Intelligence-Based Medicine 1-2 (2020) 100001

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.journals.elsevier.com/intelligence-based-medicine

What we talk about when we talk about trust: Theory of trust for AI in healthcare

Felix Gille, Anna Jobin, Marcello Ienca

Eidgenössische Technische Hochschule Zürich, Department for Health Sciences and Technology, Health Ethics & Policy Lab, Zürich, Switzerland

ARTICLE INFO

Keywords:
Artificial intelligence
Trust theory
Healthcare
Ethics
Policy guidelines

ABSTRACT

Artificial intelligence (AI) is at the forefront of innovation in medicine. Researchers and AI developers have often claimed that "trust" is a critical determinant of the successful adoption of AI in medicine. Despite the pivotal role of trust and the emergence of an array of expert-informed guidelines on how to design and implement "trustworthy AI" in medicine, we found little common understanding across these guidelines on what constitutes user trust in AI and what the requirements are for its realization. In this article, we call for a conceptual framework of trust in health-related AI which is based not just on expert opinion, but first and foremost on sound empirical research and conceptual rigor. Only with a well-grounded and comprehensive understanding of the trust construct, we will be able to inform AI design and acceptance in medicine in a meaningful way.

Artificial intelligence (AI) applications, including advanced machine learning (ML), are central to healthcare innovation. Examples are the application of deep learning and computer vision to radiology and dermatology, natural language processing approaches to mental health screening, the use of AI-led health chat-bots for telemedicine, and intelligent assistive technologies for elderly and dementia care [1–5]. While promising great opportunities, AI applications in medicine raise social, legal and economic challenges [6]. These include the reliability and accountability of predictive AI systems, their transformative impact on clinical decision-making and doctor-patient dynamics and the trustworthiness of AI-powered devices. Researchers and AI developers have argued that a trust relationship between the user (clinician or patient) and the AI could help overcome these challenges [7,8]. We concur that trust is paramount for the well-functioning of healthcare systems and, consequently, for the acceptance of AI by physicians and within healthcare more broadly [9]. Yet, determining how such a trust relationship can actually be realized in healthcare is difficult for several reasons.

First, we lack conceptual clarity on the meaning and dynamics of trust [10]. "Trust" is relational, highly complex and involves at least two actors: one actor trusts the other actor to do, or not to do, an activity. This relationship is influenced by diverse framing factors — culture, belief systems, context, to name a few—and by the traits of the individual actors in the relationship. Therefore, trust is highly situational and difficult to develop as a "general concept."

Second, the trait "trustworthiness" and the relational construct "trust" are often conflated in both policy and research. This conflation does not only lead to conceptual confusion, but may also foster false hopes among AI users and developers. Trust and trustworthiness are different concepts, and trustworthiness does not lead *per se* to a trust relationship. This is perhaps best illustrated by the situation where several trustworthy alternatives exist: In that case, the trusting actor may choose to engage in a relationship based on different grounds than trustworthiness alone (e.g. an application's purchase price, or intuitive user interface). Therefore, any debate on trust and AI should focus on the entire relationship-building process and not on the trait of trustworthiness alone.

Third, the complex nature of AI challenges current theories and practices of trust in healthcare. One prime challenge is the so-called "black box'' problem inherent in certain forms of AI [11]. This metaphor describes the difficulty of deciphering how certain algorithms learn what they learn and produce certain outputs. Algorithms of this type are usually called 'opaque' and are typically observable in approaches to AI such as artificial neural networks, unsupervised ML and deep learning. In the latter process, artificial neural nets process information through a hierarchy of interconnected layers and create a model (i.e. a structured set of relationships) that can classify information under conditions it had not previously encountered. However effective, these systems offer few clues as to how they arrive at their conclusions, hence raise questions of transparency, accountability and responsibility — three fundamental factors to build trust in AI as highlighted in the 2019 Code of conduct for data-driven health and care technology by the UK Department of Health and Social Care [12]. Technical approaches to developing 'explainable AI' are critical to address the black-box problem (also called

E-mail address: felix.gille@hest.ethz.ch (F. Gille).

https://doi.org/10.1016/j.ibmed.2020.100001
Received 12 February 2020; Received in revised form 4 June 2020; Accepted 18 June 2020

'interpretability problem'). In particular, approaches that use counterfactual probes to review which factors affect a neural network's output, such as the local-interpretable model-agnostic explanations (LIME) program, hold great promise for enhancing explainability and thereby strengthening transparency and trust [13]. However, the 'interpretability problem' cannot be solved by technical approaches alone [14]. In order to provide the desired 'interpretability', technical methodologies to 'explainable AI' need to be combined with ethical and legal expertise that accounts for the nuances between the notions of explanation, interpretation and understanding [15].

Fourth, research focusing on physicians' trust in AI highlights several concerns that can decrease trust in AI and ML as applied in clinical settings. Such concerns include, among others, the low number of randomized clinical trials to test the performance of AI systems, the lack of transparency of information flows within AI applications, the risk of inequity and discrimination introduced by algorithmic biases, and insufficient regulatory clarity [16,17].

Last, limited public literacy about AI further complicates the build-up of trust. Public perceptions of AI are shaped more pronouncedly by science-fiction writers than by scientists [18]. If media contributions inflate misconceptions and/or unrealistic expectations about AI, public trust will inevitably decrease after the AI-bubble bursts. Vice versa, unfounded fears towards AI might misdirect public debate [19]. On a professional level, a South Korean physicians survey indicates that only 5.9% reported to have a good familiarity with AI [20]. In general, familiarity that develops from previous positive experiences with the to-be-trusted is vital to the establishment of trust [10]. This low percentage is alarming as physicians' trust is key for the adoption of AI and ML [21].

These conceptual challenges result in a paucity of research attempting to measure the baseline level of AI users' trust in AI. This lack of baseline data is a fundamental shortcoming to a meaningful debate of AI users' trust in AI.

How do we develop a theory of trust in AI in healthcare?

The concept of trust in AI needs to be defined based on sound scientific evidence [22]. Only when it is based on a comprehensive understanding of the concept of trust in AI we will be able to develop meaningful trust-promoting policies. Otherwise, such attempts will merely be a lucky shot.

Current prominent examples of conceptual and practical guides to establishing trust in AI fall short. A conceptual framework for trustworthy AI based on industry-led expert opinions, as recently developed by the European Commission's 'Ethics Guidelines for Trustworthy AI' (2019), appears to be ill-suited to guide trustworthy AI design. The high-level expert group who developed these guidelines consisted of four ethicists, alongside 48 non-ethicists, most of which were working for the industry. As argued by Thomas Metzinger, one member of the expert group, the guidelines were watered down because "AI ethics" was basically defined by the private sector alone [23]. From a conceptual view, these guidelines omit the establishment of a trusting relationship and jump from postulating trustworthiness abstractly to assessing trustworthiness in AI, ignoring the trust-establishment phase entirely.

A recent review of 84 AI ethics guidelines showed vast divergence on how the underlying principles of these guidelines are interpreted. Although one in three documents specifically address issues of trust, no coherent understanding of this notion is observable. Different guidelines provide contradictory advice on how trust in AI can be achieved [24]: some imply, for example, that trust can be fostered through transparency, whereas others suggest building or sustaining trust through education, reliability or accountability. Consistent with the scholarly literature on the topic, some guidelines describe trust as a necessary, beneficial condition for the adoption of AI. Others, however, take a contradictory stance and caution against trusting AI. These divergent interpretations generate uncertainty about how trust relationships function, or should function—as to who exactly should trust whom. This requires us to rethink our theoretical approaches to trust and AI.

To reduce uncertainty and conceptual confusion, we suggest that at least four fundamental questions need to be clarified:

1. What is a fitting conceptualization of trust in AI in the healthcare domain?

Well established guidelines on construct development stress the importance of robust research [25,26]. There is a need to conduct sound qualitative research with relevant stakeholders, in particular AI users, developers, legislators and the public. Further, there is a need to review and synthesize existing research and link AI ethics with wider trust theory to accurately conceptualize trust in AI in healthcare systems. Relying on expert opinion alone will not serve justice to the complexity of the concept and elude adequate and systematic scientific vetting.

2. What specific contextual factors frame trust in the healthcare domain?

Public trust research in the healthcare system revealed that trust is prone to contextual factors and susceptible to spill-over effects of mistrust from other areas which are not necessarily associated with healthcare [10]. For example, parts of the public do not trust the government to safeguard personal data in the banking sector, so they do not trust the same government to safeguard personal data within medical research. Hence, we need to understand what these influential factors are and, consequently, which actors from outside the domains of both health and AI can influence trust in the sphere of AI in healthcare. This way, we are able to better safeguard such trust from unforeseen negative influences.

3. *Who is the trusting and the trusted actors in a trust relationship?* In particular: can a meaningful trust relationship occur between a human and an AI system, or is this relationship always between patients and health professionals on the one hand, and human providers of AI services on the other hand?

Similar to other trust networks within the healthcare system [27], trust in AI may not only be built between an AI system and the user, if it is located there at all, but also between humans. Indeed, traditional trust theory suggests that trust is a construct developing between humans, although we know today that humans can also trust a technology [28,29]. Therefore, to build trust in AI and to intervene appropriately if trust levels are low or decreasing, we need a precise understanding of the AI trust relationship to be able to address the involved parties with targeted measures.

4. What strategies are empirically proven to be effective in building trust relationships in the context of AI?

Building on the previous three points, we need to develop and validate measures that aid the buildup of trust in AI. Such measures may range from guidance for AI designers or regulatory advice, to training and education of AI users, the creation of third-party AI accreditation authorities, or new AI implementation policies. In a nutshell, the appropriate measures need to cover the entire lifecycle of AI, including development, approval, implementation, use and evaluation.

A theory of trust in health-related AI will need to answer these questions and resolve the intricate relationships between all relevant actors. It will also need to consider both the context of application as well as the challenges emerging from the application of AI. To provide conceptual clarity on what trust in health-related AI actually means, we need to ask the right questions and conduct sound empirical research. Only based on corroborated evidence we will be able to build trust in health-related AI applications among all relevant actors (health professionals, patients, informal caregivers, health service managers etc.) or even create a truly trustworthy AI.

Author contributions

All authors equally contributed to the content and drafting of this article. All authors approved the submitted manuscript.

Declaration of competing interest

We confirm that there are no conflicts of interest to disclose.

References

[1] Stewart J, Sprivulis P, Dwivedi G. Artificial intelligence and machine learning in emergency medicine [Internet] EMA - Emerg Med Austr 2018 Dec 1. https://doi.org/10.1111/1742-6723.13145 [cited 2018 Dec 5];30(6):870–4..

[2] Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study [Internet] Digit Heal 2019 Jan 1. https://doi.org/10.1177/2055207619871808.

[3] Ienca M, Jotterand F, 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):1301–40.

[4] Ting DSW, Liu Y, Burlina P, Xu X, Bressler NM, Wong TY. AI for medical imaging goes deep. Nat Med 2018 May;24(5):539–40.

[5] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence [Internet] Nat Med 2019;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7.

[6] Nundy S, Montgomery T, Wachter RM. Promoting trust between patients and physicians in the era of artificial intelligence [Internet] JAMA 2019 Aug 13;322(6):497–8. https://doi.org/10.1001/jama.2018.20563.

[7] Winfield AFT, Jirotka M. Ethical governance is essential to building trust in robotics and artificial intelligence systems [Internet] Philos Trans R Soc A Math Eng Sci 2018 Nov 28 [cited 2019 Mar 1];376(2133):20180085. Available from, http://rsta.royalsocietypublishing.org/lookup/doi/10.1098/rsta.2018.0085.

[8] LaRosa E, Danks D. Impacts on trust of healthcare AI roles for healthcare AI. In: AAAI/ACM conference on artificial intelligence. Ethics, and Society; 2018.

[9] Gille F, Smith S, Mays N. Why public trust in health care systems matters and deserves greater research attention [Internet] J Health Serv Res Pol 2014;20(1):62–4. https://doi.org/10.1177/1355819614543161.

[10] Gille F, Smith S, Mays N. What is public trust in the healthcare system? A new conceptual framework developed from qualitative data in England [Internet] Soc Theory Heal 2020. https://doi.org/10.1057/s41285-020-00129-x.

[11] Ferretti A, Schneider M, Blasimme A. Machine learning in medicine [Internet] Eur Data Prot Law Rev 2018;4. https://doi.org/10.21552/edpl/2018/3/10.

[12] Department of Health & Social Care. Guidance: Code of conduct for data-driven health and care technology [Internet] [cited 2019 May 31]. Available from, https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology; 2019.

[13] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: automated decisions and the GDPR [cited 2020 Jun 2]; Available from, https://arxiv.org/abs/1711.00399; 2017 Nov 1.

[14] Krishnan M. Against interpretability: a critical examination of the interpretability problem in machine learning [Internet] Philos Technol 2019. https://doi.org/10.1007/s13347-019-00372-9.

[15] Mittelstadt B, Russell C, Wachter S. Explaining explanations in AI. In: Proceedings of the conference on fairness, accountability, and transparency [Internet]. New York, NY, USA: Association for Computing Machinery; 2019. p. 279–88.

[16] Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and AI research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness [cited 2020 May 26]; Available from, http://arxiv.org/abs/1812.10404; 2018 Dec 21.

[17] Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies [Internet] BMJ 2020. Mar 25;368:m689. Available from, http://www.bmj.com/content/368/bmj.m689.abstract.

[18] Polonski V. AI trust and AI fears: a media debate that could divide society. Oxford Internet Institute-University of Oxford; 2018.

[19] The Royal Society. Portrayals and perceptions of AI and why they matter. The Royal Society; 2018.

[20] Oh S, Kim JH, Choi S-W, Lee HJ, Hong J, Kwon SH. Physician confidence in artificial intelligence: an online mobile survey [Internet] J Med Internet Res 2019 Mar 25;21(3). e12422–e12422. Available from, https://pubmed.ncbi.nlm.nih.gov/30907742.

[21] Faes L, Liu X, Wagner SK, Fu DJ, Balaskas K, Sim DA, et al. A clinician's guide to artificial intelligence: how to critically appraise machine learning studies [Internet] Transl Vis Sci Technol 2020 Feb 12;9(2):7. https://doi.org/10.1167/tvst.9.2.7.

[22] Green J, Browne J. Principles of social research. Maidenhead: Open University Press; 2005. p. 172.

[23] Metzinger T. Ethics washing made in europe. Der tagesspiegel [Internet]. online. Available from, https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html; 2019 Apr 8.

[24] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines [Internet] Nat Mach Intell 2019;1(9):389–99. https://doi.org/10.1038/s42256-019-0088-2.

[25] U.S. Department of health and human services FDA center for drug evaluation and research, Department of health and human services FDA center for biologics evaluation and research, Department of health and human services FDA center for devices and radiological health. Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims [Internet]. Available from, http://www.fda.gov/downloads/Drugs/Guidances/UCM193282.pdf; 2009.

[26] Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, et al. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research [Internet] Qual Life Res 2013 Oct;22(8). 1889–905. Available from, http://link.springer.com/10.1007/s11136-012-0344-y.

[27] Gille F, Smith S, Mays N. Towards a broader conceptualisation of 'public trust' in the health care system [Internet] Soc Theory Heal 2017;15(1):25–43. https://doi.org/10.1057/s41285-016-0017-y.

[28] Luhmann N. Vertrauen: ein Mechanismus der Reduktion sozialer Komplexität. UTB für Wissenschaft Soziologie fachübergreifend. 4th Edition. Stuttgart: Lucius & Lucius; 2009.

[29] Frevert U. Vertrauensfragen - Eine Obsession der Moderne. Munich: C.H. Beck; 2013.