Towards a Framework for Trust in Clinical AI: Expanding the Unit of Analysis

JACOB T BROWNE

Philips Experience Design, jacob.browne@philips.com

SASKIA BAKKER

Philips Experience Design, saskia.bakker@philips.com

BIN YU

Philips Experience Design, bin.yu@philips.com

JEROEN RAIJMAKERS

Philips Experience Design, jeroen.raijmakers@philips.com

JON PLUYTER

Philips Experience Design, jon.pluyter@philips.com

NICK RUIJS

Philips Experience Design, nick.ruijs@philips.com

EVA DECKERS

Philips Experience Design, eva.deckers@philips.com

From diagnosis to patient scheduling, AI is being considered across different clinical applications. Despite increasingly powerful clinical AI, uptake into actual clinical workflows remains limited. One of the major challenges is developing appropriate trust with clinicians. In this paper, we investigate trust in clinical AI in a wider perspective beyond the user's interactions with the AI. We offer several points in the clinical AI development, usage, and monitoring process that can have a significant impact on the clinician's trust. Future work will look at how these points affect trust and how to calibrate trust at these different points.

CCS CONCEPTS **•Human-centered computing~Human computer interaction (HCI)~HCI theory, concepts and models**

Additional Keywords and Phrases: Trust, Clinical AI

ACM Reference Format:

Jacob T Browne, Saskia Bakker, Bin Yu, Jeroen Raijmakers, Jon Pluyter, Nick Ruijs, Eva Deckers. 2022. Towards a Framework for Trust in Clinical AI: Expanding the Unit of Analysis. In CHI Conference on Human Factors in Computing Systems, Workshop on Trust and Reliance in AI-Human Teams (CHI '22), April 29–May 05, 2022, New Orleans, LA, USA. ACM, New York, NY, USA, 10 pages.

1 INTRODUCTION

AI is increasingly being considered across different healthcare areas: from patient facing applications to clinical workflow enhancements [14, 69]. AI can enable healthcare providers towards realizing what's known as the "quadruple aim": improved patient experience, better health outcomes, improved staff experience, and lower cost of care [7]. AI has the potential to give doctors more time in engaging with patients by taking care of the menial, non-critical tasks [68, 71, 74]. The promise of AI for healthcare is rife with hype, with many machine learning models outperforming clinician performance in diagnosis in controlled settings, spiking some concerns of professional autonomy among clinicians [25, 27, 54, 64, 71].

The major focus of the AI community has been on crafting better performing models, rather than the effects of AI implementation in the wild and challenges associated with adoption [63, 66]. Despite ever better AI models, the adoption of AI into actual clinical practice is limited [74]. Trust remains a critical challenge in deploying clinical AI, being influenced by many factors (user education, experience, bias, system controllability, complexity, risks, etc.) [2, 5]. Even if the AI system is deployed into an existing workflow, whether the clinician trusts it enough to adopt in usage remains a challenge [31].

Research investigating how trust is formed in clinical AI in the wild is lacking. Many studies investigating trust in AI-assisted decision making do not go beyond evaluations of the AI's UI in a laboratory setting (often focused on XAI). Vereschak et al., in a systematic review of methodologies to study trust in AI-assisted decision making, call for better methods to investigate trust that are more ecologically sound [70]. Similarly, in a survey of AI-decision making research, Lai et al. call for more investigations of AI beyond discrete, laboratory trials [39].

This paper makes two contributions: an expansion of the unit of analysis of trust in clinical AI beyond XAI and a review of relevant points of analysis within the clinical AI development process. The distributed processes of trusting within clinical AI are rendered as points of investigation for future work.

2 EXPANDING TRUST

Trust is a multifaceted, multidisciplinary, and challenging theoretical concept to study and define, with many definitions from different fields abounding [42, 48, 60]. Despite this, there are some common, important considerations that bind them. To define trust, we follow Vereschak et al.'s systematic review of trust definitions within Human-AI literature [70]. They cement Lee and See's seminal definition of trust as compromising all the relevant elements of trust (vulnerability, positive expectations, and trust rendered as an attitude). Trust is "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [42, 70].

In an integration of empirical research on trust, Hoff and Bashir build upon three variables of trust: dispositional, situational, and learned trust [31, 50]. They render trust as being influenced by multiple factors: culture, age, personality, type of system used, difficulty of the task, workload, team dynamic, decisional freedom, etc. Similarly, Chiou and Lee advocate for a broader analysis of trust to encompass the sociotechnical aspects of trusting [13]. In recent work on human-robot teaming, Huang et al. define a Distributed Dynamic Team Trust (D2T2) model, establishing trust as "a distributed, networked state that is constantly in flux" [32]. This framework includes interpersonal and technical factors that relate to trust in a dynamic, transitive light. In other work, de Visser et al. emphasize the longitudinal, relationship-like nature of trust, offering methods for trust dampening and repairing [16, 17].

These directions in trust research offer potential expansions of our current research of trust in clinical AI. There is an expansion of the unit of analysis beyond the individual and AI at one moment in time. As Möllering argues, "people's trust should be conceptualized and operationalized as a continuous process of forming and reforming the attitudes static surveys have measured so far and, crucially, as part of larger social processes" [53]. What's often missing in investigations of trust in clinical AI is the cultural and organizational processes in which clinical AI is implemented. Integrating AI into a clinical workflow involves a complex web of stakeholders and processes [18]. Instead, studies tend to focus on individual interactions of a clinician and a prototype. While these investigations are valuable, they dodge the larger picture of how trusting occurs. The process of trusting spans longer time horizons, distributed across time, material, and social worlds. In the next section, we'll use recent work in deploying clinical AI to showcase this.

3 EXPANDING THE UNIT OF ANALYSIS OF TRUST IN CLINICAL AI

Much of the research on trust in clinical AI focuses on measures of trust during the initial interaction with a prototype, often through some representation of XAI [29, 61]. Rarely is the case of trust before the prototype considered crucial [63]. However, as evidenced by studies from practitioners developing clinical AI, the process of trusting begins way before clinicians ever see an interface. We need to investigate trust in clinical AI beyond what is most apt to the domain of HCI and instead consider the larger complexities of integrating into a clinical environment. We'll consider 4 areas of trust development: pre-development, during development, during usage, and during monitoring.

3.1 Pre-Development

3.1.1 Personal Differences in How Clinicians Trust

Even before AI development begins, trust formation can occur. People have different inclinations to trust, often known as dispositional trust [6, 31, 67]. Culture, age, attachment styles, and other personal differences all count towards this dispositional trust [26, 31]. These differences can greatly affect trust and reliance in ways not related to the properties of the AI [42]. Henry et al. found that provider characteristics were closely associated with the likelihood of evaluating an alert from clinical AI: those who have used it before we're more likely to use it again [30]. Clinicians may have positive expectations (or lack thereof) in clinical AI, informed by their past experiences, culture, expertise, gossip, relevant industry news, mental models, affect etc. [6, 12, 70]. These differences are rarely considered when investigating trust in clinical AI, despite having important implications for how to calibrate appropriate trust.

3.2 During Development

3.2.1 Trust Through the AI Development Team

Trust also develops during the development of the AI, starting as early as the assembling of the team developing the AI. Releasing clinical AI is an extensive, complex process [15, 18]. The development team having the right professional credibility and experience to engage in this process indicates positive expectations being formed: that a negative outcome will be unlikely with these stakeholders [44, 56]. Clinicians may trust an AI simply because they value the brand based on prior experiences or cultural approval of a brand.

3.2.2 Trust Through the Training of the AI

Once the development of the AI begins, the training the selection of the AI's dataset, training, and metrics influence trust [72]. Reviewing the inputs of the AI with clinicians can help foster trust and catch quality issues [6]. Cai et al. note that some pathologists wanted to know the "quantity and diversity of the training data" to understand the generalizability and capacities of the AI within the clinician's local context [10]. Pathologists wouldn't trust the AI unless it were trained on judgements made by well-respected clinicians [10]. Pathologists insisted on knowing “a summary of the volume and types of clinical cases that the algorithm was created from” and “from diverse sources would be more representative” [10].

3.2.3 Trust Through Clinical Involvement

The degree of clinical involvement has an impact on trust in several ways [6]. Firstly, the better the development team’s understanding of the clinical context, the better the outcome will be for integration into the workflow, the better their understanding of clinician’s mental models, and thus, higher positive expectations from the clinical team [13, 33, 40]. Cai et al. found that dissimilar mental models between pathologists and the AI system degraded trust [9].

Secondly, more clinical involvement means clinicians will have more of a stake in designing the task allocation or division of labor of the AI [33, 47]. As a side effect of these sessions, the clinical team can further understand the purpose of the AI, how to use it, and its limitations by being active participants in the design process. They can start to calibrate their trust in the AI before any actual performance with it and prevent misuse [6, 42].

Lastly, the AI team will be able to develop relationships with the clinical team. Sendak et al. offer us a profound insight: “trust in a technology is rooted in relationships - not in a technical specification or feature” [63]. Trust is developed through developing relationships with the clinical team: meeting with the clinicians and staff involved throughout the process of development and integration. Without these relationships, it will be much more difficult to garner trust. This incorporating of different users into the development process will help create a better understanding of how to develop trust at their local sites [6]. Barda et al. emphasize that different needs arise from different clinical stakeholders, especially when considering XAI [3].

3.2.4 Trust Through Augmentation, Not Automation

Different levels of automation and the level of control the operator has impacts trust [31]. Implicit in this agreement from clinicians to the AI development team is that the AI would not be developed to replace them. If the AI developers wanted to replace the clinicians, the clinician likely would not hold that as a positive expectation of the use of AI, nor would this be an ethical, productive endeavor [57]. As Gichoya et al. argue, a more productive focus is upon the clinician-AI team (rather than replacing clinicians) given the complexities of the clinical context [25]. Kiani et al. found that there was an increased performance in the joint teaming of a pathologist and a deep learning-based assistant in the histopathologic classification of liver cancer [36]. Lee et al. found a similar improvement in therapists using AI in a rehabilitative assessment context [41]. Wang et al. found clinicians want to maintain decisional power over the AI and verify any decisions it would make [71]. They also found that clinicians didn’t believe AI could replace them, as one participant stated, “you will have to stay in medical school for 3 years in order to understand this [clinical decision support system].” [71]. Instead, the diagnosis process is a “highly interactive, communicative, and social event”, thus not up for automation anytime soon [71].

To protect autonomy and agency in clinicians, Cai et al. found it crucial to give clinicians tools to refine the system, allowing the clinician to improve the AI, remain in control, and disambiguate mistakes [9]. This is further evidenced by Strohm’s interviews with radiologists regarding AI integration, where radiologists were having to “reframe their professional identity and responsibilities…framing AI applications as “co-pilots” enabling radiologists to perform better while staying in control.” [65]. This is like airplane automation, where pilots need to be trained on how to be better monitors of automation [11]. Clinicians need to be reminded that while they are working in an environment with AI agents, they are responsible for their decisions.

3.2.5 Trust Through a Clinical Champion

Clinical and AI development teams working together are often accompanied by a clinical champion, a key to gaining the positive expectations of the clinical staff [65]. For instance, Lu et al. emphasize the need for a clinical champion, or someone to affirm the clinical utility of the AI system and promote the project within the "complex social hierarchies and regulations… that would be impenetrable to outsiders" [46]. Strohm et al. point out that these insiders share information about AI to other clinicians and promote opportunities for experimentation [65].

3.2.6 Trust Through Clinical Trials and Peer Reviewed Publications

Prior to releasing AI in a clinical workflow, these models need to be evaluated according to rigorous clinical standards [15, 74]. For instance, Sendak's team had both the sepsis definition and model peer-reviewed and disseminated in clinical and technical venues [22, 23, 45]. Sendak et al. tailored different ways to convey trustworthiness to clinicians depending on what was meaningful to them, using a form of model card and model performance presentations during meetings [52, 63]. Cai et al. also indicate the need for "evidence of FDA approval and published validation in peer-review journals, social endorsement by well-respected medical leaders" [10]. In co-designing a clinical DST, Jacobs et al. found that clinicians would primarily trust the AI based on its validation through randomized controlled trials and endorsement from other clinicians, rather than at different decision points in use [33]. This finding dovetails nicely with recent work deflating the importance of XAI in clinical AI, notably by Ghassemi et al., who advocate more for rigorous external and internal validation of models [24].

3.2.7 Trust Through Public Accountability

A form of public accountability during trial phases can also affect trust. Although seemingly rare in the literature, Sendak et al. offer an account of this [63]. Sendak et al. enabled mechanisms of public accountability by conducting a clinical trial with specified goals and outcomes, combined with an external data safety monitoring board to oversee the safety and efficacy of the system [63]. This enabled positive expectations to be built and exchanges of vulnerability between the clinical and development team.

3.3 Use in context

3.3.1 Trust Through Training and Onboarding Sessions

Training sessions with clinicians also modulate trust: both active training on how to use it and allowing clinicians to test the AI out, understand it's limitations, often through onboarding [15, 31, 35]. This onboarding is essential: developers are introducing a sociotechnical system, transforming the complex distributed workflow of clinicians. During initial onboarding and throughout usage, explanations of model predictions, transparency into higher level objectives, global behavior, and tendencies can be needed [10, 31, 62]. First impressions with AI systems greatly influence later trust development [13, 31, 55, 67].

Wang et al. found that a lack of training lessened trust in the AI, as clinicians had to learn alone how to use and understand the system [71]. Henry et al. discuss how in a sepsis alert system, clinicians might dismiss the alert if there aren't clear signs of sepsis, and the patient has a less common presentation of sepsis [30]. Training clinicians on how the system could detect sepsis despite the typical warning signs would be crucial [30]. This training and onboarding are crucial forms of trust calibration [13].

3.3.2 Trust Through Performance

As clinicians use the AI in practice, their trust will modulate dynamically based on system performance and different sociotechnical factors [31]. The goal is to calibrate appropriate trust in real-time through trust dampening and repair mechanisms to increase human-AI teamwork performance. This is where further investigations of trust in clinical AI are needed: how does trust develop over time in actual usage [35]?

Levy et al. emphasize the importance of long-term investigations of AI's impact on trust [43]. How the AI functions within the workflow has a large impact on trust and reliance. How well does the AI work with live clinical data, in an actual clinical workflow over time? Does the AI miss new edge cases arising within clinical contexts [49]? Was the data used in training similar to how it's gathered in this live clinical context? What unintended consequences arise [8, 66]? How do users accept improper results, lose engagement in the task, or take less initiative [43]?

Given appropriate trust is built upon having knowledge of the AI's performance, it's important to inform clinicians on the AI's past performance [21, 35, 70, 75]. The AI could perform better on certain tasks than others (e.g., diagnosis of different types of lesions) and the clinician would need to adjust their trust as these cases arose [35]. Further, specific environmental contexts can cause the AI to not perform well and reduce trust [35, 42, 58, 71]. For instance, Beede et al. found that poor image quality severely impacted usage of their deep learning system that made assessments based on the image of the eye [4]. How the system responds to such failures can have an effect whether trust is repaired (e.g., how the AI expresses regret) [38].

How are trust dampening and trust repairing mechanisms released as needed, and do they even work as intended [16]? Trust repair would repair trust after a trust violation, while trust dampening would lower expectations as needed [16]. How people respond to different trust repair strategies varies person and context [17, 19]. De Visser argues that designers ought to use different repair strategies for different respective violations and to be mindful of how different contexts and timings affect trust repair strategies [17]. McDermott refers to calibration points, or points in time where AI performance is degraded or improved, and trust needs to be increased or dampened [51]. Through the human-centered design process, designers could know what information the system needs to show at different calibration points (e.g., confidence scores after analyzing an image) [34, 35].

Trust during actual usage will also be influenced by other stakeholders in the clinical team. Clinicians will have varying degrees of experience with the AI: some holding positive expectations, others dismissing the AI because of adverse situations. Each stakeholder will have different predispositions to trusting AI and different levels of meta trust with each other, or: "the trust a person has that the other person's trust in the automation is appropriate" [42]. These networks of trust relations will be impacted by each other [32]. Whether or not a trusted, senior clinician trusts an AI will influence the rest of the team's trust in the AI [28]. Similarly, clinicians may be required to interact with an EHR system more, increasing situational trust [30]. The implementation of AI will have a broader effect on trust between healthcare professionals, and in turn, will affect trust in the AI [59].

3.4 Auditing and monitoring

3.4.1 Trust Through Continual Monitoring

The development team needs to continually monitor performance to ensure the clinical AI is performing effectively [15]. The "you build it, you own it" mentality by Sendak et al. 's team creates positive expectations that the AI will be improved as new information is surfaced based on real clinical data [63]. Medical procedures and best practices are similarly dynamic. Does the AI reflect up to date knowledge of clinical practice [49]? The AI will have systemic, emergent, impossible to predict effects upon the sociotechnical context of the clinical setting and this needs to be observed continuously [49]. Feedback from clinicians after deployment will be critical to maintaining relationships and maintaining trust [20].

4 CONCLUSION

We need empirical research investigating trust building in clinical AI as it occurs in the wild [63, 73]. In this paper, we outlined several points in which the development, usage, and continual monitoring of a clinical AI project can affect trust. As Chiou and Lee mention, we must go beyond "necessary but insufficient guidelines" that argue for transparency (e.g., "make clear what the system can do"), and instead focus on guidelines for how trust is developed across interactions and situations [1, 13]. Indeed, trust cannot be an afterthought, but rather it must be a central design aim of the project [37]. Future work will more closely examine how each of these points affect trust and how to calibrate appropriate trust at each stage.

ACKNOWLEDGMENTS

This work is part of the DCODE project. The project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955990.

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* Association for Computing Machinery, New York, NY, USA, Paper 3, 1–13. DOI:https://doi.org/10.1145/3290605.3300233

[2] Sahil Sandhu, Anthony L. Lin, Nathan Brajer, Jessica Sperling, William Ratliff, Armando D. Bedoya, Suresh Balu, Cara O'Brien, and Mark P. Sendak. 2020. Integrating a Machine Learning System Into Clinical Workflows: Qualitative Study. J Med Internet Res 2020;22(11):e22421. DOI: 10.2196/22421

[3] Amie J. Barda, Christopher M. Horvat, and Harry Hochheiser. 2020. A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. BMC Med Inform Decis Mak 20, 257 (2020). https://doi.org/10.1186/s12911-020-01276-x

[4] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. DOI:https://doi.org/10.1145/3313831.3376718

[5] Natalie C. Benda, Lala Tanmoy Das, Erika L. Abramson, Katherine Blackburn, Amy Thoman, Rainu Kaushal, Yongkang Zhang, and Jessica S Ancker. 2020. "How did you get to this number?" Stakeholder needs for implementing predictive analytics: a pre-implementation qualitative study. J Am Med Inform Assoc. 2020;27(5):709-716. doi:10.1093/jamia/ocaa021

[6] Natalie C Benda, Laurie L Novak, Carrie Reale, and Jessica S Ancker. 2022. Trust in AI: why we should be designing for APPROPRIATE reliance, Journal of the American Medical Informatics Association, Volume 29, Issue 1, January 2022, Pages 207–212, https://doi.org/10.1093/jamia/ocab238

[7] Thomas Bodenheimer and Christine Sinsky. 2014. From triple to quadruple aim: care of the patient requires care of the provider. Ann Fam Med. 2014;12(6):573-576. doi:10.1370/afm.1713

[8] Federico Cabitza, Raffaele Rasoini, Gian Franco Gensini. 2017. Unintended Consequences of Machine Learning in Medicine. JAMA. 2017;318(6):517–518. doi:10.1001/jama.2017.7797

[9] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 4, 1–14. DOI:https://doi.org/10.1145/3290605.3300234

[10] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 104 (November 2019), 24 pages. DOI:https://doi.org/10.1145/3359206

[11] Stephen M. Casner and Edwin L Hutchins. 2019. What Do We Tell the Drivers? Toward Minimum Driver Training Standards for Partially Automated Cars. Journal of Cognitive Engineering and Decision Making, 13(2), 55–66. https://doi.org/10.1177/1555343419830901

[12] Shih-Yi Chien, Michael Lewis, Katia Sycara, Asiye Kumru, and Jyi-Shane Liu. 2020. Influence of Culture, Transparency, Trust, and Degree of

Automation on Automation Use, in IEEE Transactions on Human-Machine Systems, vol. 50, no. 3, pp. 205-214, June 2020, doi: 10.1109/THMS.2019.2931755.

[13] Erin K. Chiou and John D. Lee. 2021. Trusting Automation: Designing for Responsivity and Resilience. Human Factors, Apr. 2021, doi:10.1177/00187208211009995.

[14] Thomas Davenport and Ravi Kalakota. 2019. The potential for artificial intelligence in healthcare. Future Healthcare Journal 6, 2: 94–98. https://doi.org/10.7861/futurehosp.6-2-94

[15] Anne A. H. de Hond, Artuur M. Leeuwenberg, Lotty Hooft, Ilse M. J. Kant, Steven W. J. Nijman, Hendrikus J. A. van Os, Jiska J. Aardoom, Thomas P. A. Debray, Ewoud Schuit, Maarten van Smeden, Johannes B. Reitsma, Ewout W. Steyerberg, Niels H. Chavannes and Karel G. M. Moons. 2022. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. npj Digit. Med. 5, 2 (2022). https://doi.org/10.1038/s41746-021-00549-7

[16] Ewart de Visser, Marieke M.M. Peeters, Malte Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark Neerincx. 2020. Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. International Journal of Social Robotics. 12. 10.1007/s12369-019-00596-x.

[17] Ewart J de Visser, Richard Pak, and Tyler H Shaw. 2018. From 'automation' to 'autonomy': the importance of trust repair in human-machine interaction. Ergonomics. 2018;61(10):1409-1427. doi:10.1080/00140139.2018.1457725

[18] Madeleine Clare Elish. 2018. The Stakes of Uncertainty: Developing and Integrating Machine Learning in Clinical Care. Ethnographic Praxis in Industry Conference Proceedings, 2018: 364-380. https://doi.org/10.1111/1559-8918.2018.01213

[19] Md Abdullah Al Fahim, Mohammad Maifi Hasan Khan, Theodore Jensen, Yusuf Albayram, and Emil Coman. 2021. Do Integral Emotions Affect Trust? The Mediating Effect of Emotions on Trust in the Context of Human-Agent Interaction. Designing Interactive Systems Conference 2021. Association for Computing Machinery, New York, NY, USA, 1492–1503. DOI:https://doi.org/10.1145/3461778.3461997

[20] Ross W Filice and Raj M Ratwani. 2020. The Case for User-Centered Artificial Intelligence in Radiology. Radiology. Artificial intelligence vol. 2,3 e190095. 13 May. 2020, doi:10.1148/ryai.2020190095

[21] Anna-Katharina Frison, Philipp Wintersberger, Andreas Riener, Clemens Schartmüller, Linda Ng Boyle, Erika Miller, and Klemens Weigl. 2019. In UX We Trust: Investigation of Aesthetics and Usability of Driver-Vehicle Interfaces and Their Impact on the Perception of Automated Driving. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* Association for Computing Machinery, New York, NY, USA, Paper 144, 1–13. DOI:https://doi.org/10.1145/3290605.3300374

[22] Joseph Futoma, Sanjay Hariharan, and Katherine Heller. 2017. Learning to detect sepsis with a multitask Gaussian process RNN classifier. In Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17). JMLR.org, 1174–1182.

[23] Joseph Futoma, Sanjay Hariharan, Katherine Heller, Mark Sendak, Nathan Brajer, Meredith Clement, Armando Bedoya, and Cara O'brien. 2017. An Improved MultiOutput Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection. Technical Report. arXiv:1708.05894v1

[24] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 3, 11 (2021), e745–e750. DOI:https://doi.org/10.1016/S2589-7500(21)00208-9

[25] Judy W. Gichoya, Siddhartha Nuthakki, Pallavi G. Maity, and Saptarshi Purkayastha. 2021. Phronesis of AI inradiology: superhuman meets natural stupidity. arXiv. Available at:https://arxiv.org/abs/1803.11244. Accessed Feb 7, 2021

[26] Omri Gillath, Ting Ai, Michael S. Branicky, Shawn Keshmiri, Robert B. Davison, and Ryan Spaulding. 2021. Attachment and trust in artificial intelligence. Comput. Hum. Behav. 115, C (Feb 2021). DOI:https://doi.org/10.1016/j.chb.2020.106607

[27] Bo Gong, James P Nugent, William Guest, William Parker, Paul J Chang, Faisal Khosa, and Savvas Nicolaou. 2018. Influence of Artificial Intelligence on Canadian Medical Students' Preference for Radiology Specialty: A National Survey Study. Acad Radiol. 2019;26(4):566-577. doi:10.1016/j.acra.2018.10.007

[28] Trisha Greenhalgh, Glenn Robert, Fraser Macfarlane, Paul Bate, and Olivia Kyriakidou. 2004. Diffusion of innovations in service organizations: systematic review and recommendations. Milbank Q. 2004;82(4):581-629. doi:10.1111/j.0887-378X.2004.00325.x

[29] Jianxing He, Sally L Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. 2019. The practical implementation of artificial intelligence technologies in medicine. Nat Med. 2019;25(1):30-36. doi:10.1038/s41591-018-0307-0

[30] Katharine E. Henry, Roy Adams, Cassandra Parent, Anirudh Sridharan, Lauren Johnson, David N. Hager, Sara E. Cosgrove, Andrew Markowski, Eili Y. Klein, Edward S. Chen, Maureen Henley, Sheila Miranda, Katrina Houston, Robert C. Linton II, Anushree R. Ahluwalia, Albert W. Wu, and Suchi Saria. 2021. Evaluating Adoption, Impact, and Factors Driving Adoption for TREWS, a Machine Learning-Based Sepsis Alerting System. 10.1101/2021.07.02.21259941.

[31] Kevin Anthony Hoff and Masooda Bashir. 2014. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. Human Factors 57, no. 3 (May 2015): 407–34. https://doi.org/10.1177/0018720814547570.

[32] Lixiao Huang, Nancy J. Cooke, Robert S. Gutzwiller, Spring Berman, Erin K. Chiou, Mustafa Demir, and Wenlong Zhang. 2020. Distributed dynamic team trust in human, artificial intelligence, and robot teaming. In C. S. Nam, & J. B. Lyons (Eds.), Trust in Human-Robot Interaction (pp. 301-319). Academic Press. https://doi.org/10.1016/B978-0-12-819472-0.00013-7

[33] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 659, 1–14. DOI:https://doi.org/10.1145/3411764.3445385

[34] Theodore Jensen. 2021. Disentangling Trust and Anthropomorphism Toward the Design of Human-Centered AI Systems". In Artificial Intelligence in HCI, 41–58. Cham: Springer International Publishing, 2021. http://dx.doi.org/10.1007/978-3-030-77772-2_3.

[35] Wiard Jorritsma, Fokeltje Cnossen, Peter M.A. van Ooijen. 2015. Improving the radiologist-CAD interaction: designing for appropriate trust. Clin

Radiol. 2015;70(2):115-122. doi:10.1016/j.crad.2014.09.017

[36] Amirhossein Kiani, Bora Uyumazturk, Pranav Rajpurkar, Alex Wang, Rebecca Gao, Erik Jones, Yifan Yu, Curtis P. Langlotz, Robyn L. Ball, Thomas J. Montine, Brock A. Martin, Gerald J. Berry, Michael G. Ozawa, Florette K. Hazard, Ryanne A. Brown, Simon B. Chen, Mona Wood, Libby S. Allard, Lourdes Ylagan, Andrew Y. Ng, and Jeanne Shen. 2020. Impact of a deep learning assistant on the histopathologic classification of liver cancer. npj Digit. Med. 3, 23 (2020). https://doi.org/10.1038/s41746-020-0232-8

[37] Bran Knowles, Mike Harding, Lynne Blair, Nigel Davies, James Hannon, Mark Rouncefield, and John Walden. 2014. Trustworthy by design. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW '14). Association for Computing Machinery, New York, NY, USA, 1060–1071. DOI:https://doi.org/10.1145/2531602.2531699

[38] E. S. Kox, J. H. Kerstholt, T. F. Hueting, and P. W. de Vries. 2021. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. Auton Agent Multi-Agent Syst 35, 30 (2021). https://doi.org/10.1007/s10458-021-09515-9

[39] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. arXiv preprint arXiv:2112.11471 (2021).

[40] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Co-Design and Evaluation of an Intelligent Decision Support System for Stroke Rehabilitation Assessment. Proc. ACM Hum.-Comput. Interact. 4, CSCW2, Article 156 (October 2020), 27 pages. DOI:https://doi.org/10.1145/341522

[41] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2021. A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 392, 1–14. DOI:https://doi.org/10.1145/3411764.3445472

[42] John Lee and Katrina See. 2004. Trust in Automation: Designing for Appropriate Reliance. Human factors. 46. 50-80. 10.1518/hfes.46.1.50.30392.

[43] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 72, 1–13. DOI:https://doi.org/10.1145/3411764.3445522

[44] J. David Lewis and Andrew Weigert. 1985. Trust as a Social Reality. Social Forces63, 4 (1985), 967–985. http://www.jstor.org/stable/2578601

[45] A.L. Lin, M. Sendak, A. Bedoya, M. Clement, J. Futoma, M. Nichols, M. Gao, K. Heller, C. O'Brien. 2018. What Is Sepsis: Investigating the Heterogeneity of Patient Populations Captured by Different Sepsis Definitions. American Journal of Respiratory and Critical Care Medicine. (May 2018).

[46] Charles Lu, Ken Chang, Praveer Singh, Stuart Pomerantz, Sean Doyle, Sujay Kakarmath, Christopher Bridge, Jayashree Kalpathy-Cramer. 2022. Deploying clinical machine learning? Consider the following…. arXiv. https://doi.org/10.48550/arXiv.2109.06919

[47] David Lyell, Enrico Coiera, Jessica Chen, Parina Shah, and Farah Magrabi. 2021. How machine learning is embedded to support clinician decision making: an analysis of FDA-approved medical devices. BMJ health & care informatics, 28(1), e100301. https://doi.org/10.1136/bmjhci-2020-100301

[48] Fergus Lyon, Guido Möllering, and Mark Saunders. 2015. Handbook of Research Methods on Trust: Second Edition. Edward Elgar Publishing, Cheltenham, United Kingdom. 1–343 pages. https://doi.org/10.4337/9781782547419

[49] Farah Magrabi, Elske Ammenwerth, Jytte Brender McNair, Nicolet F De Keizer, Hannele Hyppönen, Pirkko Nykänen, Michael Rigby, Philip J Scott, Tuulikki Vehko, Zoie Shui-Yee Wong, and Andrew Georgiou. 2019. Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications. Yearbook of medical informatics, 28(1), 128–134. https://doi.org/10.1055/s-0039-1677903

[50] Stephen Marsh and Mark R. Dibben. 2003, The role of trust in information science and technology. Ann. Rev. Info. Sci. Tech., 37: 465-498. https://doi.org/10.1002/aris.1440370111

[51] Patricia L. McDermott and Ronna ten Brink. 2019. Practical Guidance for Evaluating Calibrated Trust. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 63 (2019): 362 - 366.

[52] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 220–229. DOI:https://doi.org/10.1145/3287560.3287596

[53] Guido Möllering. 2013. Process Views of Trusting and Crises. Handbook of advances in trust research: 285-306. https://doi.org/10.4337/9780857931382.00024

[54] Myura Nagendran, Yang Chen, Christopher A Lovejoy, Anthony C Gordon, Matthieu Komorowski, Hugh Harvey, Eric J Topol, John P A Ioannidis, Gary S Collins, and Mahiben Maruthappu. 2020. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies BMJ 2020; 368 :m689 doi:10.1136/bmj.m689

[55] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertisein user trust and the impact of first impressions with intelligent systems. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing,Vol. 8. 112–121.

[56] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. Human factors 39, 2 (1997), 230–253.

[57] Samuele Lo Piano. 2020. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. Humanities and Social Sciences Communications7, 1 (2020), 1–7. https://doi.org/10.1057/s41599-020-0501-9

[58] Makenzie Pryor, Doug Ebert, Vicky Byrne, Khalaeb Richardson, Qua Jones, Richard Cole, and Anne Collins McLaughlin. 2019. Diagnosis Behaviors of Physicians and Non-Physicians When Supported by an Electronic Differential Diagnosis Aid. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 63, no. 1 (November 2019): 68–72. https://doi.org/10.1177/1071181319631420.

[59] Minakshi Raj, Adam S. Wilk, and Jodyn E. Platt. 2019. Dynamics of Physicians' Trust in Fellow Health Care Providers and the Role of Health Information Technology. Medical Care Research and Review 78, no. 4 (August 2021): 338–49. https://doi.org/10.1177/1077558719892349.

[60] Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt and Colin Camerer. 1998. Not So Different After All: A Cross-Discipline View Of Trust. Acad. Manag. Rev. 23(3), 393–404 (1998). https://doi.org/10.5465/amr.1998.926617

[61] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. (May 2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[62] Beau G. Schelble, Christopher Flathmann, Nathan J. McNeese, Guo Freeman, and Rohit Mallick. 2022. Let's Think Together! Assessing Shared Mental Models, Performance, and Trust in Human-Agent Teams. Proc. ACM Hum.-Comput. Interact. 6, GROUP, Article 13 (January 2022), 29 pages. DOI:https://doi.org/10.1145/3492832

[63] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "The human body is a black box": supporting clinical decision-making with deep learning. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 99–109. DOI:https://doi.org/10.1145/3351095.3372827

[64] Jiayi Shen, Casper J P Zhang, Bangsheng Jiang, Jiebin Chen, Jian Song, Zherui Liu, Zonglin He, Sum Yi Wong, Po-Han Fang, and Wai-Kit Ming. 2019. Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. JMIR medical informatics, 7(3), e10010. https://doi.org/10.2196/10010

[65] Lea Strohm, Charisma Hehakaya, Erik R Ranschaert, Wouter P C Boon, and Ellen H M Moors. 2020. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. European radiology, 30(10), 5525–5532. https://doi.org/10.1007/s00330-020-06946-y

[66] Mark Sujan, Dominic Furniss, Kath Grundy, Howard Grundy, David Nelson, Matthew Elliott, Sean White, Ibrahim Habli, and Nick Reynolds. 2019. Human factors challenges for the safe use of artificial intelligence in patient care. BMJ health & care informatics, 26(1), e100081. https://doi.org/10.1136/bmjhci-2019-100081

[67] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second Chance for a First Impression? Trust Development in Intelligent System Interaction. Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization. Association for Computing Machinery, New York, NY, USA, 77–87. DOI:https://doi.org/10.1145/3450613.3456817

[68] Eric Topol. 2019. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books.

[69] Eric J. Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. Nature medicine. 25, 1 (Jan. 2019), 44–56. https://doi.org/10.1038/s41591-018-0300-7

[70] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 327 (October 2021), 39 pages. DOI:https://doi.org/10.1145/3476068

[71] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. “Brilliant AI Doctor” in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 697, 1–18. DOI:https://doi.org/10.1145/3411764.3445432

[72] Anthony Wilson, Haroon Saeed, Catherine Pringle, Iliada Eleftheriou, Paul A Bromiley, and Andy Brass. 2021. Artificial intelligence projects in healthcare: 10 practical tips for success in a clinical environment. BMJ health & care informatics, 28(1), e100323. https://doi.org/10.1136/bmjhci-2021-100323

[73] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F. Antaki. 2016. Investigating the Heart Pump Implant Decision Process: Opportunities for Decision Support Tools to Help. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16). Association for Computing Machinery, New York, NY, USA, 4477–4488. DOI:https://doi.org/10.1145/2858036.2858373

[74] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 238, 1–11. DOI:https://doi.org/10.1145/3290605.3300468

[75] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 279, 1–12. DOI:https://doi.org/10.1145/3290605.3300509