Beyond ‘Our product is trusted!’ – A processual approach to trust in AI healthcare

Michael Strange[1]

[1] *Department of Global Political Studies, Malmö University, Niagara, Nordenskiöldsgatan 1, Malmö, 250 06, Sweden*

Abstract

Trust in AI healthcare technologies is often treated as an obtainable end-state enforceable by regulation, in which developers can claim their product to be ‘trusted’. The article shows the limits of this approach, arguing instead for a processual understanding in which trust is understood to be dynamic and forever a state ‘to come’. The argument is developed by considering several types of trust relations amongst key stakeholders in AI healthcare, including where developers often distrust users. Drawing on political theory and Coactive Design, the article argues that trust relations as a negotiation are integral to a well-functioning design process that not only supports the moral acceptability of AI healthcare technologies but also their innovation and efficacy.

Keywords

Trust, Healthcare, Artificial Intelligence, Process, Coactive Design

1. Introduction

The application of artificial intelligence technologies to assist human health carers is subject to significant levels of hype – both positive claims it can save lives and cut costs, as well as negative warnings that it risks a corporate ‘sell-out’ as hospitals become increasingly dependent on large technological firms [1]. Yet, beyond the hype there is evidence that the integration of AI within healthcare can bring benefits from improved diagnostics, more data-based resource management, precision treatments, and reaching far beyond the clinical context to help enhance preventative treatments such as better lifestyle management [2]. Healthcare is a sensitive and high-stakes field. Consequently, trust is known to be a key factor impacting investment in development of AI healthcare technologies. Indeed, trust has emerged as a pivotal factor influencing the deployment and acceptance of AI in healthcare. However, recent trends indicate a significant decline in trust in general AI technology, as reported by the 2024 Edelman Trust Barometer, which notes a pervasive scepticism towards AI [3]. For example, in Sweden, a country often seen as a European leader in AI innovation, three-quarters of the population say they distrust AI. This

[1]* Corresponding author.
michael.strange@mau.se (M. Strange)
0000-0002-2903-7267 (M. Strange)

paper seeks to advance understanding on the conditions needed for trust in AI healthcare technologies by considering trust as a relational negotiation requiring participation between various stakeholders, including patients, their families, healthcare professionals, technology developers and engineers, patient associations, civil society, as well as many less easily identified actors but who nonetheless impact that process. As will be argued, trust relations need to be more consciously conceived as part of the design process in AI healthcare. Discussions on trust in AI healthcare typically centre around how patients (also sometimes just referred to as the 'public', 'consumers', or 'users') and healthcare practitioners relate to the technology, with an implicit assumption that they need to be persuaded to accept and adopt AI healthcare products. How individuals come to (dis)trust technology is a complex negotiation between multiple factors, from their past experiences, expected benefits of the technology, the reputation of the producer but also immediate provider, and a host of other variables often only indirectly connected to the technology itself. The paper is structured as follows. It first provides an overview of the factors driving distrust in AI healthcare technologies. To develop a relational understanding of trust in AI healthcare, the paper then turns to critical theoretical approaches on both democracy and human rights that highlight the ways in which these 'social goods' are, despite common parlance, not fixed states or even end goals, but are constituted as processes that remain constant if they are to have normative value. The same applies to trust in AI, including within its healthcare application. That perspective will be further developed through connecting to the Coactive Design model. The article then concludes.

2. Distrust in AI healthcare technologies

Healthcare is a sensitive field due to multiple factors including not only the obvious fact that it concerns life and death matters, but also the high cost of facilities coupled with often scarce resources in which technology fails have extremely serious and long-term consequences. Adopting a new technology like AI requires trust from a diverse range of stakeholders not limited to only healthcare professionals and patients. Trust impacts not only the willingness of actors to utilize AI but also how they experience the efficacy of those technologies. It is common to see trust in AI as concerning only how the technology is perceived by potential users, but this overlooks the wider relations in which the AI operates. If the context in which AI is both engineered and utilized is considered, that requires asking after a wider set of actors. For example, do developers trust users when they design AI? Are users seen as passive recipients or an integral part of a wider dynamic and ongoing relationship? In the field of healthcare, and for the purposes of simplicity, the paper considers four categories of actor: patients, healthcare professionals, society, and developers. These categories are simplifications but serve to illustrate the variety of trust relations at play within the implementation of AI technologies in healthcare.

2.1. Patients

Critical research on the impact of AI healthcare tools on patients covers several different aspects coalescing around fear of discrimination and privacy violations. Obermeyer et al's

[4] much cited study revealing the risk that AI locks-in racial data biases within healthcare has been joined by further research identifying other form of discrimination including gender that are potentially reinforced as AI usage becomes prevalent within healthcare [5]. In large part the problem of discrimination stems from already existing discrimination within healthcare, including unequal access to healthcare and underinvestment in, for example, women's health [6]. In addition, healthcare outcomes are known to largely depend on factors outside of the clinical context – the so-called 'societal determinants of health' – such as housing, job security, and education [7]. The societal determinants of health greatly complicate the roll-out of AI in healthcare because, unless controlled for in datasets or somehow else explicitly considered, the data used to train AI tools in healthcare will be skewed. This point is central to Obermeyer et al's study already mentioned, where they show how treating healthcare data separately of such complex – and often highly political – non-clinical factors when training AI tools will mean outputs suggest certain populations are best down prioritized due to having little likely benefit from expensive treatment despite the cause of such outcomes having not to do with the treatment's efficacy but, rather, the intervening variables of the population's poor housing or job precarity, for example. Whilst the point that such factors need to be controlled for when training AI may seem an obvious solution, the complexity and scale of data needed to train AI healthcare tools coupled with the difficulty of identifying precisely the role of societal determinants in healthcare means there is scepticism as to whether AI healthcare tools can be non-discriminatory.

The scale of data needed to train AI appears to necessitate mass transferability of patient data, with examples being developed by the EU, for example, that promise security. Yet, several recent and high-profile data breaches leave an open question as to whether such security is possible. For example, hacks of healthcare data in the UK impacting blood banks led to cancelled operations including heart surgery. Elsewhere in Europe, illicit access to and exposure of data from a psychologist's clinic led to black mail of former and present patients faced with the threat of sensitive information being made public. In both cases, the data breach occurred within a private firm sharing data with other healthcare organisations. AI healthcare tools typically rely on private firms collaborating, and sharing data, within a vast ecosystem of other actors that might not always be transparent to the individual patient. This was a problem during the pandemic, for example, where the sudden shift to online consultations meant some migrants in Europe no longer felt able to contact their local doctors [8]. Having lived in authoritarian regimes, such migrants associated digital sharing of sensitive of data with being controlled by actors hostile to their wellbeing. The fear of how health data might be used is also reflected in discussions over abortion law changes in the United States, with a credible concern that data from period tracking apps can be collected by the police to enforce those new restrictions. Whilst AI tools have potential to bring benefits to patients, there are also legitimate and rational reasons why patients and patient associations should distrust the use of AI in healthcare.

2.2. Healthcare professionals

The relationship between healthcare professionals and the system in which they work puts them in a position of authority and responsibility, whilst also being required to adhere to clearly specified and evidence-based professional standards. A challenge for healthcare professionals trusting AI systems in their work is the 'black box' nature of the technology, where the outputs of large data models cannot be easily explained or understood. It is likely that the average doctor or nurse cannot explain the detailed workings of the pharmaceuticals and other more traditional technologies that are intrinsic to their work, yet authoritative individuals within their professional environment are able to do that on their behalf to the satisfaction of being evidence-based. It is often remarked that advancements in machine learning have moved beyond the understanding of all but those most actively engaged in the design of those systems. And, increasingly, the patterns AI can see within data are said to be undecipherable to even that small elite. Any obstacle to explainability brings risk within a sensitive field like healthcare as it makes it harder to ascertain acceptable standards of liability, including both premeditating against potential harms as well as allocating blame and punitive costs should harm occur.

A common solution proposed for mitigating risk of harm from AI, in healthcare but also other sensitive contexts, is the so-called 'human-in-the-loop' in which a human operative remains the ultimate decision-maker reviewing AI outputs at key action stages. For example, whilst AI might propose a diagnosis, it is the human healthcare professional that formally makes the diagnosis and on which any treatment and communication to the patient is based. In that scenario, the AI is an information tool to be judged and utilized within the framework of human expertise. Yet, it is far from clear that healthcare professionals can remain active humans 'in-the-loop'. First, the blackbox nature of AI already discussed means healthcare professionals are being asked to take responsibility for output they cannot themselves explain. When faced with scarce resources and severe time constraints, doctors and nurses will be under pressure to reject or accept AI outputs without having sufficient overview of their merits. Second, there is also concern of the 'autopilot' effect in which reliance on AI tools in areas of their work traditionally left to human judgement will lead to a loss of their own abilities such that they have difficulties judging correctly when to step in and take back control. This is what Reubin Binns has called the 'human-in-the-loophole' where the notion of a human supervisor is used to downplay risks but, in practice, the humans involved feel unable to play that active role [9]. For healthcare professionals, AI may well help them better understand a patient's needs and save them time if it can process paperwork on their behalf but may also disconnect them from their role as experts offering evidence-based care.

2.3. Society

The design of healthcare systems varies globally with an obvious difference being the underlying funding model, including whether healthcare is provided on a 'free-at-point-of-use' basis as well as the role of the private for-profit sector. The use of AI in healthcare has far reaching political-economic consequences due to both the need for large datasets that often transcend national borders, but also the dominant structure through which AI technologies are developed. That structure includes both large corporations that have a near-monopoly over the cloud computing facilities needed to process big datasets used for AI healthcare technologies, as well as a highly skewed geopolitical pattern in which the United States and China are global leaders with most national contexts left in the position of consumers only. Many societies are distanced from both development and ownership of the technologies and underlying infrastructure needed for healthcare systems to use AI. Whilst EU politicians and policy documents speak of a need to protect its 'digital sovereignty' by, for example, trying to develop data processing centres within the EU space that are controlled by EU-based interests, it is far from achieving that goal and, for much of the rest of the world, such a target is unobtainable.

Through processing sensitive data but also becoming the channel by which healthcare data becomes accessed, AI in healthcare will often function as an infrastructure supporting other parts of the healthcare system as opposed to operating as a standalone tool. Increasing usage will make it harder to discard within annual procurement rounds, and where specific AI healthcare technologies become pervasive as infrastructures within a hospital they are likely to impact what and how other healthcare technologies may be utilized in that context. The roll-out of AI in healthcare comes at the same time where there is heightened uncertainty over how to fund public healthcare systems in Europe, for example, with increasing usage of for-profit care firms. Many countries with state-funded healthcare systems have seen significant shifts in state-market relations in recent years. The roll-out of a technology that is dominated by for-profit corporations and that takes on an infrastructural role in healthcare systems cannot avoid being part of that wider political-economic context. For society, AI healthcare technologies are therefore not neutral tools focused only on healthcare but relate to broader questions over how to best organize society. That is a fundamental and highly political question, greatly complicating society's levels of (dis)trust towards AI in healthcare.

2.4. Developers

How patients, healthcare professionals, and society relate to AI healthcare technologies and learn to (dis)trust those tools speaks to how users view the technology. However, trust relations also concern how those building AI systems in healthcare view the users and wider society. University courses for technology students will often include some aspect of participatory design focused on, for example, how to engage and incorporate user feedback. Yet, whilst feedback from users may be helpful, the rapid escalation in complexity that comes with AI technologies coupled with the sensitivities when applied to healthcare means there is a growing gap between the relative knowledge levels of developers and users. There is also very limited space allocated for users and the wider public to enter the design stage. Diversity within developer teams has dropped with fewer women getting to enter high-level decision-making positions within technology firms [10]. Anecdotally, many developers of AI and other advanced technological products are often sceptical towards involving the public within the design stage, seeing such engagement as a hindrance to, rather than a driver of, innovation. Approaching this in the context of trust relations, advancements in AI are leading to developers disconnecting from users and other stakeholders. This is arguably the most serious form of break within trust relations since it undermines attempts to build trust with the other categories of actor already discussed, whilst also further narrowing down the pool of individuals able to understand and impact AI healthcare technologies. If developers have difficulty forming trust relations towards other actors involved in healthcare, it is harder to ensure transparency and accountability in the development process.

3. Trust relations as a negotiation process in AI healthcare

Highlighted in the introduction, discussions on trust in AI healthcare are often guided by an implicit goal to persuade users – whether patients, healthcare professionals, or wider society – to accept and adopt AI healthcare products. That perspective treats trust as a one-way process, with AI developers (and owners) tasked to somehow convince users of the safety and efficacy of their products. However, this approach overlooks the complexity of trust as a complex and dynamic process that involves continuous negotiation between multiple stakeholders. Reframing trust as a negotiation requires recognizing that trust is not merely an outcome to be achieved but an integral part of the development and implementation of AI healthcare systems. Trust is shaped by a variety of factors, including past experiences, the perceived benefits of the technology, the reputation of the producer, and the credibility of the immediate provider. Moreover, trust is influenced by a host of other variables that are often only indirectly related to the technology itself, such as cultural values, social norms, and broader concerns. Central to this negotiation process is the space for dialogue between multiple stakeholders, which in healthcare includes not only patients but also their families, clinicians, device producers, as well as civil society such as patient associations.

Facilitating that negotiation is no easy task since it is shaped by two significant knowledge asymmetries between – first, healthcare professionals and patients; and second, AI developers and everyone else. The knowledge gap between healthcare professionals and patients is not new and, indeed, may also be said to give doctors and nurses the authority required to do their job. However, it does also create obstacles to communication and collaboration when that authority is questioned in times of uncertainty and rapid change, as we see with the sudden emergence of AI technologies. Given the challenges involved, then, is it possible to achieve the type of negotiation needed to build trust relations between key actors in AI healthcare, at what stage, and in what form? In response, the paper has no easy answers but offers some tentative thoughts by emphasizing the processual nature of negotiation and drawing on critical political theory.

3.1. Trust as a process rather than an obtainable goal

Treating trust as a goal that can be obtained implies a point at which it has been obtained. In other words, the goal of obtaining trust suggests a future statement in which one can claim 'our technology is trusted'. Jacques Derrida's work on democracy is pertinent here, taking a critical approach to claims that any nation-state is 'democratic' [11]. In contrast to those who treat democracy as a finished project, Derrida emphasized its processual character with the notion 'democracy to come' that drew out its ongoing, evolving character that, for it to exist, must always be open to contestation and reinterpretation. A similar emphasis on process over finality can be seen in the theoretical work of Illan Rua Wall [12] in relation to the concept of human rights. Wall argues against what he sees as the classic approach to human rights and its privileging of a legislative model, treating it also as a fixed state dependent on the law whilst ignoring the dynamic that gives it substance. Rather, Wall points to the necessarily seditious character of human rights – the demand for rights that go beyond what the authorities are willing to give – that means it should be seen first and foremost as a form of activism that, to be human rights, must always challenge the status quo to create space for justice and equity. Applying critical political theory of this nature to trust in AI healthcare, trust is not an end-state but must be continually renegotiated as new technologies emerge and the needs and interests of stakeholders change. Rather than being confined to legal and regulatory frameworks, trust-building must include multiple stakeholders in ways that meaningfully contest and shape how AI is utilized in healthcare.

3.2. Trust in AI healthcare requires a democratic grammar of conduct

Network governance models in policymaking offer another useful framework for understanding how trust can be fostered in the AI healthcare ecosystem, adding detail to how one might build the kind of 'trust to come' approach suggested above. Network governance speaks to the importance of involving diverse stakeholders in complex governance processes, recognizing that effective policymaking requires input from a wide range of perspectives. The concept of a democratic grammar of conduct [13] highlights the need for not only rules but also that all actors involved in a policy space adopt inclusive and participatory practices. Such practices must be learnt and mutually reinforced as part of an ongoing process of negotiation. In the context of technology ethics, Trevisan et al [14] have likewise argued the need to spend more time listening to problems rather than jumping into solutions. Specifically, they propose a deliberative method – consisting of literature reviews, surveys, expert interviews, and participatory workshops – where technology and policy designers work to embrace societal concerns. By first becoming aware of such concerns, designers are better able to understand the consequences of their work in ways that are necessary if sensitive AI developments like in healthcare are to be done appropriately.

3.3. Trust relations as a form of Coactive Design in AI Healthcare

There is overlap between seeing trust relations in AI healthcare as a negotiation and the type of negotiation contained within Matthew Johnson's model of 'Coactive Design' [15]. Acknowledging interdependence between humans and machines, Johnson challenges a simple reading of 'human-in-the-loop' and shows how in many cases it is unclear how much space there is for the human to be cognitively independent when utilizing advanced technology. If taken to its full extent, questioning the space for human independence from technology in a healthcare setting challenges the basis of medical liability as well as the authority of healthcare professionals. Johnson's Coactive Design model provides a practical framework for managing human-AI relations – so-called ‘hybrid intelligence’ – listing three key principles: observability, predictability, and directability. All three principles are essential for there to be a sustainable and productive negotiation between humans and AI. In the example of an AI-based medical assistant, Johnson illustrates these principles as follows. First, observability requires that the human knows which vitals (e.g. heart rate, blood sugar levels, etc) are assessed and how. Second, predictability means the human can anticipate what types of issues the AI might alert one to, as well as how it might direct a patient to other specialists if needed. Third, directability means the human has the option to ask the AI to include additional variables and factors in its analysis. Returning to the example of societal determinants in healthcare, it might be that a human healthcare practitioner is able to ask that AI to factor in those non-clinical variables into its analysis. Being able to observe, predict, and direct AI means that the human involved is an active interlocutor with the technology. Precisely what form of communication takes place within the human-AI negotiation will vary according to the relative knowledge levels and needs of the human, given different knowledge levels depending on if the human is a patient or highly educated healthcare professional. Either way, though, there should be a form of active negotiation and to achieve that requires building in those steps where it is possible for the human concerned to observe, predict, and direct.

4. Building trust relations in AI healthcare

Critical political theory on democracy and human rights helps us better understand trust in AI healthcare through drawing out the processual character of trust as a normative concept. Like those other lofty political goals, if treated as an end-state the quest for trust in AI healthcare becomes limited to a questionable endeavour of persuasion in which developers seek to claim their product is the ‘most trustworthy’. In contrast, if we ask how developers conduct their relations with various stakeholders, including how open they are to hear societal concerns, interest turns not to trust as an end-state but, instead, as an ongoing and dynamic process. Rather than asking how to persuade patients and healthcare professionals to use AI, attention is placed on what kind of relations are needed for there to be trust in this highly sensitive domain.

Trust is also seen not just as a factor impacting whether people use technology, but as part of the ecosystem in which the technology is developed and operated. Asking how much developers trust users, for example, speaks to some of the significant barriers that prevent wider societal engagement and diversity within the design process. The fields of healthcare and technological development are not separate from the societies in which they operate, impacting and being impacted by wider political and economic issues. However, to be aware of those issues and how they impact the efficacy of AI technologies in healthcare requires design processes that do not limit trust relations until only the point of implementation. Instead, building trust relations in AI healthcare means using those relations as part of the design process.

A Coactive Design approach when combined with the processual approach to trust argued for in the paper positions negotiation as necessary in the everyday operation of AI healthcare technology. In other words, we see trust relations in AI healthcare as requiring negotiation throughout design, adoption, and usage of the technology. That is not a moral add-on but, as Johnson's model shows, it is necessary to optimal functioning and efficacy of the technology.

5. Conclusion

Given the risks involved, healthcare is one of the most sensitive fields in which AI technology is being adopted, but also one in which there is considerable excitement with claims it will cut costs and improve outcomes by identifying conditions at points when treatment is cheaper and more effective, for example, as well as other benefits. Yet, such promise may well benefit only the already privileged depending on how AI is adopted in healthcare. To overcome potential worries that would slow the roll-out of AI healthcare technologies there is considerable focus on encouraging patients and healthcare professionals to trust it. The EU AI Act, for example, is part of that process to establish a basis for trust. However, as the paper argues, such efforts are mistaken where they treat trust as a fixed state that can be achieved through persuasion or regulation alone. Instead, the paper argues for a processual understanding of trust in which it is seen as a dynamic and ongoing negotiation requiring participation and collaboration between multiple stakeholders. Trust needs to be reframed as part of the design process, with trust relations empowering not only the moral acceptability of technology but as integral to innovation and, in healthcare, patient health and well-being.

References

[1] M. Strange, Three different types of AI hype in healthcare. AI Ethics (2024). doi:10.1007/s43681-024-00465-y.

[2] Z Ayaz, S. Naz, N.H. Khan, I Razzak & M Imran, Automated methods for diagnosis of Parkinson's disease and predicting severity level, Neural Computing and Applications, 35(20), 14499-14534. doi:10.1007/s00521-021-06626-y. B. Hunter, S. Hindocha, & R.W. Lee, The role of Artificial Intelligence in early cancer diagnosis, Cancers, 14(6), 1524. doi:10.3390/cancers14061524.

[3] Edelman Trust Institute, Edelman Trust Barometer Global Report, 2024. URL: https://www.edelman.com/trust/2024/trust-barometer.

[4] Z. Obermeyer, B. Powers, C. Vogeli, & S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science (2019), 447-453. doi: 10.1126/science.aax2342.

[5] H. Chung, C. Park, W.S. Kang, & J. Lee, Gender Bias in Artificial Intelligence: Severity Prediction at an Early Stage of COVID-19, Frontiers in Physiology (2021). doi:10.3389/fphys.2021.778720. M. Mittermaier, M. M. Raza, & J.C. Kvedar, Bias in AI-based models for medical applications: challenges and mitigation strategies, NPJ Digital Medicine (2023) 6(1):113. doi:10.1038/s41746-023-00858-z.

[6] C. Criado Perez, Invisible Women: Exposing Data Bias in a World Designed for Men, Vintage Press, 2020.

[7] World Health Organization, Social determinants of health, 2023. URL: www.who.int/health-topics/social-determinants-of-health.

[8] M. Strange, H. Gustafsson, E. Mangrio, & S. Zdravkovic, Report #1 PHED Commission on the future of healthcare post covid-19: Social inequity makes us vulnerable to pandemics. PHED Commission on the Future of Healthcare Post Covid-19, 2021. URL: https://phed.uni.mau.se/phed-commission-reports.

[9] R. Binns, AI and the human in the loophole, 2022. URL: https://www.youtube.com/watch?v=XkU1-AHG1qk.

[10] S. White, Women in tech statistics: The hard truths of an uphill battle, March 8th 2024. URL: https://www.cio.com/article/201905/women-in-tech-statistics-the-hard-truths-of-an-uphill-battle.html.

[11] J. Derrida, The Politics of Friendship, tr. G. Collins, London: Verso, 1997.

[12] I. R. Wall On a radical politics for human rights. In: Douzinas C, Gearty C (eds) The Meanings of Rights: The Philosophy and Social Theory of Human Rights. Cambridge: Cambridge University Press, 2014, pp.106–120.

[13] E. Sørensen and J. Torfing, The Democratic Anchorage of Governance Networks. Scandinavian Political Studies (2005), 28: 195-218. doi.org/10.1111/j.1467-9477.2005.00129.x.

[14] F. Trevisan, P. Troullinou, D. Kyriazanos, D. *et al,* Deconstructing controversies to design a trustworthy AI future. Ethics and Information Technology (2024) 26(35), doi.org/10.1007/s10676-024-09771-9.

[15] M. Johnson, M., & A. Vera No AI Is an Island: The Case for Teaming Intelligence. AI Magazine (2019), 40(1), 16-28. doi:10.1609/aimag.v40i1.2842. M. Johnson & J. M. Bradshaw, The Role of Interdependence in Trust. In Trust in Human-Robot Interaction (2020) (pp. 379-403). Academic Press.