trust the model's output.

AI for Assistive Neurotechnology

In comparison to the examples discussed previously, a quite different trust relationship can be found in the case of assistive neurotechnologies based on brain-computer interfaces (BCIs). BCIs can, for instance, be used for neurorehabilitation, restoring lost function, or augmenting and enhancing existing cognitive or motor abilities [7]. AI, especially data-driven machine learning with deep neural networks, plays a crucial role in their development and use as these methods enable and facilitate the modeling and decoding of complex neural signals as well as the adaptation to individual users [88]. To delineate differences in the trust relationship between AI-based neurotechnology and other medical AI, we focus on a specific example: a BCI-guided robotic hand orthosis for rehabilitative purposes.

Orthoses are external wearable devices that can be "used to compensate for impairments of the structure and function of the neuromuscular and skeletal systems" [89]. In line with numerous existing research projects at the level of preclinical prototypes [90], let us consider a specific BCI-based robotic orthosis that allows patients with stroke to restore some volitional control of hand-grasping movements. To do so, the system records electrical activity of the dominant motor cortex noninvasively through an electroencephalography cap and extracts features from the measured signals using machine learning–based classifiers to extract patient-specific neural markers of intended hand movements, which are then translated into mechanical movements of the robotic hand orthosis.

There are several aspects of such personalized AI-based assistive neurotechnologies for neurorehabilitation that set them apart from medical AI used for diagnosis, prediction, or public health surveillance. One crucial difference that stands out is that the AI in question is *embodied* in the sense that the AI used for decoding neural signals and for translating them to mechanical movement is inextricably linked to a physical object—namely, the hand orthosis. This implies that physical design aspects play a crucial role on whether trust is expedited to the device [91].

However, there are further differences that impact the trust relationship involving actors, the environment, and framing. First, assistive rehabilitative neurotechnology is designed for patients with impairment of cognitive or motor functions who need to engage with the technology. These patients are also the main trusters of this technology, not health care professionals or public health experts. Second, suppose the neurotechnology also helps patients in their everyday life. In that case, it should not only work under highly constrained and controlled laboratory or clinical conditions but also in a patient's home. This implies additional demands for the devices' ease of use and their security in a more open environment, which would be crucial for the trust of bystanders in these devices. Third, due to the BCIs' physicality, trust in assistive neurotechnology may also be framed differently, invoking widely known images from science fiction [92]. Fourth, users may be concerned that decoding their brain activity could be used not only to control the robotic orthosis but also to infer other types of information that would not be available using other sensing technologies, thus raising potential threats to their privacy. Therefore, addressing potential fears of patients and designing AI that assures human control (in the sense of "human in the loop") at any moment adds to the complexities of building trust in such embodied AI.

To summarize, the assistive AI system discussed here is embedded in an environment of clinical neurorehabilitation. Its trusters are primarily patients and rehabilitation experts who trust in the developing engineers and the pertinent regulatory bodies. In the context of embodied AI, many specific framing factors influence its perception, especially from science fiction literature and cinema. It is also shaped by public attitudes and policies on related technologies such as robotics. To build trust in this environment and prove the accuracy of the AI, a lack of conflicts of interest on the part of the developers; sufficient understanding of the underlying technology by its users and rehabilitation experts; and independent, long-term technical support seem key. Successful trust building will result in wider acceptance and adherence by users, as well as potential inclusion of the technology in public health programs, facilitated regulatory compliance, and improved insurance reimbursement. Taken together, these elements can facilitate access and increase the technology's affordability by expanding the number of its users.

AI for Health Care Resource Allocation

A 2019 study in *Science* revealed that software provided by the health service company Optum, and which was being used to manage the care of >200 million patients at hospital centers across the United States, was significantly—even if unintentionally—biased against Black patients [93]. The machine learning algorithm aimed to predict the future health care needs of patients and direct extra medical care toward the most vulnerable. However, it was shown to systematically underestimate the needs of Black patients. It did this because it used health care costs as a proxy for need in its risk score despite health care costs not being a neutral and reliable proxy for health needs in this context. This led it to assign consistently lower risk scores to Black patients compared to White patients who were equivalently sick. This was despite the model being "race blind" in the sense that race was not specified in the input data [94].

When considering how this case study relates to trust in and trustworthiness of AI in health care, it is important to note that this is not a case of AI being introduced into a dyadic

physician-patient relationship but, rather, of its integration into the operations of complex health systems involving many actors and institutions. Therefore, philosophical accounts of trust that focus on the properties of individuals—these accounts traditionally comprise reliability *plus* some appropriate motivational state on the part of the trustee toward the truster, for example, goodwill [95]—are insufficient to capture the conditions necessary for trustworthy AI in these contexts.

Accounts of institutional trustworthiness are more useful here. They often emphasize features such as transparency as well as competence and reliability [96]. Accounts of trustworthiness that do not necessarily require a phenomenological state such as goodwill may also be useful. For example, the trust responsiveness account by McGeer and Pettit [97] emphasizes the importance of an agent—or, in our case, institutions—responding appropriately to the reasons for doing what they are being relied upon to do. Alternatively, one might provide a deflationary account that either equates trustworthiness with reliability [48] or rejects trust as an appropriate attitude altogether [98].

What does it mean to meet these conditions in this and similar cases of integrating machine learning into resource allocation decisions across an entire hospital or health system? Trust responsiveness demands that systems such as the one provided by Optum be introduced to promote health and well-being and improve patient outcomes and not to meet other goals such as cutting costs to maximize profits for insurance companies and health systems. Using a system such as Optum may not necessarily require internal algorithmic transparency, which is often the focus of work considering the ethics of machine learning algorithms in health care [99], yet may be difficult to translate into a genuine understanding of the AI system by clinicians or patients. However, there needs to be openness regarding design decisions and value choices involved in the algorithm's construction and implementation [100]. Demonstrating these may require engaging in ethical and algorithmic impact assessments during the design process [101,102], welcoming algorithmic auditing after implementation [49], and abiding by regulatory frameworks throughout the product life cycle [103]. Ideally, transparency for AI systems for resource allocation would highlight both global explanations [104], given that the goal of the algorithm is to distribute resources across a population [105], and local explanations for individual decisions as the biased program impacts individual care. Both are potentially useful for promoting fairness [104].

In addition, it is crucial to consider contextual factors impacting such trust relationships. The program was developed by a for-profit health service provider in the US health care system and used to determine resource allocation. Key actors were the health service company itself, which provided the algorithm; the health systems; clinicians interacting with the algorithm's recommendations; patients having their care influenced by the algorithm's recommendations; and, finally, regulatory bodies and (internal and external) algorithmic auditors. Important framing factors include media reporting on algorithms and their impact and, on a conceptual level, theories of institutional trust. Such trust, though unwarranted here, would result in the model's acceptance and use. It is supported by an AI's reliability and competence as well as by the transparency of the developers concerning design choices and the values reflected in them.

Asynchronous Step: Comparison and Discussion of Case Studies

As the 5 cases highlight, medical AI is a multifaceted concept encompassing a diverse group of systems, environments, actors, and framing factors. Talking about trust in medical AI in general can only be a first approximation to the phenomenon. However, some general elements of trust were accepted by all experts. First, they agreed that trust provides a way to deal with complex situations in the face of uncertainty [106,107] and is established in anticipation of a beneficial outcome. Therefore, in a rather minimal fashion, and in line with existing literature, an agreed upon definition of trust in AI in health care focuses on giving discretionary power to an AI system with respect to a specific health care–related task [29,108]. At the same time, it was agreed that situations of trust were characterized by risks, rendering the trusting agent vulnerable to betrayal of trust [21]. To minimize such risks, potentially of life and death in health care, experts agreed that we need to only promote warranted trust in medical AI (ie, trust that is justified, plausible, and well grounded [19,28,49]). In the literature as well as in regulatory contexts, AI systems that warrant such trust have been described as trustworthy [49,109,110]. Therefore, in the eyes of the experts, a complete assessment of trust in AI includes epistemological and practical considerations of trust as much as examining the trustworthiness of the AI in question (ie, the functionalities of an AI system that morally and practically vindicate a relationship of trust [111]).

The involved experts agreed on many practical shortfalls of current AI in health care potentially impacting its trustworthiness. These include the lack of explainability of many AI algorithms [28,66,112]; the difficulties of benchmarking model performance [113]; or the longtail of exceptional situations that will be difficult to anticipate at the development stage, from nonstandard data inputs to the system's deployment in a health care system for which it was not originally trained [114]. In the eyes of the experts, anticipatory difficulty also relates to risks of AI perpetuating, hiding, or reinforcing statistical and social biases as well as other data problems [115,116]. Finally, it was stressed that many medical AI tools so far lack proper engagement with stakeholders in their development, which is crucial for ensuring that they adequately address a real clinical problem and identifying their relevant ethical implications [117].

When looking at the individual case studies, commonalities and differences among them arose at micro and macro levels. Table 2 provides an overview of the contextual factors of each case study.

Table 2. Contextual factors for the 5 case studies.

	Environment	Actors	Framing factors	Causes of trust	Effects of trust
Diagnostic AI[a] (chest x-rays)	Image-driven diagnostics (radiology)	Medical professionals and AI system; patients to a limited extent	Discourse regarding job security and potential AI replacement	Accuracy, design transparency, and human competencies and virtues	Acceptance of systems by physicians, potentially at the cost of deskilling
Predictive AI (ICU[b] setting)	Clinical setting of an ICU	Physicians, nurses, and AI system; patients to a limited extent; and potentially caregivers and family members	Stressful situations potentially with a need to act under time pressure, risk of severe consequences, the need to synthesize too much information, and alert fatigue	Accuracy, transparency, and explainability; fairness; exclusion of harm; and rigorous testing (eg, in the form of an RCT[c])	Acceptance and use of the system, potentially at the risk of erroneous clinical decisions following misleading predictions
Public health AI (disease outbreak model)	Nonclinical setting—publicly accessible web-based tool for the analysis of heterogeneous data	Developers, public health practitioners, policy makers, and the public	Stage and severity of disease outbreak; usability aspects (eg, intuitive interface or data visualization); and, potentially, antiscience sentiments and conspiracy theories with regard to the disease and health service providers	Historical accuracy and endorsement by authorities	Acceptance and use of the system by public decision makers (public health experts and policy makers)
Assistive AI (neurorehabilitation)	Clinical neurorehabilitation; elective use of different technologies for different activities, potentially every day	Patients and their caregivers and social circle, potentially including employers, engineers, and regulators	Clinical setting, science fiction literature and cinema, and public attitudes and policies on related technologies	Accuracy, privacy, lack of conflicts of interest, independence, long-term technical support, and user understanding of the underlying technology	Technology acceptance by users, health care professionals, and health care providers; potentially facilitated reimbursement and increased affordability and accessibility
Resource-allocating AI (predicting costs and needs)	Health service providers and health care system	The developing company providing the algorithm, the health system implementing it, the clinicians interacting with it, the patients having their care influenced by the algorithm, and regulatory bodies and algorithmic auditors	Media reporting on algorithms and their impact and theories of institutional trust	Reliability, accuracy, transparency, design, and model-centric explanations	Acceptance and use in health care systems

[a]AI: artificial intelligence.

[b]ICU: intensive care unit.

[c]RCT: randomized controlled trial.

Most commonalities among the different cases can be found with regard to their causes and effects, some of which arise from the very structure of trust. A common effect of trust was found in it leading to wider acceptance of the technology in question. Such acceptance implies an increased uptake by its intended end users, resulting in a larger role in health care systems, facilitating access, and potentially increasing the systems' affordability. In line with the literature, common aspects with regard to trust-supporting features included both aspects intrinsic to the system, such as transparency and explicability, and extrinsic factors, such as proper external validation and assessment of potential biases [25,29].

Across the discussed cases, some forms of transparency were identified as supportive of establishing trust. Such transparency included openness by the developers about their measures of a well-working system, enabling an alignment with the goals and values of patients and clinicians. Promoting transparency was also assessed as a good way of promoting trust among clinicians using the systems. For instance, there is some empirical evidence that being transparent about the functioning of a triage system makes clinicians more likely to trust it [118]. The extent to which this is possible depends on the exact nature of the system, with clear differences among algorithms that are interpretable by design [119], algorithms that can be explained in terms of their exact functioning on the level of predictors and their weights, and algorithms whose decision process cannot be interpreted causally.

Informational openness was also recognized as central to promoting a system's external trustworthiness (eg, if a system's reliability and performance metrics are publicly available). From the discussed cases, it also emerged that rendering information accessible to all relevant auditing bodies is crucial for mechanisms of accountability and, in doing so, for creating the necessary epistemic basis for trust at an individual level as much as among the public, who can have confidence that the AI systems have been subject to meaningful scrutiny. In this sense,

transparency was deemed to support trust among affected communities and stakeholders by providing the means to evaluate claims made about these systems with regard to bias and discrimination. AI systems reproducing inequalities and exacerbating injustices may, in turn, erode public trust not only in the AI-based tools themselves but also in stakeholders involved in building, deploying, and using such systems. Addressing and testing for such biases was deemed particularly relevant considering that "personal and collective experiences with discrimination or degradation-along lines of race, class, gender, or other personal characteristics especially create reasons for suspicion if not outright distrust" [99]. However, as the experts agreed, such testing needs to go beyond the sole measure of excluding apparently discriminatory information. Deep learning models may predict a patient's race from medical images such as chest and hand x-rays and mammograms [100] or identify patient self-reported race from redacted clinical notes [101] despite human experts being unable to do the same. Therefore, as indicated by the resource allocation case study, discrimination can occur even when models are apparently *race blind* through implicit proxy features, as is well understood and documented in algorithmic fairness research [120]. Moreover, discrimination through differential impact across subpopulations might occur precisely because models are insufficiently sensitive to inequitable distribution of social determinants of health along ethnic or racial lines [121].

Beyond these common themes, our cases highlight important contextual differences among trust in different kinds of medical AI. In particular, these relate to (1) framing factors, (2) previous knowledge of the targeted trusters demanding different levels of explicability and transparency, and (3) the different risk-benefit trade-offs in different environments. Sometimes, the risks in medicine are grave for individual patients, such as in an ICU setting. However, in the resource allocation setting, the risk-benefit trade-offs concern both individual-level consequences of allocating or denying a resource to an individual and group- or population-wide benefits or harms due to better or worse resource use.

Discussion

Principal Findings

Using conceptual framework analysis, our study provides the first consensus-based publication investigating trust in AI in the domain of health care. Commonalities and differences emerged from the examination of our case studies, particularly with respect to causes and effects of user trust. Across all cases, trust was found to enhance acceptance and adoption of AI technologies, and some factors such as a system's accuracy were similarly deemed to be crucial for the trustworthiness of all discussed systems. Developers' openness about a system's internal workings, assumptions, and value judgments underlining the technical choices were also considered crucial across the case studies. Such transparency fosters alignment with the goals and values of health care professionals and patients; supports accountability mechanisms; and may help address biases and discrimination, which is essential for building trust among affected communities and stakeholders. However, our analysis also highlights contextual differences among trust in various forms of AI in health care. These differences relate to a multitude of case-specific aspects, such as framing factors, previous knowledge of the trusters, and risk-benefit trade-offs in different environments, necessitating tailored approaches to establish trust and trustworthiness. This finding has implications for anyone aiming to foster trust in AI within the health care domain, from developers and health care professionals using AI systems to public health experts and regulators—trust-building measures should distinguish among different types of AI systems considering not only implied risk levels and legal responsibilities in case of errors but also the factors outlined in our conceptual framework: context, actors, discourse, and the mechanisms of trust building.

Implications

Our findings have implications for developers, patients and health care professionals, and regulators aiming to increase trust in medical AI systems. The developing companies should communicate openly about the algorithmic design and training of their AI systems and tailor their level of transparency to the communicative needs of their respective audiences [122]. Such needs will be very different depending on whether the AI is used by patients directly (assistive AI case), trained physicians (diagnostic and predictive AI cases), health economists and policy makers (resource allocation AI case), or potentially the interested public (public health AI case). Individual AI end users need to be educated on the technology's limitations and carefully consider who bears the risk and responsibility of involving the AI—especially if it is not themselves.

At a regulatory level, distinguishing among different kinds of AI systems in medicine seems also crucial, regulating them according to the level of implied risk as currently proposed by the European Union AI Act—or possibly at a finer-grained level specific for health care [123]. Regulation is also needed to clarify who bears legal responsibility in the case of specific errors. While there is a clear responsibility by developers to avoid systematic errors as much as possible, for the foreseeable future, it will not be possible to design an AI system without any errors. Therefore, some responsibility also falls on health care providers who use the systems as part of their workflows to use them appropriately and with the right level of human oversight. This requires strengthening both the technical capabilities of health care practitioners to evaluate AI systems and the responsibilities placed on the developers of systems to openly document where a system may be expected to generate errors. Given the limits of both, proper certification of AI systems is furthermore crucial, allowing for trust in an overseeing institution to be extended to the system itself [124].

A final important point is that there should be a focus on institutions implementing these tools in a trustworthy manner, not on increasing trust. There are several reasons for this. First, trust could be increased through marketing and presentational gimmicks targeted at patients and clinicians, which would do nothing to increase trustworthiness [105]. Second, trust may be hard to achieve given prevailing and justified distrust in health service providers such as Optum and private health systems and marginalized communities' distrust in health care systems due

to historic or entrenched disparities in outcomes or experiences of care [125]. In contrast, this may be less of a problem where there are higher levels of trust in health institutions that might implement AI systems [106]. Finally, the size and complexity of health systems complicate questions of where, why, and in whom the individual patient should place trust within these systems [87]. This creates challenges in assessing how far the design and implementation of an algorithmic system for resource allocation impacted the attitudes of trust in affected patients.

Limitations

While our findings aim to provide a starting point for fruitful further work on AI in health care, there are several limitations to our consensus process and its results. Of these limitations, 3 are related to the selection of experts for our consensus process. First, owing to the recruitment of participants working on trust in AI in health care, our work is biased toward a position that considers trust a useful and meaningful concept in the context of AI in the first place. While this position is supported by a substantive part of the literature [28,31,49,52], it should be noted that there are also outspoken critics of using the notion of trust with respect to AI in the first place [98,126,127]. Second, given the location of our workshop, our recruitment focused largely on experts from Europe, with only a small addition of participants from North America and Oceania. While this may reflect the central role that trust and trustworthiness have taken in regulatory debates across the European Union, it limits the generalizability of our findings to other contexts. A recent systematic review of empirical research on trust in AI also revealed a lack of diversity in the discussion surrounding this topic [26]. This review underscored the need for a broader range of perspectives to more comprehensively understand and address the complexities of trust in AI systems [26]. Third, to address the lack of agreement on trust in AI in health care among experts, our work focused on expert agreement, leaving out some other relevant stakeholders. As the backgrounds of our participants highlight, some stakeholders were actively involved, from developers to practicing clinicians, but others were not included, such as patients, nurses, or hospital administrators. Further work is needed to address this gap.

There were 2 additional limitations more conceptual in nature, and we thank an anonymous reviewer for flagging these. First, our framework does not distinguish between trust *before* and trust *during* use of an AI-based system, yet causal and framing factors of trust may evolve with the experience of using a technology over time. In fact, all our case studies focus more on the initial adoption of a technology and less on trust evolving during its use, reflecting the current early-stage integration of AI systems into health care. That being said, we do hold that our framework can adequately reflect the relevant factors of trust in different AI systems in health care—even if they may change over time. Second, our work did not focus on the phenomenon of distrust despite its undebatable influence on user acceptance and its irreconcilability with trust. We consciously omitted distrust as it is often considered to not be a mere absence of trust but a more complex, richer phenomenon [19,23,95]. Therefore, we believe that a consensus on distrust deserves a paper of its own. However, it should be noted that distrust toward a specific technology, or potentially against AI in general, can of course play a crucial role in inhibiting trust building at a causal level.

Finally, when considering wider conceptual work on trust and the focus on warranted trust in this paper, we need to acknowledge a body of literature that understands trust to be motivated by emotions rather than calculated decisions [128]. The conceptual thinking in this paper leans more toward a cognitive approach to trust building. Thereby, the effect of affective states on trust might be somehow undervalued. However, these affective states can play an important role in trusting and accepting AI in medicine, although their contribution to trusting AI can vary greatly across different instances of human-AI interactions [129]. Therefore, it will be necessary to extend the conceptual thinking beyond cognitive approaches to encompass traits of affective trust.

In summary, our findings suggest that achieving trustworthy AI systems in health care requires a multifaceted approach bridging human-centered and technology-centered approaches. While regulatory precision and boundaries help provide the legal basis needed to develop trustworthy AI, and while technological design features are critical to successful AI development, communication and positive lived experiences with AI systems are at the forefront of user trust. From a user perspective, especially for those who are not AI aficionados, trustworthy technological features and legal compliance should be a sure thing, whereas communication and positive experiences point beyond the technical sphere of AI.

Conclusions

This paper examines 5 diverse AI systems in health care, revealing the complex landscape of trust in this field. These cases highlight how AI in health care is influenced by various environments, actors, and framing factors. While discussing trust in medical AI in a broad sense may serve as an initial overview, the specific nuances uncovered in these 5 cases demand a more detailed understanding to characterize trusting AI in health care.

Ultimately, however, focus should shift toward ensuring that institutions implement these tools in a *trustworthy* manner rather than merely aiming to indefinitely increase *trust*—only warranted trust is needed for valuable AI implementation. This approach recognizes the complexity of trust dynamics, the risk of superficial increase in trust without a substantial enhancement of trustworthiness, and the importance of trustworthiness in health care institutions. Despite all domain-specific variations, addressing these challenges and embracing transparency and accountability by design can help build a foundation of trust that will underpin the successful integration of AI into health care for the benefit of patients, physicians, and society at large.

Acknowledgments

This paper was produced within the framework of the workshop titled “To trust or not to trust: When should AI be allowed to make decisions?” funded by the Swiss National Science Foundation (SNSF) under grant IZSEZ0_213480 (applicant: MI; coapplicants: A Facchini and FG) and with local support from the École Polytechnique Fédérale de Lausanne. At the time of the workshop, A Ferrario was affiliated with the Mobiliar Lab for Analytics at Federal Institute of Technology Zürich and gratefully acknowledges their support. FG acknowledges the support of the Digital Society Initiative, University of Zurich. Unrelated and outside of the work leading to this paper, FG acknowledges funding from Novartis International AG, Sanitas Krankenversicherung (Stiftung), the Swiss Academy of Engineering Sciences, and the World Health Organization. RC acknowledges the support of the Digitalization Initiative of the Zurich Higher Education Institutions fellowship program of Zurich University of Applied Sciences digital. KJ acknowledges funding from the Wilhelmina Onderzoeksfonds (grant wkz22040701). ER’s research is supported by a Medical Sciences Graduate School Studentship, issued by the Nuffield Department of Population Health, and the Baillie Gifford-Institute for Ethics in AI Scholarship, both associated with the University of Oxford. PT acknowledges funding from Warsaw University of Technology within the Excellence Initiative – Research University program (grant 1820/97/Z01/2023). A Facchini and AT acknowledge funding from the University of Applied Sciences and Arts of Southern Switzerland Best4EthicalAI research program. MI and GS also acknowledge funding from the ERA-NET NEURON network and the SNSF under grant 32NE30_199436 (HybridMinds). GS acknowledges generous support from Fondation Brocher, Hermance, to finish a first full draft during a research residence. The work reported in this paper was funded by the SNSF (grant 32NE30_199436).

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request. The asynchronous consensus-building process was tracked in repeatedly updated documents.

Authors' Contributions

GS, FG, AT, A Facchini, and MI conceived and planned the workshop and the consensus process. All authors provided input on the choice of case studies and the framework. GS coordinated the consensus process and compiled the individual authors’ contributions for a first full draft of the manuscript. GS, AT, A Facchini, and MI drafted a first introduction, and GS, FG, AT, YSJA, JH, KJ, BK, EP, PT, JW, and A Facchini contributed text to the Methods section. KJ led the consensus for the first case study, JH led the consensus for the second case study, YSJA led the consensus for the third case study, GS led the consensus for the fourth case study, and JW led the consensus for the fifth case study. The discussion was jointly developed by GS, FG, AT, A Facchini, and MI. A Ferrario further contributed substantially to the development of the manuscript by providing extensive critical feedback during the review phase. All authors provided repeated critical feedback and read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. Jan 2019;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
2. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. Feb 02, 2017;542(7639):115-118. [FREE Full text] [doi: 10.1038/nature21056] [Medline: 28117445]
3. Lång K, Josefsson V, Larsson AM, Larsson S, Högberg C, Sartor H, et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. Lancet Oncol. Aug 2023;24(8):936-944. [doi: 10.1016/S1470-2045(23)00298-X] [Medline: 37541274]
4. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. Nov 22, 2018;24(11):1716-1720. [FREE Full text] [doi: 10.1038/s41591-018-0213-5] [Medline: 30349085]
5. Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med. Mar 09, 2020;26(3):364-373. [doi: 10.1038/s41591-020-0789-4] [Medline: 32152583]
6. Ienca M, Fabrice J, Elger B, Caon M, Scoccia Pappagallo A, Kressig RW, et al. Intelligent assistive technology for Alzheimer's disease and other dementias: a systematic review. J Alzheimers Dis. 2017;56(4):1301-1340. [doi: 10.3233/JAD-161037] [Medline: 28222516]
7. Valeriani D, Santoro F, Ienca M. The present and future of neural interfaces. Front Neurorobot. Oct 11, 2022;16:953968. [FREE Full text] [doi: 10.3389/fnbot.2022.953968] [Medline: 36304780]

8. Zeng D, Cao Z, Neill DB. Artificial intelligence–enabled public health surveillance—from local detection to global epidemic monitoring and control. In: Xing L, Giger ML, Min JK, editors. Artificial Intelligence in Medicine: Technical Basis and Clinical Applications. New York, NY. Academic Press; 2021:2021-2053.
9. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med. Aug 17, 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
10. Shani C, Zarecki J, Shahaf D. The lean data scientist: recent advances toward overcoming the data bottleneck. Commun ACM. Jan 20, 2023;66(2):92-102. [doi: 10.1145/3551635]
11. van Velsen L, Ludden G, Grünloh C. The limitations of user-and human-centered design in an eHealth context and how to move beyond them. J Med Internet Res. Oct 05, 2022;24(10):e37341. [FREE Full text] [doi: 10.2196/37341] [Medline: 36197718]
12. Gama F, Tyskbo D, Nygren J, Barlow J, Reed J, Svedberg P. Implementation frameworks for artificial intelligence translation into health care practice: scoping review. J Med Internet Res. Jan 27, 2022;24(1):e32215. [FREE Full text] [doi: 10.2196/32215] [Medline: 35084349]
13. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med. Jan 2019;25(1):30-36. [FREE Full text] [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
14. Shaw J, Rudzicz F, Jamieson T, Goldfarb A. Artificial intelligence and the implementation challenge. J Med Internet Res. Jul 10, 2019;21(7):e13659. [FREE Full text] [doi: 10.2196/13659] [Medline: 31293245]
15. Chekroud AM, Hawrilenko M, Loho H, Bondar J, Gueorguieva R, Hasan A, et al. Illusory generalizability of clinical prediction models. Science. Jan 12, 2024;383(6679):164-167. [doi: 10.1126/science.adg8538] [Medline: 38207039]
16. Crossnohere NL, Elsaid M, Paskett J, Bose-Brill S, Bridges JFP. Guidelines for artificial intelligence in medicine: literature review and content analysis of frameworks. J Med Internet Res. Aug 25, 2022;24(8):e36823. [FREE Full text] [doi: 10.2196/36823] [Medline: 36006692]
17. Maddox TM, Rumsfeld JS, Payne PR. Questions for artificial intelligence in health care. JAMA. Jan 01, 2019;321(1):31-32. [doi: 10.1001/jama.2018.18932] [Medline: 30535130]
18. Tucci V, Saary J, Doyle TE. Factors influencing trust in medical artificial intelligence for healthcare professionals: a narrative review. J Med Artif Intell. Mar 2022;5:4. [doi: 10.21037/jmai-21-25]
19. McLeod C. Stanford encyclopedia of philosophy. The Metaphysics Research Lab. URL: https://plato.stanford.edu/archives/fall2021/entries/trust/ [accessed 2024-04-29]
20. McKnight DH, Chervany NL. What is trust? A conceptual analysis and an interdisciplinary model. In: Proceedings of the 2000 Americas Conference on Information Systems. 2000. Presented at: AMCIS '00; August 10-13, 2000:382; Long Beach, CA. URL: https://aisel.aisnet.org/amcis2000/382
21. Baier A. Tanner lectures on human values. Princeton University. 1991. URL: https://uchv.princeton.edu/events/tanner-lectures-human-values [accessed 2024-04-29]
22. Luhmann N. Vertrauen : ein Mechanismus der Reduktion sozialer Komplexita¨t. Stuttgart, Germany. F. Enke; 1968.
23. Hawley K. Trust, distrust and commitment. Noûs. Oct 25, 2012;48(1):1-20. [FREE Full text] [doi: 10.1111/nous.12000]
24. Castelfranchi C, Falcone R. Trust Theory - A Socio-Cognitive and Computational Model. Hoboken, NJ. John Wiley & Sons; 2010.
25. Jacovi A, Marasović A, Miller T, Goldberg Y. Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021. Presented at: FAccT '21; March 3-10, 2021:624-635; Virtual Event. URL: https://dl.acm.org/doi/10.1145/3442188.3445923 [doi: 10.1145/3442188.3445923]
26. Benk M, Kerstan S, von Wangenheim F, Ferrario A. Twenty-four years of empirical research on trust in AI: a bibliometric review of trends, overlooked issues, and future directions. AI Soc. Oct 02, 2024:25. [FREE Full text] [doi: 10.1007/S00146-024-02059-Y]
27. Jones C, Thornton J, Wyatt JC. Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability. Med Law Rev. Nov 27, 2023;31(4):501-520. [FREE Full text] [doi: 10.1093/medlaw/fwad013] [Medline: 37218368]
28. Ferrario A, Loi M. How explainability contributes to trust in AI. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022. Presented at: FAccT '22; June 21-24, 2022:1457-1466; Seoul, Republic of Korea. URL: https://dl.acm.org/doi/10.1145/3531146.3533202 [doi: 10.1145/3531146.3533202]
29. Starke G, Ienca M. Misplaced trust and distrust: how not to engage with medical artificial intelligence. Camb Q Healthc Ethics. Oct 20, 2022;33(3):1-10. [doi: 10.1017/S0963180122000445] [Medline: 36263755]
30. Choung H, David P, Ross A. Trust in AI and its role in the acceptance of AI technologies. Int J Hum Comput Interact. Apr 20, 2022;39(9):1727-1739. [doi: 10.1080/10447318.2022.2050543]
31. Gille F, Jobin A, Ienca M. What we talk about when we talk about trust: theory of trust for AI in healthcare. Intell Based Med. Nov 2020;1-2:100001. [doi: 10.1016/j.ibmed.2020.100001]
32. Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors. 2004;46(1):50-80. [doi: 10.1518/hfes.46.1.50_30392] [Medline: 15151155]

33. Gibson GR, Hutkins R, Sanders ME, Prescott SL, Reimer RA, Salminen SJ, et al. Expert consensus document: the International Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on the definition and scope of prebiotics. Nat Rev Gastroenterol Hepatol. Aug 14, 2017;14(8):491-502. [FREE Full text] [doi: 10.1038/nrgastro.2017.75] [Medline: 28611480]
34. Giustina A, Chanson P, Kleinberg D, Bronstein MD, Clemmons DR, Klibanski A, et al. Acromegaly Consensus Group. Expert consensus document: a consensus on the medical treatment of acromegaly. Nat Rev Endocrinol. Apr 25, 2014;10(4):243-248. [doi: 10.1038/nrendo.2014.21] [Medline: 24566817]
35. Ienca M, Fins JJ, Jox RJ, Jotterand F, Voeneky S, Andorno R, et al. Towards a governance framework for brain data. Neuroethics. Jun 03, 2022;15(2):20. [doi: 10.1007/S12152-022-09498-8]
36. Ives J, Dunn M, Molewijk B, Schildmann J, Bærøe K, Frith L, et al. Standards of practice in empirical bioethics research: towards a consensus. BMC Med Ethics. Jul 10, 2018;19(1):68. [FREE Full text] [doi: 10.1186/s12910-018-0304-3] [Medline: 29986689]
37. Gattrell WT, Logullo P, van Zuuren EJ, Price A, Hughes EL, Blazey P, et al. ACCORD (ACcurate COnsensus reporting document): a reporting guideline for consensus methods in biomedicine developed via a modified Delphi. PLoS Med. Jan 23, 2024;21(1):e1004326. [FREE Full text] [doi: 10.1371/journal.pmed.1004326] [Medline: 38261576]
38. La Brooy C, Pratt B, Kelaher M. What is the role of consensus statements in a risk society? J Risk Res. Jul 25, 2019;23(5):664-677. [doi: 10.1080/13669877.2019.1628094]
39. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care. Dec 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
40. Ancker JS, Benda NC, Reddy M, Unertl KM, Veinot T. Guidance for publishing qualitative research in informatics. J Am Med Inform Assoc. Nov 25, 2021;28(12):2743-2748. [FREE Full text] [doi: 10.1093/jamia/ocab195] [Medline: 34537840]
41. Jabareen Y. Building a conceptual framework: philosophy, definitions, and procedure. Int J Qual Methods. Dec 01, 2009;8(4):49-62. [doi: 10.1177/160940690900800406]
42. U.S. Department of HealthHuman Services FDA Center for Drug EvaluationResearch, U.S. Department of HealthHuman Services FDA Center for Biologics EvaluationResearch, U.S. Department of HealthHuman Services FDA Center for DevicesRadiological Health. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health Qual Life Outcomes. Oct 11, 2006;4(1):79. [FREE Full text] [doi: 10.1186/1477-7525-4-79] [Medline: 17034633]
43. OECD guidelines on measuring trust. Organisation for Economic Co-operation and Development. URL: https://www.oecd.org/en/publications/oecd-guidelines-on-measuring-trust_9789264278219-en.html [accessed 2024-04-29]
44. Gille F, Smith S, Mays N. Why public trust in health care systems matters and deserves greater research attention. J Health Serv Res Policy. Jan 17, 2015;20(1):62-64. [doi: 10.1177/1355819614543161] [Medline: 25038059]
45. Ferrario A, Loi M, Viganò E. In AI we trust incrementally: a multi-layer model of trust to analyze human-artificial intelligence interactions. Philos Technol. Oct 23, 2019;33(3):523-539. [doi: 10.1007/S13347-019-00378-3]
46. Ferrario A, Loi M, Viganò E. Trust does not need to be human: it is possible to trust medical AI. J Med Ethics. Nov 25, 2020;47(6):437-438. [FREE Full text] [doi: 10.1136/medethics-2020-106922] [Medline: 33239471]
47. Gille F. What Is Public Trust in the Health System?: Insights into Health Data Use. Bristool, UK. Policy Press; 2023.
48. Durán JM, Formanek N. Grounds for trust: essential epistemic opacity and computational reliabilism. Minds Mach. Oct 29, 2018;28(4):645-666. [doi: 10.1007/S11023-018-9481-6]
49. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics. Mar 18, 2021:22. [doi: 10.1136/medethics-2020-106820] [Medline: 33737318]
50. Loi M, Ferrario A, Viganò E. How much do you trust me? A logico-mathematical analysis of the concept of the intensity of trust. Synthese. May 23, 2023;201(6):186. [doi: 10.1007/S11229-023-04169-4]
51. Chong D, Druckman JN. Framing theory. Annu Rev Polit Sci. Jun 01, 2007;10(1):103-126. [doi: 10.1146/annurev.polisci.10.072805.103054]
52. Starke G, van den Brule R, Elger BS, Haselager P. Intentional machines: a defence of trust in medical artificial intelligence. Bioethics. Feb 18, 2022;36(2):154-161. [doi: 10.1111/bioe.12891] [Medline: 34142373]
53. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. Nat Rev Cancer. Aug 2018;18(8):500-510. [FREE Full text] [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
54. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. Dec 2017;2(4):230-243. [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
55. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur Radiol Exp. Oct 24, 2018;2(1):35. [FREE Full text] [doi: 10.1186/s41747-018-0061-6] [Medline: 30353365]
56. McDonald ES, McCarthy AM, Akhtar AL, Synnestvedt MB, Schnall M, Conant EF. Baseline screening mammography: performance of full-field digital mammography versus digital breast tomosynthesis. AJR Am J Roentgenol. Nov 2015;205(5):1143-1148. [doi: 10.2214/AJR.15.14406] [Medline: 26496565]

57. Tang A, Tam R, Cadrin-Chênevert A, Guest W, Chong J, Barfett J, et al. Canadian Association of Radiologists (CAR) Artificial Intelligence Working Group. Canadian association of radiologists white paper on artificial intelligence in radiology. Can Assoc Radiol J. May 2018;69(2):120-135. [FREE Full text] [doi: 10.1016/j.carj.2018.02.002] [Medline: 29655580]
58. Bejnordi BE, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, the CAMELYON16 Consortium, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. Dec 12, 2017;318(22):2199-2210. [FREE Full text] [doi: 10.1001/jama.2017.14585] [Medline: 29234806]
59. Ghafoorian M, Karssemeijer N, Heskes T, Bergkamp M, Wissink J, Obels J, et al. Deep multi-scale location-aware 3D convolutional neural networks for automated detection of lacunes of presumed vascular origin. Neuroimage Clin. 2017;14:391-399. [FREE Full text] [doi: 10.1016/j.nicl.2017.01.033] [Medline: 28271039]
60. Oxipit awarded CE mark for first autonomous AI medical imaging application. Oxipit.ai. URL: https://oxipit.ai/news/first-autonomous-ai-medical-imaging-application/ [accessed 2024-04-29]
61. Keski-Filppula T, Nikki M, Haapea M, Ramanauskas N, Tervonen O. Using artificial intelligence to detect chest X-rays with no significant findings in a primary health care setting in Oulu, Finland. arXiv. Preprint posted online May 17, 2022. [FREE Full text] [doi: 10.48550/arXiv.2205.08123]
62. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. Mar 2020;46(3):205-211. [FREE Full text] [doi: 10.1136/medethics-2019-105586] [Medline: 31748206]
63. Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. Nat Biomed Eng. Dec 11, 2022;6(12):1346-1352. [doi: 10.1038/s41551-022-00914-1] [Medline: 35953649]
64. Tizhoosh HR, Pantanowitz L. Artificial intelligence and digital pathology: challenges and opportunities. J Pathol Inform. 2018;9:38. [FREE Full text] [doi: 10.4103/jpi.jpi_53_18] [Medline: 30607305]
65. Pena GP, Andrade-Filho JS. How does a pathologist make a diagnosis? Arch Pathol Lab Med. Jan 2009;133(1):124-132. [FREE Full text] [doi: 10.5858/133.1.124] [Medline: 19123724]
66. Arbelaez Ossa L, Starke G, Lorenzini G, Vogt JE, Shaw DM, Elger BS. Re-focusing explainability in medicine. Digit Health. Feb 11, 2022;8:20552076221074488. [FREE Full text] [doi: 10.1177/20552076221074488] [Medline: 35173981]
67. Starke G, Elger BS, De Clercq E. Machine learning and its impact on psychiatric nosology: findings from a qualitative study among German and Swiss experts. PhiMiSci. Apr 11, 2023;4:1-17. [doi: 10.33735/phimisci.2023.9435]
68. Walker MJ, Rogers WA. Diagnosis, narrative identity, and asymptomatic disease. Theor Med Bioeth. Aug 5, 2017;38(4):307-321. [doi: 10.1007/s11017-017-9412-1] [Medline: 28681328]
69. Drogt J, Milota M, Vos S, Bredenoord A, Jongsma K. Integrating artificial intelligence in pathology: a qualitative interview study of users' experiences and expectations. Mod Pathol. Nov 2022;35(11):1540-1550. [FREE Full text] [doi: 10.1038/s41379-022-01123-6] [Medline: 35927490]
70. Starke G, Schmidt B, De Clercq E, Elger BS. Explainability as fig leaf? An exploration of experts' ethical expectations towards machine learning in psychiatry. AI Ethics. Jun 07, 2022;3(1):303-314. [doi: 10.1007/S43681-022-00177-1]
71. Yang L, Ene IC, Arabi Belaghi R, Koff D, Stein N, Santaguida P. Stakeholders' perspectives on the future of artificial intelligence in radiology: a scoping review. Eur Radiol. Mar 21, 2022;32(3):1477-1495. [doi: 10.1007/s00330-021-08214-z] [Medline: 34545445]
72. Aquino YS, Rogers WA, Braunack-Mayer A, Frazer H, Win KT, Houssami N, et al. Utopia versus dystopia: professional perspectives on the impact of healthcare artificial intelligence on clinical roles and skills. Int J Med Inform. Jan 2023;169:104903. [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104903] [Medline: 36343512]
73. Patel D, Kher V, Desai B, Lei X, Cen S, Nanda N, et al. Machine learning based predictors for COVID-19 disease severity. Sci Rep. Feb 25, 2021;11(1):4673. [FREE Full text] [doi: 10.1038/s41598-021-83967-7] [Medline: 33633145]
74. Jauk S, Kramer D, Avian A, Berghold A, Leodolter W, Schulz S. Correction to: technology acceptance of a machine learning algorithm predicting delirium in a clinical setting: a mixed-methods study. J Med Syst. Mar 10, 2021;45(4):52. [FREE Full text] [doi: 10.1007/s10916-021-01728-5] [Medline: 33740133]
75. van Royen FS, Moons KG, Geersing GJ, van Smeden M. Developing, validating, updating and judging the impact of prognostic models for respiratory diseases. Eur Respir J. Jun 21, 2022;60(3):2200250. [doi: 10.1183/13993003.00250-2022]
76. Mikhael PG, Wohlwend J, Yala A, Karstens L, Xiang J, Takigami AK, et al. Sybil: a validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. J Clin Oncol. Apr 20, 2023;41(12):2191-2200. [FREE Full text] [doi: 10.1200/JCO.22.01345] [Medline: 36634294]
77. Marinovich ML, Wylie E, Lotter W, Pearce A, Carter SM, Lund H, et al. Artificial intelligence (AI) to enhance breast cancer screening: protocol for population-based cohort study of cancer detection. BMJ Open. Jan 03, 2022;12(1):e054005. [FREE Full text] [doi: 10.1136/bmjopen-2021-054005] [Medline: 34980622]
78. King OC, Mertens M. Self-fulfilling prophecy in practical and automated prediction. Ethic Theory Moral Prac. Jan 16, 2023;26(1):127-152. [doi: 10.1007/S10677-022-10359-9]
79. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing transparency through a multiverse analysis. Perspect Psychol Sci. Sep 29, 2016;11(5):702-712. [doi: 10.1177/1745691616658637] [Medline: 27694465]
80. Marx CT, Calmon FD, Ustun B. Predictive multiplicity in classification. In: Proceedings of the 37 th International Conference on Machine Learning. 2020. Presented at: PMLR '20; July 12-18, 2020:1-10; Vienna, Austria. URL: https://proceedings.mlr.press/v119/marx20a.html

81. Kulynych B, Hsu H, Troncoso C, Calmon FP. Arbitrary decisions are a hidden cost of differentially private training. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023. Presented at: FAccT '23; June 12-15, 2023:1609-1623; Chicago, IL. URL: https://dl.acm.org/doi/10.1145/3593013.3594103 [doi: 10.1145/3593013.3594103]
82. Kim A, Chung KC, Keir C, Patrick DL. Patient-reported outcomes associated with cancer screening: a systematic review. BMC Cancer. Mar 01, 2022;22(1):223. [FREE Full text] [doi: 10.1186/s12885-022-09261-5] [Medline: 35232405]
83. Dickinson JA, Pimlott N, Grad R, Singh H, Szafran O, Wilson BJ, et al. Screening: when things go wrong. Can Fam Physician. Jul 2018;64(7):502-508. [FREE Full text] [Medline: 30002025]
84. Peng Y, Liu E, Peng S, Chen Q, Li D, Lian D. Using artificial intelligence technology to fight COVID-19: a review. Artif Intell Rev. Jan 03, 2022;55(6):4941-4977. [FREE Full text] [doi: 10.1007/s10462-021-10106-z] [Medline: 35002010]
85. MacIntyre CR, Lim S, Quigley A. Preventing the next pandemic: use of artificial intelligence for epidemic monitoring and alerts. Cell Rep Med. Dec 20, 2022;3(12):100867. [FREE Full text] [doi: 10.1016/j.xcrm.2022.100867] [Medline: 36543103]
86. Bhargavi BS, Moa A. Global outbreaks of zika infection by epidemic observatory (EpiWATCH), 2016-2019. Glob. Biosecurity. Oct 21, 2020;2:55. [doi: 10.31646/gbio.83]
87. Lesmanawati DAS, Adam DC, Hooshmand E, Moa A, Kunasekaran MP, MacIntyre CR. The global epidemiology of hepatitis a outbreaks 2016-2018 and the utility of EpiWATCH as a rapid epidemic intelligence service. Glob Biosecurity. Mar 23, 2021;3(1):25. [doi: 10.31646/gbio.100]
88. Kuhner D, Fiederer LD, Aldinger J, Burget F, Völker M, Schirrmeister RT, et al. A service assistant combining autonomous robotics, flexible goal formulation, and deep-learning-based brain–computer interfacing. Robot Auton Syst. Jun 2019;116:98-113. [doi: 10.1016/j.robot.2019.02.015]
89. ISO 8549-3:2020: prosthetics and orthotics — vocabulary: part 3: terms relating to orthoses. International Organization for Standardization. 2022. URL: https://www.iso.org/standard/79497.html [accessed 2024-04-29]
90. Baniqued PD, Stanyer EC, Awais M, Alazmani A, Jackson AE, Mon-Williams MA, et al. Brain-computer interface robotics for hand rehabilitation after stroke: a systematic review. J Neuroeng Rehabil. Jan 23, 2021;18(1):15. [FREE Full text] [doi: 10.1186/s12984-021-00820-8] [Medline: 33485365]
91. Kellmeyer P, Mueller O, Feingold-Polak R, Levy-Tzedek S. Social robots in rehabilitation: a question of trust. Sci Robot. Aug 15, 2018;3(21):eaat1587. [doi: 10.1126/scirobotics.aat1587] [Medline: 33141717]
92. Carmena JM. Brain versus machine control. PLoS Biol. Dec 14, 2004;2(12):e430. [doi: 10.1371/journal.pbio.0020430]
93. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. Oct 25, 2019;366(6464):447-453. [FREE Full text] [doi: 10.1126/science.aax2342] [Medline: 31649194]
94. The Lancet Digital Health. There is no such thing as race in health-care algorithms. Lancet Digit Health. Dec 2019;1(8):e375. [FREE Full text] [doi: 10.1016/S2589-7500(19)30201-8] [Medline: 33323212]
95. Jones K. Trust as an Affective Attitude. Ethics. Oct 1996;107(1):4-25. [doi: 10.1086/233694]
96. Sheehan M, Friesen P, Balmer A, Cheeks C, Davidson S, Devereux J, et al. Trust, trustworthiness and sharing patient data for research. J Med Ethics. May 18, 2020;47(12):e26. [doi: 10.1136/medethics-2019-106048] [Medline: 32424061]
97. McGeer V, Pettit P. The empowering theory of trust. In: Faulkner P, Simpson T, editors. The Philosophy of Trust. Oxford, UK. Oxford Academic Press; 2017:14-34.
98. Hatherley JJ. Limits of trust in medical AI. J Med Ethics. Jul 2020;46(7):478-481. [doi: 10.1136/medethics-2019-105935] [Medline: 32220870]
99. Wadden JJ. What kind of artificial intelligence should we want for use in healthcare decision-making applications? Can J Bioeth. May 27, 2021;4(1):94-100. [doi: 10.7202/1077636ar]
100. Ratti E, Graves M. Explainable machine learning practices: opening another black box for reliable medical AI. AI Ethics. Feb 15, 2022;2(4):801-814. [doi: 10.1007/S43681-022-00141-z]
101. McLennan S, Fiske A, Tigard D, Müller R, Haddadin S, Buyx A. Embedded ethics: a proposal for integrating ethics into the development of medical AI. BMC Med Ethics. Jan 26, 2022;23(1):6. [FREE Full text] [doi: 10.1186/s12910-022-00746-3] [Medline: 35081955]
102. Algorithmic impact assessment: a case study in healthcare. Ada Lovelace Instititue. URL: https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/ [accessed 2024-04-29]
103. Markus AF, Kors JA, Rijnbeek PR. 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 Inform. Jan 2021;113:103655. [FREE Full text] [doi: 10.1016/j.jbi.2020.103655] [Medline: 33309898]
104. Gerke S. Health AI for good rather than evil? The need for a new regulatory framework for AI-based medical devices. Yale J Health Policy Law Ethics. 2012;20(2):432. [FREE Full text]
105. Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ. Mar 12, 2019;364:l886. [FREE Full text] [doi: 10.1136/bmj.l886] [Medline: 30862612]
106. Möllering G. The nature of trust: from Georg Simmel to a theory of expectation, interpretation and suspension. Sociology. 2001;35(2):403-420. [doi: 10.1017/S0038038501000190]

107. Simmel G. Soziologie: Untersuchungen über die Formen der Vergesellschaftung. Berlin, Germany. Duncker und Humblot; 1908.
108. Nickel PJ. Trust in medical artificial intelligence: a discretionary account. Ethics Inf Technol. Jan 24, 2022;24(1):e56. [doi: 10.1007/S10676-022-09630-5]
109. Ethics guidelines for trustworthy AI. European Commission: Directorate-General for Communications Networks, Content and Technology. URL: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1/language-en [accessed 2024-04-29]
110. Braun M, Bleher H, Hummel P. A leap of faith: is there a formula for "trustworthy" AI? Hastings Cent Rep. May 19, 2021;51(3):17-22. [doi: 10.1002/hast.1207] [Medline: 33606288]
111. O'Neill O. Linking trust to trustworthiness. Int J Philos Stud. Apr 25, 2018;26(2):293-300. [doi: 10.1080/09672559.2018.1454637]
112. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. Nov 30, 2020;20(1):310. [FREE Full text] [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
113. Mincu D, Roy S. Developing robust benchmarks for driving forward AI innovation in healthcare. Nat Mach Intell. Nov 15, 2022;4(11):916-921. [doi: 10.1038/S42256-022-00559-4]
114. Strickland E. IBM Watson, heal thyself: how IBM overpromised and underdelivered on AI health care. IEEE Spectr. Apr 2019;56(4):24-31. [doi: 10.1109/mspec.2019.8678513]
115. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. JAMA. Dec 24, 2019;322(24):2377-2378. [FREE Full text] [doi: 10.1001/jama.2019.18058] [Medline: 31755905]
116. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. PLoS Med. Nov 2018;15(11):e1002689. [FREE Full text] [doi: 10.1371/journal.pmed.1002689] [Medline: 30399149]
117. Char DS, Abràmoff MD, Feudtner C. Identifying ethical considerations for machine learning healthcare applications. Am J Bioeth. Nov 2020;20(11):7-17. [FREE Full text] [doi: 10.1080/15265161.2020.1819469] [Medline: 33103967]
118. Powell J. Trust me, I'm a chatbot: how artificial intelligence in health care fails the Turing test. J Med Internet Res. Oct 28, 2019;21(10):e16222. [FREE Full text] [doi: 10.2196/16222] [Medline: 31661083]
119. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. May 13, 2019;1(5):206-215. [FREE Full text] [doi: 10.1038/s42256-019-0048-x] [Medline: 35603010]
120. Tschantz MC. What is proxy discrimination? In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022. Presented at: FAccT '22; June 21-24, 2022:1993; Seoul, Republic of Korea. URL: https://dl.acm.org/doi/fullHtml/10.1145/3531146.3533242 [doi: 10.1145/3531146.3533242]
121. Suleman M, Qureshi Z. Should medicine be colour blind? J Med Ethics. Nov 23, 2023;49(11):725-726. [doi: 10.1136/jme-2023-109634] [Medline: 37871944]
122. Starke G. The emperor's new clothes? Transparency and trust in machine learning for clinical neuroscience. In: Friedrich O, Wolkenstein A, Bublitz C, Jox RF, Racine E, editors. Clinical Neurotechnology meets Artificial Intelligence: Philosophical, Ethical, Legal and Social Implications. Cham, Switzerland. Springer; 2021:2021-2096.
123. Reddy S. Navigating the AI revolution: the case for precise regulation in health care. J Med Internet Res. Sep 11, 2023;25:e49989. [FREE Full text] [doi: 10.2196/49989] [Medline: 37695650]
124. Scharowski N, Benk M, Kühne SJ, Wettstein L, Brühlmann F. Certification labels for trustworthy ai: Insights from an empirical mixed-method study. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023. Presented at: FAccT '23; June 12-15, 2023:248-260; Chicago, IL. URL: https://dl.acm.org/doi/10.1145/3593013.3593994 [doi: 10.1145/3593013.3593994]
125. Lee MK, Rich K. Who is included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021. Presented at: CHI '21; May 8-13, 2021:1-14; Yokohama, Japan. URL: https://dl.acm.org/doi/10.1145/3411764.3445570 [doi: 10.1145/3411764.3445570]
126. Ryan M. In AI we trust: ethics, artificial intelligence, and reliability. Sci Eng Ethics. Oct 10, 2020;26(5):2749-2767. [FREE Full text] [doi: 10.1007/s11948-020-00228-y] [Medline: 32524425]
127. Metzinger T. Ethics washing made in Europe. Der Tagesspiegel. URL: https://www.tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html [accessed 2024-04-29]
128. PytlikZillig LM, Kimbrough CD. Consensus on conceptualizations and definitions of trust: are we there yet? In: Shockley E, Neal TM, PytlikZillig LM, Bornstein BH, editors. Interdisciplinary Perspectives on Trust: Towards Theoretical and Methodological Integration. Cham, Switzerland. Springer; 2016:17-47.
129. Kyung N, Kwon HE. Rationally trust, but emotionally? The roles of cognitive and affective trust in laypeople's acceptance of AI for preventive care operations. Prod Oper Manag. Jul 31, 2022:13785. [FREE Full text] [doi: 10.1111/poms.13785]

Abbreviations

AI: artificial intelligence

BCI: brain-computer interface
ICU: intensive care unit

Edited by N Cahill, T Leung; submitted 12.01.24; peer-reviewed by H Burkhardt, P Nickel, L Weinert; comments to author 29.04.24; revised version received 31.07.24; accepted 28.11.24; published 19.02.25

Please cite as:
Starke G, Gille F, Termine A, Aquino YSJ, Chavarriaga R, Ferrario A, Hastings J, Jongsma K, Kellmeyer P, Kulynych B, Postan E, Racine E, Sahin D, Tomaszewska P, Vold K, Webb J, Facchini A, Ienca M
Finding Consensus on Trust in AI in Health Care: Recommendations From a Panel of International Experts
J Med Internet Res 2025;27:e56306
URL: https://www.jmir.org/2025/1/e56306
doi: 10.2196/56306
PMID: 39969962