Viewpoint

Trust, Trustworthiness, and the Future of Medical AI: Outcomes of an Interdisciplinary Expert Workshop

Melanie Goisauf[1*], PhD; Mónica Cano Abadía[1*], PhD; Kaya Akyüz[1*], PhD; Maciej Bobowicz[2*], Dr med; Alena Buyx[3*], Prof Dr; Ilaria Colussi[1*], PhD; Marie-Christine Fritzsche[3,4*]; Karim Lekadir[5*], Prof Dr; Pekka Marttinen[6*], PhD; Michaela Th Mayrhofer[1*], PhD; Janos Meszaros[7*], PhD

[1]Department of ELSI Services and Research, Biobanking and Biomolecular Resources Research Infrastructure Consortium, Graz, Austria

[2]2nd Department of Radiology, Gdańsk Medical University, Gdansk, Poland

[3]Institute of History and Ethics in Medicine, TUM School of Medicine and Health, Technical University of Munich, Munich, Germany

[4]Department of Science, Technology and Society (STS), School of Social Sciences and Technology, Technical University of Munich, Munich, Germany

[5]Institució Catalana de Recerca i Estudis Avançats (ICREA), Department of Mathematics and Computer Science, Universitat de Barcelona, Barcelona, Spain

[6]Department of Computer Science, Aalto University, Espoo, Finland

[7]Division of Clinical Pharmacology and Pharmacotherapy, Department of Pharmaceutical and Pharmacological Sciences, KU Leuven, Leuven, Belgium

[*]all authors contributed equally

Corresponding Author:
Melanie Goisauf, PhD
Department of ELSI Services and Research
Biobanking and Biomolecular Resources Research Infrastructure Consortium
Neue Stiftingtalstrasse 2/B/6
Graz, 8010
Austria
Phone: 43 664 88 72 18 73
Email: melanie.goisauf@bbmri-eric.eu

Abstract

Trustworthiness has become a key concept for the ethical development and application of artificial intelligence (AI) in medicine. Various guidelines have formulated key principles, such as fairness, robustness, and explainability, as essential components to achieve trustworthy AI. However, conceptualizations of trustworthy AI often emphasize technical requirements and computational solutions, frequently overlooking broader aspects of fairness and potential biases. These include not only algorithmic bias but also human, institutional, social, and societal factors, which are critical to foster AI systems that are both ethically sound and socially responsible. This viewpoint article presents an interdisciplinary approach to analyzing trust in AI and trustworthy AI within the medical context, focusing on (1) social sciences and humanities conceptualizations and legal perspectives on trust and (2) their implications for trustworthy AI in health care. It focuses on real-world challenges in medicine that are often underrepresented in theoretical discussions to propose a more practice-oriented understanding. Insights were gathered from an interdisciplinary workshop with experts from various disciplines involved in the development and application of medical AI, particularly in oncological imaging and genomics, complemented by theoretical approaches related to trust in AI. Results emphasize that, beyond common issues of bias and fairness, knowledge and human involvement are essential for trustworthy AI. Stakeholder engagement throughout the AI life cycle emerged as crucial, supporting a human- and multicentered framework for trustworthy AI implementation. Findings emphasize that trust in medical AI depends on providing meaningful, user-oriented information and balancing knowledge with acceptable uncertainty. Experts highlighted the importance of confidence in the tool's functionality, specifically that it performs as expected. Trustworthiness was shown to be not a feature but rather a relational process, involving humans, their expertise, and the broader social or institutional contexts in which AI tools operate. Trust is dynamic, shaped by interactions among individuals, technologies, and institutions, and ultimately centers on people rather than tools alone. Tools are evaluated based on reliability and credibility, yet trust fundamentally relies on human connections. The article underscores the development of AI tools that are not only technically sound but also ethically robust and broadly accepted by end users, contributing to more effective and equitable AI-mediated health care. Findings highlight that building AI trustworthiness in health care requires a human-centered, multistakeholder approach with diverse and inclusive engagement. To promote equity, we recommend that

AI development teams involve all relevant stakeholders at every stage of the AI lifecycle—from conception, technical development, clinical validation, and real-world deployment.

(J Med Internet Res 2025;27:e71236) doi: 10.2196/71236

KEYWORDS

artificial intelligence; trust; trustworthy AI; medicine; ethics of AI; interdisciplinarity; stakeholder engagement; human-centered AI

Introduction

Trustworthiness has become a key requirement in the development and application of ethical artificial intelligence (AI). This is highlighted by the European Commission High-Level Expert Group on AI (AI HLEG) Guidelines for Trustworthy AI [1] and emphasized by the World Health Organization (WHO) [2] for effective AI integration in health care. In addition, cross-cutting initiatives such as FUTURE-AI (Fairness, Universality, Traceability, Usability, Robustness, and Explainability–artificial intelligence) aim to guide AI developments toward trustworthiness, focusing on principles of fairness, universality, traceability, usability, robustness, and explainability [3]. The discourse on trustworthy AI often focuses on defining the conditions under which it can be achieved. Scholars are developing practical approaches for realizing trustworthy AI through guidelines or regulations [4], and some institutions even offer certification possibilities (eg, IEEE CertifAIEd). Such efforts neglect that ethical AI cannot be guaranteed solely through following principles [5,6], or that trustworthy AI cannot only be defined in terms of technical conditions and solutions [7], such as fairness objectives [8]. Indeed, some authors have challenged the idea of achieving trustworthy AI by merely meeting technical criteria and have argued that trusting in AI is based on more human than technical aspects [9,10]. Besides, guidelines on the implementation of “trustworthy AI” in medicine lack consensus on what defines a user’s trust in AI [11].

The significance of the discourses around trust in AI and the trustworthiness of AI within the scientific discussion on ethical AI is rarely reflected in its conceptualization [12,13], which is crucial as AI “can increase systemic risks of harm, raise the possibility of errors with severe consequences, and amplify complex ethical and societal issues” [14]. In scientific literature, the concepts of “trust” and “trustworthiness” are often used interchangeably, despite repeated efforts to define them [15,16]. A more thorough analysis is needed on the situatedness of trust in medical AI and the trustworthiness of medical AI, given that trust plays a critical role in situations of risk, vulnerability, and uncertainty, circumstances frequently encountered in the medical context. In this context, trust needs to be understood as a complex relational concept that involves several trustor-trustee relationships, such as trust in technology, institutions, and persons, for example, scientists who trust each other, patients who trust scientists, and health care professionals [17,18].

Trustworthiness of medical AI also depends on how certain challenges are being managed: for instance, biases in the training data that lead to biases in algorithms, lack of transparency or explainability in how an AI system decides, securing sensitive data such as medical data, as well as ensuring human oversight and continuous monitoring [19]. Against the backdrop of evidence showing that algorithms employed in health care can encode, reinforce, and exacerbate existing inequalities within the health care system [20], which poses a particular risk to vulnerable patients [21], identifying and mitigating biases, especially racial and gender biases [22], is key for trust [23,24]. Failure to do so could erode public trust in health systems and hinder the adoption of AI systems in health care [25]. Consequently, addressing bias must go beyond the prevailing focus on computational factors and fairness of machine learning algorithms and must take all forms of potential bias into account, including human, institutional, and societal factors [26].

This viewpoint article presents an interdisciplinary approach to analyzing trust in AI and trustworthy AI within the medical context and proposes a practice-oriented conceptualization. Adopting a multidisciplinary understanding of trust and trustworthiness, it builds on the outcomes of a workshop that brought together experts from various disciplines involved in the development and application of medical AI, particularly in oncological imaging and genomics. The aim was to address analytical gaps and explore the concepts of trust and trustworthiness. This interdisciplinary and contextualized standpoint informs the analysis of trust and trustworthiness in medical AI from knowledge production to technology development and use. In doing so, the article puts an emphasis on the social and legal conceptualizations of trust and their implications for trustworthy AI in medicine.

Trust in Humanities and Social Sciences

While there is a plethora of literature on trust and trustworthiness of AI [15,16,27] and in medicine in particular [11,28-30], a deeper conceptualization of the terms used is needed for specific practices in medicine and health care.

Across different philosophical approaches (eg, virtue ethics or deontological ethics), trust is considered interpersonal and relational, involving risk and moral responsiveness to patients’ vulnerability. Carter [31] notes that illness induces feelings of vulnerability in patients and caregivers, highlighting the need for a mature exploration of trust, distinct from related concepts like reliance, confidence, or faith. Baier [32] defines trust as a mix of reliance, confidence, and dependence. Reliance involves trusting another’s competence, while trust involves depending on their goodwill. Baier asserts that everyone is interdependent, relying on others for care. This aligns with more recent literature, which emphasizes our constitutive vulnerability and interdependence [33-36]. Baier [32] also highlights the role of power relations, stressing that morality requires trust and consideration of potential exploitation in unequal power dynamics.

Sociological perspectives on trust, particularly those of Giddens and Luhmann [37], explore the dynamics of institutional and interpersonal trust. Institutional trust, such as in the health care system, is influenced by trust in its representatives, like doctors. Giddens emphasizes the role of the health care professional's appearance and professionalism in shaping patient expectations and legitimizing the medical system. He asserts that trust in "flesh-and-blood" representatives informs trust in the system and is necessary to manage partial understanding due to uncertainty. Trust is not needed in situations with complete knowledge. Luhmann, however, views trust as a medium that reduces social complexity, enabling interactions within and by the system. He distinguishes between trust and confidence: trust involves past experiences and perceived risks, whereas confidence relies solely on expectation without considering alternatives [37].

Trust has been a central topic in the social sciences' examination of science and knowledge production, emphasizing the interplay between individuals and social institutions. Understanding trust requires recognizing its foundation in social relations and exploring its interactive nature, including the role of technology and the importance of tacit, embodied, and situated knowledge [38,39]. These elements incorporate social, practical, human, and bodily factors into the comprehension of knowledge production and mutual trust.

Science and technology studies (STS) conceptualize trust in relational terms, particularly focusing on the publics of science and technology. A significant shift has been the critique of the "deficit model," which assumes public mistrust stems from a lack of understanding [40,41]. This critique has led to a transformation in building trust through communicative practices, moving from one-directional communication to models of engagement that recognize the diversity of publics. Despite this, critics note the persistent reemergence of the deficit model even in newer efforts [42]. Concerns about emerging technologies' risks and uncertainties highlight the need to reflect on modernity's inherent crises and their relation to policymaking, accountability, transparency, and expertise [43]. STS scholars emphasize the importance of considering various forms of expertise, including lay expertise [41,44,45]. Cases of public acceptance or rejection of technologies like nuclear energy, nanotechnology, and genetically modified organisms provide insights into contemporary AI developments [46,47]. One-way communication does not automatically increase public trust; technologies often carry risks not immediately apparent to technologists, necessitating reflexivity [48] and a global perspective [49]. Hence, building trust would require transparent, adaptable governance systems and strong trust relations between individuals and institutions, respecting the variety of expertise.

In examining the intersection of trust and AI within medicine, it may be relevant to translate insights into trust offered by the humanities and social sciences into this emerging field. Table 1 presents a synthesis of classical approaches to trust outlined above, highlighting their relevance and application of AI in health care.

Table 1. Translation of selected classical approaches to trust in the humanities and social sciences into the field of AI[a] and medicine.

Approach	Translation
Trust as a relational and moral concept (Carter [31], Baier [32])	Trust in AI is viewed as a relationship in which ethical behavior, transparency, and accountability are essential. Trust is not only about the system's performance but also about the intentions and actions of those who design and manage AI systems.
Trust involving risk and moral responsiveness (Pellegrino & Thomasma [50], Emanuel & Emanuel [51])	Trust in AI involves stakeholders assuming risks, such as data privacy risks or misdiagnoses, with the expectation that AI systems will function ethically, respecting user autonomy and promoting well-being.
Foundational trust in doctor-patient relationships (Kittay & Meyers [52])	In AI-enhanced health care, trust is crucial for integrating AI tools into patient care, affecting how patients perceive and cooperate with AI-driven diagnostics and treatment plans.
Institutional trust (Giddens, Luhmann in Meyer et al [37])	Trust in AI within institutions depends on the trustworthiness of those who deploy and manage AI systems, influencing public and professional trust in the technology's utility and safety.
Dynamic trust shaped by social relations, technology, and experience (Collins [39], Haraway [38], Meyer et al [37])	Trust in AI is dynamic, shaped by ongoing interactions between users, developers, and AI systems. This includes how technology adapts to social expectations and how it is embedded in social practices.

[a]AI: artificial intelligence.

Legal Perspectives on Trust and Trustworthiness

In legal studies, the concept of trust is often overlooked [53,54] except for "trust law" in common law systems, which involves a fiduciary relationship where a "trustor" transfers property or rights to a trustee [55]. "Good faith" and "due diligence" embody elements of trust, implying confidence in a person or entity. For example, in contract law, good faith is essential, presuming honest and fair dealings between parties. Thus, trust is implicitly part of good faith. When trust is violated, liability rules restore balance [56,57].

Some authors argue that the law substitutes for trust [58] or emerges where trust is lacking [59]. Law's coercive and controlling nature can crowd out trust, especially "personal trust" in individuals or firms, which could mean that legislation cannot effectively promote trust. Instead, the law can define and establish rules for trustworthy behaviors. Greco [60] challenges the idea that law and trust are incompatible, arguing that law can play a constructive role in fostering trust and serving as a bridge between citizens and institutions, thereby generating and sustaining social trust. In other words, the law not only imposes rules but also helps people feel safe, respected, and thus more inclined to trust the system in which they live. From a legal philosophy perspective, trust underpins the law's existence: human relationships, which the law regulates, rely on trust. Sanctions rebuild violated trust, and rules emerge from trust, with the expectation of compliance. Thus, trust is foundational to the law.

Like trust, trustworthiness is not inherently a legal concept. In the context of AI, the European Union (EU) has emphasized trustworthiness, specifically through the concept of "trustworthy AI" rather than focusing on "trust" itself. The Ethics Guidelines for Trustworthy AI by the AI HLEG [1] define AI as being "lawful, ethical, and robust." Smuha et al [61] indicate three pillars of "Legally Trustworthy AI": (1) responsibility allocation, where the regulation appropriately assigns accountability for the harms and wrongs resulting from AI systems; (2) a consistent legal framework, where the regulation establishes and maintains a unified legal structure accompanied by effective and legitimate enforcement mechanisms to secure and uphold the rule of law; and (3) democratic deliberation, where the regulation places democratic discussion at its center, ensuring public participation and other information rights. These pillars were crucial when the EU AI Act was drafted, aiming to establish a robust legal framework to foster the development of secure, trustworthy, and ethical AI. The EU AI Act, which entered into force on August 1, 2024, establishes a comprehensive regulatory framework for AI within the EU [62]. The Act highlights that "in the health sector where the stakes for life and health are particularly high, increasingly sophisticated diagnostics systems and systems supporting human decisions should be reliable and accurate" (Recital 47). The AI Act takes a human-centric approach and recalls the guidelines for trustworthy AI by the AI HLEG (see Recital 27), namely: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, nondiscrimination and fairness, societal and environmental well-being, and accountability. It states that these principles should be applied in the design and use of AI models and should serve as a basis for the drafting of codes of conduct under the Regulation.

The AI Act mentions "trust" or "trustworthiness" 29 times, underscoring the significance of building trust in AI technologies. The health sector is significantly impacted by the EU AI Act, particularly due to the inclusion of AI systems. Moreover, the AI Act states that AI systems classified as medical devices under Medical Device Regulation (MDR) [63] or In Vitro Device Regulation [64]—and thus those AI systems used as medical devices—must comply with both the AI Act and existing medical device regulations. There are exemptions only for AI systems used exclusively for scientific research and premarket product development. In light of the AI Act, therefore, it seems that most commercial AI-enabled medical devices on the market used in radiology, being classified as Class IIa and up, under the MDR [65], are classified as high risk under the AI Act [66].

These developments and the prominence that "trust" and "trustworthiness" have achieved in law and policy, while being scarcely defined, highlight the need for a comprehensive analysis, particularly from an interdisciplinary and practical perspective. Toward this end, this paper advances the discussion by examining how the EU AI Act gives legal meaning to the otherwise vague concept of "trustworthy AI." It highlights how trust is not only a technical or ethical aspiration, but a legal objective shaped by rules on accountability, transparency, and human oversight, especially in health care settings.

Synthesizing Interdisciplinary Knowledge

To make better sense of issues around ethical AI as well as understandings and conditions of trust and trustworthy AI in medicine, we organized a 2-day workshop with 14 academic experts in data science, law, medicine, ethics, social science, and philosophy, which took place on May 5-6, 2022, in Berlin, Germany. The workshop participants were researchers representing four large European projects (EuCanImage, INTERVENE, Bigpicture, and BIOMAP) and other experts who contributed perspectives from philosophical and policy research. The focus was on how to implement the ethics of AI, drawing from experiences in oncological imaging and genomics. The workshop combined presentations and discussions on ethical, legal, societal, technical, and clinical aspects of medical AI.

Inspired by focus group and stakeholder engagement methods [67-69], the workshop synthesized interdisciplinary inputs on explainability and trustworthiness of AI solutions, clinical applications in cancer imaging, polygenic risk score generation and use, biases in training and calibration of AI models, and algorithmic impact assessment, as well as moderated discussions between the experts. The workshop discussions were recorded based on the informed consent of all participants, and key discussion points were put on digital sticky notes and related to each other on a digital mural during the workshop.

To deepen these discussions, the workshop participants were invited to partake in an online questionnaire consisting of five open-ended questions to share further thoughts and more specific insights into their field of expertise. Since trust and trustworthiness emerged as a crucial topic in the workshop discussions, the questions were on field-specific definitions of trust in the context of AI, criteria of trustworthy AI applications, for trusting relationships involving AI, and for trustworthiness as a valuable principle.

Aspects of trustworthy AI in medicine were also guiding the analysis of the workshop discussions and questionnaire responses. The analysis team (MG, KA, MCA) interpreted the workshop discussions and questionnaire answers in several meetings and conducted thematic analysis [70,71] to unpack the requirements, conditions, and challenges around trustworthy AI that are presented in the following sections of this article.

While topics like bias and fairness—often central and technically-focused within AI ethics discussions—have received considerable attention, we aim here to highlight equally critical yet more socially and practically oriented themes, specifically the importance of knowledge and human involvement. These dimensions, as detailed by practitioners, offer deeper insights into real-world challenges within the medical field.

The More We Know, the More We Trust?

In the scientific discourse on AI, the trustworthiness of an AI system is closely tied to the requirement that the decision-making process must be transparent, explainable, and understandable, to allow medical professionals to make informed decisions [72]. This should include that AI systems communicate their limitations, for example, their inability to provide an accurate answer for a given patient or, in general, having a greater accuracy for certain patient groups, as well as include proper uncertainty quantification for their results [73]. This allows doctors to assess the appropriateness of using the AI tool for specific patients.

The need for transparency resonates with the fact that medical knowledge production inherently involves a degree of uncertainty. In scientific knowledge production, organized skepticism is part of the ethos of science [74]; however, the temporalities of clinical practice and scientific knowledge production differ. For AI, providing comprehensive explanations of internal processes may often be unrealistic [75]. Besides, striving for more explainable AI systems could reduce their efficacy, depending on the explainability techniques applied. For example, if a model is designed to be explainable from the outset using a simple model structure, like a linear additive model, its predictive performance could be compromised. On the other hand, even a simple model based on causal relationships verified through careful interpretation and domain knowledge may be robust and generalizable [76], but could be difficult to achieve for complex data types such as images or text. Hence, the pursuit of trustworthy AI raises questions about the balance between transparency and complexity.

Workshop outcomes indicate that what makes a difference to the trustworthiness of medical AI systems is not necessarily more, but more meaningful, situation- and user-oriented information as well as a good balance between knowledge and acceptable uncertainty. Experts highlighted that it is essential to ensure that the AI tool functions exactly as it is expected. Rather than detailed insights into the inner workings of an AI system, clarifying what knowledge is needed by which actors and areas of application has been emphasized. For instance, in the clinical context, it was deemed essential to know how the AI tool was built, eg, the type of dataset that was used to train the algorithm, its limitations and biases, the accuracy of the performance, but also the uncertainty of the outcome.

User-oriented knowledge takes several information needs of different stakeholders into account, such as developers, users, and patients, and these vary. For instance, developers benefit from multiple explainability techniques to identify biases to make sure that the predictions are not based on spurious correlations in the training data. In this respect, medical doctors might benefit from highlighting the part of the x-ray image that indicates the presence of cancer. From the patient's perspective, explainability is of interest primarily through the attending physician's judgment and experience. Patients may trust their physician more than they rely on the explainability of the AI system.

The requirement for specific types of knowledge reflects the relational aspects associated with explainability and AI trustworthiness. Several quantitative and qualitative explainability methods are used in AI development to understand why and how AI systems make specific predictions [77]. However, it is crucial to involve physicians during the tool development and testing stages in a human-in-the-loop manner. As domain experts, physicians use explainability tools to verify the accuracy of system predictions based on correct premises.

Explainability methods are essential for AI-based software as a medical device. Daily operation of the tool does not necessarily require explainability, though it becomes useful if medical professionals disagree with its predictions. In such cases, explainability algorithms help verify the prediction premises. Local visualization methods like heat maps or attention maps, supported by uncertainty metrics, are commonly used in image-based analysis as they are easier for nondevelopers to understand. Clinicians can then make informed decisions on whether to accept or reject the predictions based on their knowledge and judgment. However, few doctors are currently aware of how to review the technical aspects of AI tools. Therefore, it is essential that all information in training materials is presented in an understandable way, using methods and metrics that are familiar to health care professionals, and that professional associations are involved as institutional stakeholders.

Medical AI technologies should include alert systems to warn users when cases fall "out of scope" or outside the tool's "intended use," ensuring the tool is not applied to patients it was not designed for, thereby preventing potential harm. For instance, if an AI system for breast cancer screening was primarily trained on elderly patients with low-density breasts, its performance may be inferior for young patients with high-density breasts. Doctors should be informed of these limitations during training and should be alerted when the system is used in these cases. In this regard, AI tools in oncological imaging must meet the AI Act's high-risk requirements, including transparency about dataset limitations, essential for building trust among clinicians. Similarly, genomics tools such as polygenic risk scores [13] and biomarkers based on multiomics data [78,79] raise concerns about bias and generalizability, and in line with the obligations for documentation and human oversight of the AI Act, these risks should be addressed, and trustworthy use in clinical practice should be supported.

The additional training for medical professionals in the use of AI should include methods for quick and reliable assessment of tools in daily practice. This will enhance technical literacy in health care, demystify AI's potential, and clarify the tools' technical capabilities, thereby increasing trust. Nonetheless, the required training should be sufficient for safe use but limited in terms of attention and time consumption. AI-driven software, as another set of tools in daily practice, should integrate seamlessly into workflows without causing major interruptions. Doctors should be able to focus on patient care and medical knowledge rather than learning AI's inner workings. Therefore, adequate education and training at the time of AI-tool implementation, along with clear documentation, explainability, uncertainty metrics, and alert systems, are essential for building trust in AI solutions among health care professionals and other users.

Situating Humans: Human-in-the-Loop, Human in the Center, Humans in Loops of Trust

Approaches from the ethics of AI perspective emphasize that AI should involve human features, that humans need to be in the loop of AI processes, or that humans need to have the last word in decision-making to ensure the trustworthiness of an AI system [80]. According to the EU AI Act, AI should be human-centric, and thus, it should involve human-in-the-loop approaches that ensure the involvement of humans in all stages of medical AI development and testing. As mentioned earlier, trust is related to several interpersonal relationships and institutional conditions, and in this case, also to automated or autonomous technologies. Therefore, rather than asking *how* humans can be integrated into the AI loop, the situatedness of humans in terms of where and when within the entanglement of social-technological and institutional relations in connection to trust should receive further attention.

Workshop outcomes emphasize that a tool's trustworthiness is established through trust in humans (and their expertise), such as developers or physicians, and relationships. Moreover, trusting relationships are framed by trust in the social system or institutional framework in which it is embedded. Therefore, trust can be considered a dynamic process involving several actors and interactions among individuals, technologies, and institutions. Tackling issues around trust in medical AI necessitates emphasizing its multifaceted character, situatedness, and contextuality. In this regard, trust evolves as a dynamic process, spanning machine reliability, human relationships, and broader scientific skepticism. In addition, it also has a temporal dimension, as past experiences significantly influence perceptions of AI reliability in medicine.

In all its complex multidimensionality, which was discussed during the workshop, trust places humans at the center with the claim that we trust people, not just tools. Tools are judged on their reliability and credibility. For instance, an AI tool is credible and reliable when the system functions in a way that the individual expects and expectations themselves are cascading, considering that the designer, producer, the user (eg, medical doctor or the patient) may have different expectations that can be tied to each other as well as their own situatedness. The integration of AI into real-world medical settings relies on its promised functionality, such as accurate disease prediction or classification. This includes defining features, protocols, and patient-centered considerations during the design and implementation phases, including quality control after implementation. Ultimately, trust is built by humans, such as developers and physicians, who collaborate in increasingly multidisciplinary and international teams is also reenacted in settings that cannot be completely predefined and restricted.

Various examples of AI tools used in clinical settings allow insights into how human is situated. For instance, against a rapidly expanding plethora of AI tools, the clearance of these by the Food and Drug Administration in the US context relies on categorization of tools as low-risk and high-risk [81]. In the clinical context, however, many factors impact how risks are understood in practice, such as whether the tool is used to assist with a process under the clinician's complete control, such as a tool assisting a cardiologist with drawing the contours of the heart for consultation. On the contrary, there are also cases where the clinician must assume the reliability of tools, where institutional structures, such as health ministries and researchers, have already considered potential biases of the datasets that informed the algorithm, as well as during the entire development, validation, and auditing process. While the algorithmic impact assessment is expected to be performed by other experts before tools make it to clinical practice, the consequences of the opposite could be drastic for the patients and doctors: for instance, a tool that does not consider the breast glandular tissue of individuals according to varying lifestyles and demographics may directly impact the outcome of diagnosis or intervention.

The integration of AI tools into human medical professionals' practices presents a challenge as AI systems and medical professionals are intrinsically different in terms of the registers that they rely on. AI systems rely on mathematical rules, functions, and statistical algorithms, enabling their performance to be assessed with mathematical calculability. However, a major concern regarding the credibility of AI in medicine is the accurate translation of these mathematical capabilities into meaningful clinical value. One aspect that is relevant here is accuracy. The possible expectation for health care professionals to verify all AI-driven recommendations as an additional task to their daily routine to avoid harm in some patients, rather than only those flagged as potentially erroneous, would pose a significant impediment to the trustworthiness and widespread adoption of these technologies as standalone tools without human oversight.

The balance between sensitivity and specificity is particularly challenging in medical domains, such as oncology. Screening tools, like those used for breast cancer detection, where the prevalence of cancer in the examined samples is low, require high sensitivity to ensure that all potential cases are identified, even at the expense of a higher rate of false positives. The sensitivity of the AI tool should at least match, if not exceed, the average sensitivity of radiologists in national screening programs [82]. In such scenarios, the false-positive results can be addressed through subsequent verification procedures, such as biopsies. While the cost of a false positive result may be less harmful than missing the cancer, it can still have a significant impact on patients receiving misleading information [83], exemplifying the multidimensionality of AI's use. Conversely, in applications, such as those aimed at distinguishing small liver tumors from other lesions, high specificity is essential for early and accurate diagnosis, enabling timely treatment and improved patient outcomes.

The question, therefore, is: what level of accuracy and precision would be deemed “good enough” for specific clinical applications to foster trust and drive the wider adoption of AI tools in medicine? Addressing this challenge requires human understanding and judgment of the trade-offs among sensitivity, specificity, and the practical implications for health care professionals and patients.

Conclusions

This article centers on a synthesis of multidisciplinary discussions on trustworthy AI during an expert workshop, which allows us to underscore the complexities surrounding this topic in the evolving landscape of medical technology. The findings highlighted the nexus of knowledge on trust, emphasizing the situatedness of the human. Thus, conceptualizing and building trustworthiness in AI requires a comprehensive, multifaceted approach.

Trust and trustworthiness are not legal concepts. However, trustworthiness can be a feature of an AI system, and trust can be a purpose and effect of a law. Such trust can be reached by providing standards for trustworthy AI, including (1) transparent rules on roles, responsibilities, and procedures for AI development and by enforcing those rules through liability norms as well as through clear consequences for violation of duties; (2) quality and security features for AI tools; and (3) opportunities for public participation and debate. Clear, well-implemented, and effectively enforced laws contribute to public trust. Therefore, AI laws should prioritize defining roles, procedures, responsibilities, and liabilities alongside establishing efficient systems for control and enforcement. For example, the EU AI Act focuses on trustworthiness to foster trust through regulation [84]. Our article demonstrates that scholars from various domains converge on the understanding that achieving trust in AI is complex and involves individuals, making this legal expectation a desirable mission reflected in global AI regulations.

Contrary to the “move fast, break things” innovation maxim, setting high standards for quality and security is essential to maintaining the reliability and integrity of AI tools. Trust in AI is not an inherent feature, but a belief held by users. Trust involves the procedures, steps, and individuals behind the creation, use, and maintenance of AI tools, including developers, health care professionals, and update teams. Therefore, AI developers must build reliable systems to earn public trust and uphold their reputation. User trust relies on the AI’s trustworthiness, which depends on the transparency and verifiability of its development processes. To incorporate elements of human oversight by design, developers, for instance, strive more and more to visualize explainability in user interfaces when users interact with AI [85].

Historically, medicine embraced the authority of individual experts. In the 20th century, the concept of the “eminent expert” was replaced by evidence-based medicine (EBM) [86], which relies on scientific evidence produced by the broader scientific community. EBM establishes trust in new diagnostic and treatment methods. Similarly, the credibility of medical AI tools will grow with increasing good-quality evidence and defined accuracy and precision levels needed for specific tasks.

Researchers in the field of ethics of AI in medicine must strive for accuracy and precision by providing clear definitions for concepts, such as trustworthiness or trust in specific contexts, and situating them within broader societal issues. It is crucial to analyze the complex relationship between trustworthiness, trust, and explainability [87,88] and to find out what kind of explanations are required for specific situations and applications in medicine to adapt procedures accordingly. Interdisciplinary research, involving social scientists and clinicians, is crucial to incorporating clinical concepts [89]. This interdisciplinary approach involves a layered understanding, where ethical, societal, and legal issues from AI and clinical applications add further risks and complexities when combined [13].

Experts highlighted that trust in AI systems is bolstered not only by understanding how these tools operate but also by ensuring a balance between knowledge and acceptable uncertainty. Users need confidence that AI functions as intended.

Explainability can also play a role in the users’ and patients’ right to know, which affects their autonomy and agency to make informed decisions [90,91]. Transparency is crucial; however, our findings highlight the importance of providing context-specific, stakeholder-relevant information. Failure to uphold this right can result in epistemic injustices [92-95], especially for marginalized groups, who may be denied knowledge that may affect their rights and well-being. Upholding the right to know helps prevent epistemic injustices and enhances patients’ ability to comprehend the rationale behind diagnoses, treatment recommendations, and the use of AI tools in their care.

Our findings emphasize that establishing AI trustworthiness in health care requires a robust, human-centered, multistakeholder approach. Historically, AI tools have been predominantly engineered with limited input beyond technical development teams, often with minimal involvement from health care professionals. Documented real-world examples that fulfill all the requirements for trustworthy AI discussed in this article are scarce, and most scientific research focuses on evaluating some dimensions of trustworthiness, for example, by providing methodology to assess AI tools in diverse populations [96]. It is worth noticing that major private AI companies do report involving ethics and safety teams as part of their development, with parallels to our discussion, including techniques such as the use of external red-teaming to ensure privacy and safety, stakeholder engagement, and frameworks to assess biases and other social and ethical risks (for instance, Google DeepMind’s three-layered framework for evaluating the social and ethical risks of AI systems). However, the fact that many of the state-of-the-art AI models still remain black-box in terms of their inner workings and training data poses a challenge to their wider use in health care. In this regard, several engagement techniques can be used to collect adequate insights and continuous feedback from stakeholders. It is crucial that engagement activities are diverse and inclusive to ensure the

AI tool is designed for all and to promote equity in AI-mediated health care.

Engaging only stakeholders with high education and digital proficiency may result in tools that are inaccessible to individuals with low digital literacy, which may impact health outcomes [97]. Including vulnerable groups and minorities in AI design and development can help develop tools that are respectful of diverse needs and contexts. Hence, we recommend that AI development teams should involve all relevant stakeholders at all stages of the AI life cycle, that is, conception, technical development, clinical validation, and real-world deployment (Textbox 1).

Textbox 1. Authors' recommendations for involving relevant stakeholders across all stages of the AI (artificial intelligence) life cycle.

The authors provide recommendations for involving relevant stakeholders at each stage of the AI life cycle, emphasizing continuous engagement to ensure transparency, accountability, and inclusiveness throughout the development and deployment of AI systems.

AI design phase:

Depending on the AI application, a range of health care professionals, such as general practitioners, specialists (domain experts), health care managers, nurses, and technicians, should be engaged for requirements elicitation. This includes defining the intended use of AI tools, clinical endpoints, success criteria, and the specific requirements for trustworthiness and transparency. Furthermore, clinicians can help specify the most adequate approaches for explainability, the types of explanations needed, and the conditions under which alerts or warnings should be issued.

Beyond clinicians, patient engagement is vital for identifying user needs, preferences, and potential barriers to trust and adherence to AI-mediated care. Ethicists and social scientists should also play a crucial role during conception, especially for anticipating the application-specific ethical and social impacts of the AI tools, such as misalignments with fundamental rights, effects on deskilling, changes in power relationships, and alterations in human behaviors as AI is integrated into care settings. Considering the approaches of trustworthy AI in Table 1, they can also apply qualitative research methods to examine the intentions and actions of the AI developers, understand the AI-mediated doctor-patient relationship, and assess the trustworthiness and perspectives of the institutions involved.

AI development phase:

At this stage, the AI team should focus on translating the stakeholder-defined requirements and identified risks into development strategies, including mitigation measures. This includes compiling diverse and representative training datasets, employing machine learning methods that minimize potential biases, and developing AI-human interfaces that enhance user interaction and comprehension of the AI system. It is also important to ensure that the AI tool's technical development considers existing care models for seamless integration into real-world practice, thereby adding value without disrupting established workflows. During the development phase, it is important that stakeholders continue to be engaged, so they can monitor the technical developments and provide continuous feedback on the AI tool's anticipated level of trustworthiness.

AI validation phase:

During the validation phase, it is important to assess the AI tool's trustworthiness across multiple dimensions, including robustness under real-world conditions, level of transparency and explainability, fairness concerning diverse groups, usability in practice, and ethical and social compliance. This phase should continue to engage all stakeholders, including social scientists to evaluate the socio-behavioral implications of the AI tools on end users, such as whether the tools and their explanations enhance or diminish user confidence and trust, affect users' ability to retain judgment when using the AI tool, and improve or degrade doctor-patient relationships.

AI deployment phase:

Once the tool is validated, certified, and deployed, a multistakeholder team must continue to monitor its performance and impact in real-world settings. This includes conducting periodic evaluations and audits to identify any performance degradation or emerging ethical issues, implementing logging systems to enhance traceability and accountability, and ensuring robust human oversight mechanisms are in place. These steps are crucial for maintaining user trust and ensuring that human autonomy is respected, demonstrating that there is adequate governance surrounding the use, maintenance, and oversight of the AI tool.

By adhering to a human-centered, multicentered framework, AI development teams can create tools that are not only technically efficient but also ethically sound and broadly accepted by all relevant users, groups, and institutions. This inclusive approach ensures that AI systems are developed with an in-depth understanding of the various contexts in which they will operate, leading to more effective and equitable AI-mediated health care.

Recent years have witnessed a surge of foundation models, that is, AI models trained with massive data and computational resources, which can also solve medical problems out-of-the-box. Though not the focus of this article, the aspects of trust discussed here can also be relevant when pretraining foundation models or fine-tuning them for deployment in health care use cases with additional local data. Studies focusing on the trustworthiness of foundation models have started to emerge [98], and we expect more work in this direction in the future.

Finally, future research should examine how the legal notion of "trustworthy AI" under the EU AI Act is interpreted and implemented across different member states, especially in clinical contexts. Comparative legal analysis could reveal how national competent authorities enforce trust-related obligations, such as human oversight or transparency in medical AI systems. Furthermore, empirical studies are needed to assess whether regulatory compliance translates into perceived trust among clinicians and patients. For instance, guidelines and recommendations, whether they have been established top down by policy makers or bottom up by the scientific community and practitioners, need to be tested to see if they are the effective and practical governance tools they are intended to be. This requires an interdisciplinary effort by asking if their uptake by individuals as well as institutions is meaningful in practice. Consider, exemplarily, the assessment of the added value of the inclusion of patient and citizen groups in the design, validation, and deployment phases of AI systems and the development of standardized frameworks for measuring trust across stakeholders and exploring how such trust is cultivated or undermined over time. These also involve evaluating the various roles of stakeholders and experts in these processes, particularly with regard to responsibility, distribution of roles, and power. To determine whether trust-building is effective, new methodologies will be required and should go beyond the quantitative measurement of key performance indicators and allow for thorough qualitative assessments that allow agile management.

Acknowledgments

This study was supported by the European Union's Horizon 2020 projects EuCanImage (grant agreement No 952103) and INTERVENE (grant agreement No 101016775), as well as by the Innovative Medicines Initiative 2 Joint Undertaking project Bigpicture (grant agreement No 945358) and BIOMAP, which has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU; grant agreement No 821511). The JU receives support from the European Union's Horizon 2020 research and innovation program and EFPIA. This publication reflects only the author's view and the JU is not responsible for any use that may be made of the information it contains. PM received funding from the Research Council of Finland (Finnish Center for Artificial Intelligence FCAI, and grants 352986, 358246, NextGenerationEU). Where authors are identified as personnel of the Biobanking and BioMolecular resources Research Infrastructure (BBMRI-ERIC), the authors alone are responsible for the views expressed in this article, and they do not necessarily represent the decisions, policy, or views of BBMRI-ERIC. Special thanks to Verena Borecký for proofreading the article.

Conflicts of Interest

None declared.

References

1. Ethics guidelines for trustworthy AI. European Commission. 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [accessed 2021-12-28]
2. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. 2021. URL: https://iris.who.int/bitstream/handle/10665/341996/9789240029200-eng.pdf?sequence=1 [accessed 3024-10-15]
3. Lekadir K, Frangi A, Porras A. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025;388:r340. [FREE Full text] [doi: 10.1136/bmj.r340] [Medline: 39961614]
4. Serban A, Blom K, Hoos H, Visser J. Practices for engineering trustworthy machine learning applications. In: 2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN). Piscataway, NJ. IEEE; 2021:97-100.
5. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell. 2019;1(11):501-507. [doi: 10.1038/s42256-019-0114-4]
6. Floridi L, Cowls J, King TC, Taddeo M. How to design AI for social good: seven essential factors. In: Floridi L, editor. Ethics, Governance, and Policies in Artificial Intelligence. Cham, Switzerland. Springer; 2021:125-151.
7. Hagendorff T. The ethics of AI ethics: an evaluation of guidelines. Minds Mach. 2020;30(1):99-120. [FREE Full text] [doi: 10.1007/s11023-020-09517-8]
8. Zink A, Rose S. Fair regression for health care spending. Biometrics. 2020;76(3):973-982. [FREE Full text] [doi: 10.1111/biom.13206] [Medline: 31860120]
9. DeCamp M, Tilburt JC. Why we cannot trust artificial intelligence in medicine. Lancet Digital Health. 2019;1(8):e390. [FREE Full text] [doi: 10.1016/s2589-7500(19)30197-9]
10. Hatherley JJ. Limits of trust in medical AI. J Med Ethics. 2020;46(7):478-481. [doi: 10.1136/medethics-2019-105935] [Medline: 32220870]
11. 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. 2020;1-2:100001. [FREE Full text] [doi: 10.1016/j.ibmed.2020.100001]
12. Goisauf M, Cano Abadía M. Ethics of AI in radiology: a review of ethical and societal implications. Front Big Data. 2022;5(850383):1-13. [FREE Full text] [doi: 10.3389/fdata.2022.850383] [Medline: 35910490]
13. Fritzsche MC, Akyüz K, Cano Abadía M, McLennan S, Marttinen P, Mayrhofer MT, et al. Ethical layering in AI-driven polygenic risk scores-new complexities, new challenges. Front Genet. 2023;14:1098439. [FREE Full text] [doi: 10.3389/fgene.2023.1098439] [Medline: 36816027]
14. Geis JR, Brady AP, Wu CC, Spencer J, Ranschaert E, Jaremko JL, et al. Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. Can Assoc Radiol J. 2019;70(4):329-334. [doi: 10.1016/j.carj.2019.08.010] [Medline: 31585825]
15. Hawley K. Trust, distrust and commitment. Noûs. 2012;48(1):1-20. [doi: 10.1111/nous.12000]
16. Jones K. Trustworthiness. Ethics. 2012;123(1):61-85. [doi: 10.1086/667838]
17. Wyatt S, Harris A, Adams S, Kelly SE. Illness online: Self-reported data and questions of trust in medical and social research. Theory Cult Soc. 2013;30(4):131-150. [doi: 10.1177/0263276413485900]
18. Bijker EM, Sauerwein RW, Bijker WE. Controlled human malaria infection trials: how tandems of trust and control construct scientific knowledge. Soc Stud Sci. 2016;46(1):56-86. [doi: 10.1177/0306312715619784] [Medline: 26983172]

19. Werner-Felmayer G, Minari J, Schicktanz S, Raz A, Sharon T. Data-intensive medicine and healthcare: ethical and social implications in the era of artificial intelligence and automated decision making. In: Frontiers in Genetics. Lausanne. Frontiers Media SA; 2023.
20. Owens K, Walker A. Those designing healthcare algorithms must become actively anti-racist. Nat Med. 2020;26(9):1327-1328. [FREE Full text] [doi: 10.1038/s41591-020-1020-3] [Medline: 32908272]
21. Quinn TP, Jacobs S, Senadeera M, Le V, Coghlan S. The three ghosts of medical AI: can the black-box present deliver? Artif Intell Med. 2022;124:102158. [FREE Full text] [doi: 10.1016/j.artmed.2021.102158] [Medline: 34511267]
22. Rasheed K, Qayyum A, Ghaly M, Al-Fuqaha A, Razi A, Qadir J. Explainable, trustworthy, and ethical machine learning for healthcare: a survey. Comput Biol Med. 2022;149:106043. [FREE Full text] [doi: 10.1016/j.compbiomed.2022.106043] [Medline: 36115302]
23. Celi LA, Cellini J, Charpignon ML, Dee EC, Dernoncourt F, Eber R, et al. for MIT Critical Data. Sources of bias in artificial intelligence that perpetuate healthcare disparities-a global review. PLOS Digit Health. 2022;1(3):e0000022. [FREE Full text] [doi: 10.1371/journal.pdig.0000022] [Medline: 36812532]
24. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27(12):2176-2182. [FREE Full text] [doi: 10.1038/s41591-021-01595-0] [Medline: 34893776]
25. DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. J Am Med Inform Assoc. 2020;27(12):2020-2023. [FREE Full text] [doi: 10.1093/jamia/ocaa094] [Medline: 32574353]
26. Schwartz R, Vassilev A, Greene K, Perine L, Burt A, Hall P. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. NIST Special Publication 1270. 2022. URL: https://www.dwt.com/-/media/files/blogs/artificial-intelligence-law-advisor/2022/03/nist-sp-1270--identifying-and-managing-bias-in-ai.pdf [accessed 2024-11-15]
27. Simion M, Kelp C. Trustworthy artificial intelligence. AJPH. 2023;2(1):8. [doi: 10.1007/s44204-023-00063-5]
28. Ienca M, Starke G. Misplaced trust and distrust: how not to engage with medical artificial intelligence. Camb Q Healthc Ethics. 2022;33(3):360-369. [doi: 10.1017/s0963180122000445]
29. Grote T. Trustworthy medical AI systems need to know when they don't know. J Med Ethics. 2021;47:337-338. [FREE Full text] [doi: 10.1136/medethics-2021-107463] [Medline: 33849959]
30. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. 2020;46(3):205-211. [FREE Full text] [doi: 10.1136/medethics-2019-105586] [Medline: 31748206]
31. Carter MA. Trust, power, and vulnerability: a discourse on helping in nursing. Nurs Clin North Am. 2009;44(4):393-405. [FREE Full text] [doi: 10.1016/j.cnur.2009.07.012] [Medline: 19850176]
32. Baier A. Trust and antitrust. Ethics. 1986;96(2):231-260. [doi: 10.1086/292745]
33. Butler J. Frames of War: When is Life Grievable? New York, NY. Verso; 2016.
34. Butler J, Gambetti Z, Sabsay L, editors. Vulnerability in Resistance. Durham, NC. Duke University Press; 2016.
35. Ferrarese E. Vulnerability and critical theory. BRP Crit Theory. 2016;1(2):1-88. [FREE Full text] [doi: 10.1163/24519529-12340002]
36. Mackenzie C. Mackenzie C, Rogers W, Dodds S, editors. Vulnerability: New Essays in Ethics Feminist Philosophy. England. Oxford University Press; 2014:33-59.
37. Meyer S, Ward P, Coveney J, Rogers W. Trust in the health system: an analysis and extension of the social theories of Giddens and Luhmann. Health Sociol Rev. 2014;17(2):177-186. [doi: 10.5172/hesr.451.17.2.177]
38. Haraway DJ. Situated knowledges: the science question in feminism and the privilege of partial perspective. Fem Stud. 1988;14(3):575-599. [doi: 10.2307/3178066]
39. Collins HM. The TEA set: Tacit knowledge and scientific networks. Science Studies. 1974;4(2):165-185. [doi: 10.1177/030631277400400203]
40. Wynne B. Public uptake of science: a case for institutional reflexivity. Public Underst Sci. 1993;2(4):321-337. [doi: 10.1088/0963-6625/2/4/003]
41. Irwin A, Wynne B. Misunderstanding science?: The Public Reconstruction of Science and Technology. New York. Cambridge University Press; 1996.
42. Wynne B. Public engagement as a means of restoring public trust in science--hitting the notes, but missing the music? Community Genet. 2006;9(3):211-220. [doi: 10.1159/000092659] [Medline: 16741352]
43. Beck U. World at Risk. Cambridge. Polity; 2009.
44. Wynne B. Misunderstood misunderstanding: social identities and public uptake of science. Public Underst Sci. 2016;1(3):281-304. [doi: 10.1088/0963-6625/1/3/004]
45. Hilgartner S. Science on Stage: Expert Advice as Public Drama. Stanford. Stanford University Press; 2000.
46. Felt U. Keeping technologies out: Sociotechnical imaginaries and the formation of Austria's technopolitical identity. In: Jasanoff S, Kim SH, editors. Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power. Chicago. University of Chicago Press; 2015:103-125.
47. Rogers-Hayden T, Pidgeon N. Moving engagement “upstream”? Nanotechnologies and the royal society and royal academy of engineering's inquiry. Public Underst Sci. 2007;16(3):345-364. [doi: 10.1177/0963662506076141]

48. Jasanoff S. Technologies of humility: citizen participation in governing science. Minerva. 2003;41(3):223-244. [doi: 10.1023/A:1025557512320]
49. Jasanoff S, Hurlbut JB. A global observatory for gene editing. Nature. 2018;555(7697):435-437. [doi: 10.1038/d41586-018-03270-w] [Medline: 29565415]
50. Pellegrino ED, Thomasma DC. The conflict between autonomy and beneficence in medical ethics: proposal for a resolution. J Contemp Health Law Policy. 1987;3(1):23-46. [FREE Full text]
51. Emanuel EJ, Emanuel LL. Four models of the physician-patient relationship. JAMA. 1992;267(16):2221-2226. [doi: 10.1001/jama.1992.03480160079038]
52. Kittay EF, Meyers DT, editors. Women and Moral Theory. Totowa, NJ. Rowman & Littlefield; 1987.
53. Cotterrell R. Trusting in law: legal and moral concepts of trust. Curr Le Probl. 1993;46(2):75-95. [doi: 10.1093/clp/46.part_2.75]
54. Cross FB. Law and trust. Georgetown Law Journal. 2004;93(5):1457-1545. [FREE Full text]
55. Trusts. Faculty of Law, University of Oxford. URL: https://www.law.ox.ac.uk/content/trusts [accessed 2024-08-12]
56. Powell R. 2 Good faith in contracts 1. Curr Leg Probl. 1956;9(1):16-38. [doi: 10.1093/clp/9.1.16]
57. Brownsword R, Hird NJ, Howells GG. Good Faith in Contract: Concept and Context. Dartmouth. Ashgate; 1999.
58. Ribstein LE. Law v. Trust. SSRN. 2001;81(3):553-590. [doi: 10.2139/ssrn.247224]
59. Hult D. Creating trust by means of legislation – a conceptual analysis and critical discussion. Theory Pract Legis. 2018;6(1):1-23. [doi: 10.1080/20508840.2018.1434934]
60. Greco T. La Legge Della Fiducia: alle Radici del Diritto. Laterza. Roma; 2021.
61. Smuha NA, Ahmed-Rengers E, Harkens A, Li W, MacLaren J, Piselli R, et al. How the EU can achieve legally trustworthy AI: a response to the European Commission's proposal for an artificial intelligence act. SSRN. Preprint posted online August 5, 2021. [FREE Full text] [doi: 10.2139/ssrn.3899991]
62. Corrigendum. European Parliament. 2024. URL: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf [accessed 2024-08-12]
63. EUR-Lex: access to European Union law. European Union. 2024. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745 [accessed 2024-09-02]
64. EUR-Lex: access to European Union law. European Union. URL: https://eur-lex.europa.eu/eli/reg/2017/746/oj [accessed 2024-09-02]
65. Products. Radiology Health AI Register. URL: https://radiology.healthairegister.com/products/ [accessed 2024-09-02]
66. Busch F, Kather JN, Johner C, Moser M, Truhn D, Adams LC, et al. Navigating the European union artificial intelligence act for healthcare. NPJ Digit Med. 2024;7(1):210. [FREE Full text] [doi: 10.1038/s41746-024-01213-6] [Medline: 39134637]
67. Murtagh MJ, Minion JT, Turner A, Wilson RC, Blell M, Ochieng C, et al. The ECOUTER methodology for stakeholder engagement in translational research. BMC Med Ethics. 2017;18(1):24. [FREE Full text] [doi: 10.1186/s12910-017-0167-z] [Medline: 28376776]
68. Bloor M, Frankland J, Thomas M, Robson K. Focus Groups in Social Research. Thousand Oaks. Sage; 2002.
69. Wilson RC, Butters OW, Clark T, Minion J, Turner A, Murtagh MJ. Digital methodology to implement the ECOUTER engagement process. F1000Res. 2016;5:1307. [doi: 10.12688/f1000research.8786.1]
70. Braun V, Clarke V. What can "thematic analysis" offer health and wellbeing researchers? Int J Qual Stud Health Well-being. 2014;9:26152. [FREE Full text] [doi: 10.3402/qhw.v9.26152] [Medline: 25326092]
71. Terry G, Hayfield N, Clarke V, Braun V. Thematic analysis. In: Willig C, Stainton R, editors. The SAGE Handbook of Qualitative Research in Psychology. Thousand Oaks. Sage; 2017:17-37.
72. Markus A, Kors J, Rijnbeek P. 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. 2021;113:103655. [FREE Full text] [doi: 10.1016/j.jbi.2020.103655] [Medline: 33309898]
73. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. Inf Fusion. 2021;76:243-297. [FREE Full text] [doi: 10.1016/j.inffus.2021.05.008]
74. Merton R. The normative structure of science. In: The sociology of science: Theoretical and empirical investigations. Chicago. University of Chicago Press; 1973:267-278.
75. Bjerring JC, Busch J. Artificial intelligence and patient-centered decision-making. Philos. Technol. 2020;34(2):349-371. [doi: 10.1007/s13347-019-00391-6]
76. Peters J, Janzing D, Schölkopf B. Elements of Causal Inference: Foundations and Learning Algorithms. Cambridge, Massachusetts. The MIT Press; 2017.
77. Champendal M, Müller H, Prior J, Dos Reis CS. A scoping review of interpretability and explainability concerning artificial intelligence methods in medical imaging. Eur J Radiol. 2023;169:111159. [FREE Full text] [doi: 10.1016/j.ejrad.2023.111159] [Medline: 37976760]
78. Hangel N, Buyx A, Fritzsche M. The interrelation of scientific, ethical, and translational challenges for precision medicine with multimodal biomarkers - a qualitative expert interview study in dermatology research. Heliyon. 2024;10(13):e31723. [FREE Full text] [doi: 10.1016/j.heliyon.2024.e31723] [Medline: 39040296]

79. Fritzsche MC, Buyx AM, Hangel N. Mapping ethical and social aspects of biomarker research and its application in atopic dermatitis and psoriasis: a systematic review of reason. J Eur Acad Dermatol Venereol. 2022;36(8):1201-1213. [doi: 10.1111/jdv.18128] [Medline: 35366351]
80. High-level expert group on Artificial Intelligence. European Commission. 2019. URL: https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_policy_and_investment_recommendations.pdf [accessed 2022-11-14]
81. Muehlematter UJ, Bluethgen C, Vokinger KN. FDA-cleared artificial intelligence and machine learning-based medical devices and their 510(k) predicate networks. Lancet Digit Health. 2023;5(9):e618-e626. [FREE Full text] [doi: 10.1016/S2589-7500(23)00126-7] [Medline: 37625896]
82. Lehman C, Arao R, Sprague B, Lee JM, Buist DSM, Kerlikowske K, et al. National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. Radiology. 2017;283(1):49-58. [FREE Full text] [doi: 10.1148/radiol.2016161174] [Medline: 27918707]
83. Rasmussen JF, Siersma V, Malmqvist J, Brodersen J. Psychosocial consequences of false positives in the Danish lung cancer CT screening trial: a nested matched cohort study. BMJ Open. 2020;10(6):e034682. [FREE Full text] [doi: 10.1136/bmjopen-2019-034682] [Medline: 32503869]
84. Tamò‐Larrieux A, Guitton C, Mayer S, Lutz C. Regulating for trust: can law establish trust in artificial intelligence? Regul Gov. 2023;18(3):780-801. [FREE Full text] [doi: 10.1111/rego.12568]
85. Holzinger A, Zatloukal K, Müller H. Is human oversight to AI systems still possible? N Biotechnol. 2025;85:59-62. [FREE Full text] [doi: 10.1016/j.nbt.2024.12.003] [Medline: 39675423]
86. Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA. 1992;268(17):2420-2425. [doi: 10.1001/jama.1992.03490170092032] [Medline: 1404801]
87. 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. 2019;33(3):523-539. [doi: 10.1007/s13347-019-00378-3]
88. Arbelaez Ossa L, Starke G, Lorenzini G, Vogt JE, Shaw DM, Elger BS. Re-focusing explainability in medicine. Digit Health. 2022;8:20552076221074488. [FREE Full text] [doi: 10.1177/20552076221074488] [Medline: 35173981]
89. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI Consortium. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. Feb 05, 2025;388:e081554. [FREE Full text] [doi: 10.1136/bmj-2024-081554] [Medline: 39909534]
90. Giesen D. The patient's right to know--a comparative law perspective. Med Law. 1993;12(6-8):553-565. [Medline: 8183063]
91. Inkeroinen S, Virtanen H, Stolt M, Leino-Kilpi H. Patients' right to know: a scoping review. J Clin Nurs. 2023;32(15-16):4311-4324. [doi: 10.1111/jocn.16603] [Medline: 36550593]
92. Fricker M. Evolving concepts of epistemic injustice. In: Kidd IJ, Medina J, Pohlhaus GJ, editors. The Routledge handbook of epistemic injustice. United Kingdom. Routledge; 2017:53-60.
93. Carel H, Kidd I. Epistemic injustice in healthcare: a philosophial analysis. Med Health Care Philos. 2014;17(4):529-540. [doi: 10.1007/s11019-014-9560-2] [Medline: 24740808]
94. Pozzi G, Durán J. From ethics to epistemology and back again: informativeness and epistemic injustice in explanatory medical machine learning. AI & Soc. 2024;40(2):299-310. [doi: 10.1007/s00146-024-01875-6]
95. Perdomo Reyes I. Injusticia epistémica y reproducción de sesgos de género en la inteligencia artificial. Rev Iberoam Cienc Tecnol Soc. 2024;19(56):89-100. [doi: 10.52712/issn.1850-0013-555]
96. Fajtl J, Welikala RA, Barman S, Chambers R, Bolter L, Anderson J, et al. Trustworthy evaluation of clinical AI for analysis of medical images in diverse populations. NEJM AI. 2024;1(9). [doi: 10.1056/aioa2400353]
97. Kemp E, Trigg J, Beatty L, Christensen C, Dhillon HM, Maeder A, et al. Health literacy, digital health literacy and the implementation of digital health technologies in cancer care: the need for a strategic approach. Health Promot J Austr. 2021;32 Suppl 1:104-114. [doi: 10.1002/hpja.387] [Medline: 32681656]
98. Shi C, Rezai R, Yang J, Dou Q, Li X. A survey on trustworthiness in foundation models for medical image analysis. arXiv:240715851. 2024.

Abbreviations

AI: artificial intelligence
AI HLEG: High-Level Expert Group on Artificial Intelligence
EBM: evidence-based medicine
EU: European Union
FUTURE-AI: Fairness, Universality, Traceability, Usability, Robustness, and Explainability–artificial intelligence
MDR: Medical Device Regulation
STS: science and technology studies
WHO: World Health Organization

Edited by A Mavragani; submitted 13.01.25; peer-reviewed by N Schlicker, S Fitzek; comments to author 16.03.25; revised version received 04.04.25; accepted 21.04.25; published 02.06.25

Please cite as:
Goisauf M, Cano Abadía M, Akyüz K, Bobowicz M, Buyx A, Colussi I, Fritzsche M-C, Lekadir K, Marttinen P, Mayrhofer MT, Meszaros J
Trust, Trustworthiness, and the Future of Medical AI: Outcomes of an Interdisciplinary Expert Workshop
J Med Internet Res 2025;27:e71236
URL: https://www.jmir.org/2025/1/e71236
doi: 10.2196/71236
PMID: