Trust in Clinical AI

Expanding the Unit of Analysis

Browne, Jacob T.; Bakker, Saskia; Yu, Bin; Lloyd, Peter; Ben Allouch, Somaya

DOI
10.3233/FAIA220192

Publication date
2022

Document Version
Final published version

Published in
HHAI2022

HHAI2022: Augmenting Human Intellect
S. Schlobach et al. (Eds.)

doi:10.3233/FAIA220192

Trust in Clinical AI: Expanding the Unit of Analysis

Jacob T. Browne[a,b,1], Saskia Bakker[a], Bin Yu[a], Peter Lloyd[b], and Somaya Ben Allouch[c,d]
[a] *Philips Experience Design, Eindhoven, 5656 AE, The Netherlands*
[b] *Delft University of Technology, Delft, 2628 CD, The Netherlands*
[c] *Amsterdam University of Applied Sciences, Amsterdam, 1097 DZ, The Netherlands*
[d] *Informatics Institute, University of Amsterdam, the Netherlands*

Abstract. From diagnosis to patient scheduling, AI is increasingly 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 user interactions with the AI. We offer several points in the clinical AI development, usage, and monitoring process that can have a significant impact on trust. We argue that the calibration of trust in AI should go beyond explainable AI and focus on the entire process of clinical AI deployment. We illustrate our argument with case studies from practitioners implementing clinical AI in practice to show how trust can be affected by different stages in the deployment cycle.

Keywords. Trust, Clinical AI, Artificial Intelligence, Trust Calibration.

1. Introduction

AI is increasingly being considered across different healthcare areas: from patient facing applications to clinical workflow enhancements [1, 2]. AI can enable healthcare providers towards realizing what is known as the "quadruple aim": improved patient experience, better health outcomes, improved staff experience, and lower cost of care [3, 4]. AI has the potential to give doctors more time engaging with patients by taking care of the menial, non-critical tasks, and being a "second reader" for some clinicians, helping aid detection and diagnosis [5, 6, 7, 8]. 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 [8, 9, 10, 11, 12, 13]. The major focus of the AI community has been on crafting better performing models, rather than the effects of AI implementation in practice and challenges associated with adoption [14, 15].

Despite ever better AI models, the adoption of AI into actual clinical practice has been limited [16, 17, 18, 19]. Aside from poor human-centered design, trust remains a critical challenge in deploying clinical AI, being influenced by many factors (education, experience, bias, system controllability, complexity, risks, etc.) [8, 20, 21, 22, 23, 24, 25,

[1] Corresponding Author: Jacob Browne Philips, Eindhoven, 5656 AE, The Netherlands; E-mail: jacob.browne@philips.com.

26]. Even if the AI system is deployed into an existing workflow, whether the clinician trusts it enough to adopt in usage remains a challenge [27]. Additionally, trust has increasingly been mentioned by regulatory bodies as crucial to the development of human-centered AI [28, 29].

Research investigating how trust is formed in clinical AI in practice is currently lacking. Many studies investigating trust in AI-assisted decision making do not go beyond evaluations of the interface in a laboratory setting (often through investigating XAI) [30]. 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 valid [31]. Similarly, in a survey of AI-decision making research, Lai et al. call for more investigations of AI beyond discrete, laboratory trials [32].

This paper makes two contributions: an expansion of the unit of analysis of trust in clinical AI and a review of different trust calibration points within the clinical AI deployment process. Through this expansion, we render the problem of trust calibration to encompass more than just XAI, instead extending throughout the entire clinical AI deployment process.

2. Defining Trust

2.1. What is Trust?

Trust is a multifaceted, multidisciplinary, and challenging theoretical concept to study and define, with many definitions from different fields abounding [33, 34, 35]. Despite this, there are some common, important components that bind them. To define trust, we follow Vereschak et al. in their systematic review of trust definitions within Human-AI literature [31]. They cement Lee and See's seminal definition of trust as comprising 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" [31, 33]. Reliance (asking for a system's recommendation) and compliance (following a system's recommendation) with a system serve as trust-related behavioral measures, as trust cannot be inferred from a specific behavior [31, 36].

An important focus within the trust literature is calibrated or appropriate trust: that the level of user trust matches the capabilities of the AI [33]. Rather than simply promoting trust in a system, trust should be calibrated to system capabilities as they modulate in usage. Overtrust will lead to overreliance and misuse of the AI (often referred to as automation bias), while undertrust will lead to underreliance and disuse of the AI (often known as algorithm aversion) [37, 38]. Poor trust calibration will lead to a lower performance in human-AI teams [25]. Clinical AI operates in a high-stakes environment, rendering appropriate trust critical to the development of clinical AI. The goal of the designer is to keep the user's trust calibrated to the capabilities of the system.

2.2. Expanding Trust

> *"...most contemporary social scientists do not view trust as a process. This can partly be explained by the fact that trust is traditionally measured through surveys and experiments, which*

are not particularly useful for depicting the dynamic nature of trust." [39]

What is 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 [40]. Instead, studies tend to focus on individual interactions of a clinician and a prototype outside of a clinical context. While these investigations are valuable, they dodge the larger picture of how the trusting process occurs. Some recent work on trust offers us an expansive path forward.

In an integration of empirical research on trust, Hoff and Bashir build upon three variables of trust: dispositional, situational, and learned trust [27, 41]. Dispositional trust refers to a personal tendency to trust. Culture, age, personality traits, etc. can all lead to different dispositions to trust at different stages of use [27, 33]. Situational refers to the context in which the interaction takes place, where aspects of the context could affect trust. This includes both external factors (type of system, difficulty of the task, team dynamics, workload, decisional freedom, etc.) and internal factors (affect, expertise, attentional capacity, etc.) [42]. Further, Meyerson et al. point to trust being affected by the temporal context in a group (whether a group formation is more temporary or permanent) [43]. Learned trust refers to trust in past interactions with the system. Learned trust is expanded into initial learned trust (trust prior to interaction) and dynamic learned trust (trust during an interaction).

Trusting occurs in a social context. Lee and See advocate that "trust between people depends on the individual, organizational, and cultural context… it affects initial levels of trust and how people interpret information regarding the agent." [33]. Chiou and Lee emphasize the social aspects of automation, where trust is "…essential to coordination, rather than relying on individual skill, static knowledge structures, or having well-defined roles" [36]. There is an increased focus on the joint activity, shared awareness, coordination, cooperation, adaptation within dynamic environments and how trust develops through interactions [36]. Their attention is drawn to a relational approach of trusting, an ongoing dialectic, rather than a static state of trust.

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" [44]. This framework includes interpersonal and technical factors that relate to trust in a dynamic, transitive light. Similarly, de Visser et al. emphasize the longitudinal, relationship-like nature of trust, offering methods for trust dampening and repairing [45, 46].

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" [47]. 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

> *"The interrelated, communicative types of interpersonal factors in complex teams are not captured by trust studies that consider only individual end-users."* [44]

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 [30, 48, 49]. Rarely is the case of trust before the prototype considered crucial [14, 30]. However, as evidenced by studies from practitioners developing clinical AI in practice, the process of trusting begins way before clinicians ever see an interface and extend beyond experiences with XAI. 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. Our unit of analysis should go beyond a clinician's interactions with the AI once deployed and expand to the entire deployment process. This attempt is reminiscent of Fitzpatrick and Ellingsen's call for a focus on the large-scale, longitudinal nature of healthcare projects and recognizing the situated nature of clinical work [50]. We'll consider 5 areas where significant trusting processes occur derived from clinical AI literature: pre-development, during development, generating evidence, during use, and post-deployment (see Figure 1).

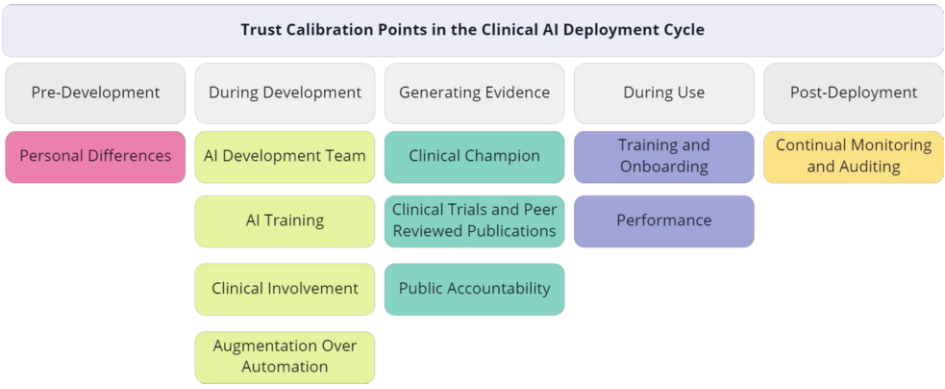

Figure 1. Different calibration points in the Clinical AI Deployment Cycle.

3.1. Pre-Development

3.1.1. Personal Differences in How Clinicians Trust

People have different inclinations to trust, known as dispositional trust [27, 51, 52]. Culture, age, attachment styles, and other personal differences all count towards this dispositional trust [27, 53]. These differences can greatly affect trust and reliance in ways not related to the properties of the AI, across the entire trusting process [33]. 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. [31, 51, 55]. These differences are rarely considered when investigating trust in clinical AI, despite having important implications for how to calibrate trust.

3.2. During Development

3.2.1. Trust Through the AI Development Team

The formation of trust in clinical AI starts as early as the assembling of the team developing the AI. Releasing clinical AI is an extensive, complex process [40, 56]. 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 [57]. Clinical stakeholders do not want to harm patients and besmirch their own career aspirations by relying on untrustworthy AI. Clinicians may trust an AI simply because they value the brand based on prior experiences or cultural approval [36, 40].

3.2.2. Trust Through the Training of the AI

The selection of the AI's dataset, type of model, training, and metrics influence trust [26, 58]. 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 [59]. Some pathologists would not trust the AI unless it were trained on judgements made by well-respected clinicians [59]. 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" [59]. To pathologists, where the algorithm received its training data is comparable to where their colleagues were trained [59]. The need to know the details of the training data can be seen as an extension of the practice of resolving clinical uncertainty by expert consultation, and thus informing positive expectations. Clinicians need to know how much the training data matches their local, live clinical data in allowing vulnerability. As noted by Engström et al., "there may be a need for local tuning of an AI before deployment in a trial, which needs to fit well with the pre-existing organization and ensure patient safety" [60].

3.2.3. Trust Through Clinical Involvement

> *"Any time you are adopting new technology which is not validated, I think there is some amount of trust building that has to go along with the project and that comes from working with an engagement right from the beginning."* [40]

The degree of clinical involvement has an impact on trust in several ways [51]. 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 [36, 61]. The more clinical involvement, the better the AI developers will be able to design for those mental models and integrate successfully into the clinical workflow. Cai et al. found that dissimilar mental models between pathologists and the AI system degraded trust [62]. The more accurate the user's mental model of the AI, the better their performance with the AI [63, 64]. As the user's mental model of the system evolves, so does their trust [65, 66]. Early in the AI training process, Benda et al. note that reviewing the inputs of the AI with clinicians can help foster trust and catch quality issues [51]. This also helps clinicians understand how the AI works. Jacobs et al. held co-design and

interview sessions with clinicians to understand their perceptions in using a DST (decision support tool) for antidepressant treatment decisions [30]. Through these sessions, they revealed that DSTs should: engage with the broader healthcare system beyond the clinician including patients and other healthcare providers, connect to existing healthcare system processes, be designed for time-constrained environments, and adapt to information contrasting clinical guidelines [30]. By increasing clinical involvement in the development process, developers will make better AI integrations and thus make their systems more likely to be used and trusted.

Secondly, more clinical involvement means clinicians will have more of a stake in designing the task allocation or division of labor of the AI [67]. 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 [33, 51].

Lastly, the AI team will be able to develop relationships with the clinical team. Sendak et al. offer us a profound insight in their development of Sepsis Watch: "trust in a technology is rooted in relationships - not in a technical specification or feature" [14]. 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. These meetings develop positive expectations to allow for instances of vulnerability. In piloting Sepsis Watch, Sendak et al. formed accountable relationships with appropriate leaders from the emergency department, establishing monthly meetings [14]. Training sessions were conducted with nurses (the primary users of Sepsis Watch) during the first two weeks of their pilot [14]. After a month of use, feedback was presented to the governance committee, and they created several workflow changes [14].

This is the development of positive expectations of clinical AI before deployment, and key to successful integration. Incorporating different users into the development process will help create a better understanding of how to develop trust at their local sites [51]. Barda et al. emphasize that different needs arise from different clinical stakeholders, especially when considering XAI [68]. This learned initial trust development through these sessions will go a long way in the successful deployment of clinical AI and the calibration of appropriate trust.

3.2.4. Trust Through Augmentation, Not Automation

Different levels of automation and the level of control the operator has impacts trust [27]. 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 a productive endeavor [69]. 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 [9]. To render this point, although BoneXpert was developed to automate the bone age rating task of a radiologist, 82% of radiologists who use it still do some degree of assessment of the radiographs [70]. Instead, it is better to think of the AI as augmenting the clinician. 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 [71]. Lee et al. found a similar improvement in therapists using AI in a rehabilitative assessment context [72].

Wang et al. found clinicians want to maintain decisional power over the AI and verify any decisions it would make [8]. They also found that clinicians did not 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]." [8]. Instead, the diagnosis process is a "highly interactive, communicative, and social event", thus not up for automation anytime soon [8]. De Boo reflects this in computer-aided diagnosis (CAD) in radiology, where CAD is more of a second reader, a complementary tool that can spot lesions that the radiologist might miss, while the radiologist can dismiss or accept different findings [73]. 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." [22]. This is indeed like airplane automation, where pilots need to be trained on how to be better monitors of automation [11, 74]. To maintain autonomy and agency in clinicians, Cai et al. found it important to give clinicians tools to refine the system, allowing the clinician to improve the system, remain in control, and disambiguate mistakes [62]. Further, clinicians need to be reminded that while they are working in an environment with AI agents, they are responsible for their decisions [75]. Being forward with this approach when working with a clinical team is important for integration and them forming positive expectations.

3.3. Generating Evidence

3.3.1. Trust Through a Clinical Champion

Clinical and development teams working together are often accompanied by a clinical champion, a key to gaining the positive expectations of the clinical staff [22, 76]. 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" [76]. This is further emphasized by Wilson et al. and Cosgriff et al., "We need visionary clinicians working with expert technical collaborators to establish the organizational structures requisite to translate technological progress into meaningful clinical outcomes… the hype around AI in healthcare will only be realized when the scattered champions of this movement emerge from their silos and begin formally working as a team under the same roof" [58, 77]. Strohm et al. point out that these insiders share information about AI to other clinicians and promote opportunities for experimentation [22]. The presence of a clinical champion serves to increase the positive expectations of necessary stakeholders in the clinical system.

3.3.2. 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 and regulatory standards (e.g., FDA) [19, 56, 78, 79]. For instance, Sendak et al. had both the sepsis definition and model peer-reviewed and disseminated in clinical and technical venues [14]. 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 [14, 80]. 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" [59]. In an evaluation of BoneXpert, Thodberg et al. found that 71% of radiologists indicated that

clinical evaluation of data through performance data and peer-reviewed publications were the most important factors in generating trust in BoneXpert [70]. Jacobs et al. also mention that DSTs going through random controlled trials have a large influence on trust, emphasizing that clinicians don't have time to evaluate trust in clinical AI at every recommendation the DST makes: "participants expected that trust in the technology will not be decided at each decision point." [30]. This is an extension of medicine's reliance on such clinical, peer reviewed processes. The fact that clinical AI should go through this practice is an indicator of positive expectations and extended vulnerability: other members of clinical practice either vow or disavow to use this technology based upon their practice [30, 56].

3.3.3. Trust Through Public Accountability

A form of public accountability through the participation of external parties during trial phases can also affect trust. Although seemingly rare in the literature, Sendak et al. offer an account of this [14]. 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 [14]. This enabled positive expectations to be built and exchanges of vulnerability between the clinical and development team.

3.4. Trust During Use

3.4.1. Trust Through Training and Onboarding Sessions

> *"…training and documentation, when done right, can make up for trust lost elsewhere"* [81]

Training sessions with clinicians also modulate trust: both active training on how to use it and allowing clinicians to test the AI out, understand its limitations, and use cases, often through onboarding [25, 27, 56]. This onboarding is essential: developers are introducing a sociotechnical system, transforming the complex, distributed workflow of clinicians. Expectation setting of AI early on has been shown to have an impact on user perceptions and trust [82, 83].

During onboarding and throughout usage, explanations of model predictions, transparency into higher level objectives, global behavior, expectations of model performance, and tendencies can be needed [27, 59, 84, 85]. First impressions with AI systems have a large effect on trust development downstream [27, 36, 52, 86, 87]. To set the AI's capabilities, it should be noted what is the AI good at, what cannot it do, and how well it can do what it does [80, 88]. De Boo shares that in radiology, the radiologist should learn how to use their CAD (computer-aided detection) to understand the optimal reading of its findings around true and false positives, "they have to become familiar with the potential and limitations of the CAD they are using and they have to build up a trust into the CAD's capability to be able to accept the true positive lesions without a too big loss of specificity by correctly ruling out the false positive lesions" [73].

Cai et al. emphasize that beyond knowing summary statistics of performance and how to use the AI, pathologists relate to the AI assistant as they would a new colleague: homing in on its medical point-of-view, strengths, weaknesses, and how it complements their skill set (e.g., relevant patient populations to use the AI on) [59]. The AI could also

be calibrated to the clinician's preferred way of operating (e.g., a radiologist setting the AI to only notify them when the AI is over 80% confident of a finding).

However, as in the case of airline automations, this added explanation is typically the first step taken by industries in mitigating risk, and notably insufficient [74]. We need larger efforts of training that expand beyond XAI: educating clinicians in training and in practice on the basics of how AI works, how humans work with AI effectively (mitigating biases, maintaining awareness, knowledge of the new task given automation, etc.), and how to respond to its alarms, predictions, etc. [67, 74].

Cai et al. note how some pathologists envisioned comparing their own diagnoses with the AI's diagnoses on a set of cases with ground truth data give insight into the AI's tendencies [59]. This also has the effect of instilling cases of low-stakes vulnerability, allowing clinicians to slowly test the system and see how it performs against themselves. Model refinement mechanisms allowed clinicians to test, understand, and grapple with opaque models [62]. These mechanisms could allow clinicians to improve the model itself, increase transparency through testing, and help better form their own mental models beyond explanations. Similarly, Lee et al. found that following an interactive machine learning approach in a rehabilitative assessment context improved therapist performance [72].

Wang et al. found that a lack of training lessened trust in the AI, as clinicians had to learn how to use and understand the system alone [8]. Henry et al. discuss how in a sepsis alert system, clinicians might dismiss the alert if there are not clear signs of sepsis and the patient has a less common presentation of sepsis [54]. Training clinicians on how the system can detect this rarity would be crucial [54]. Training and onboarding are vital forms of trust calibration [36].

3.4.2. Trust Through Performance

> *"All the team leaders knew that establishing "trust" was an essential foundation upon which everything else would rest. Only if a technology is trusted will it be used."* [40]

As clinicians use the AI, their trust will modulate dynamically based on system performance and different sociotechnical factors (known as dynamic learned trust and situational trust) [27]. The goal is to calibrate appropriate trust in real-time by communicating an accurate picture of the AI's performance and through trust dampening and repair mechanisms to increase human-AI teamwork performance. How does trust develop over time in actual usage and what trust modulation mechanisms work in context [25]?

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 [89]? Was the data used in training the AI similar to data gathered in this live clinical context? How can you successfully communicate this to the clinicians in context? Answers to these questions are difficult to predict given clinical contexts and procedures can vary from hospital to hospital. As Elish notes in one of their interviews with clinicians, "If you've seen one academic hospital, you've seen one academic hospital." [40]. Clinical environments are situated in terms of complex organizational policies, regulations, and culture which defy laboratory settings [20]. Not attending to these complexities increases non-adoption of technology within clinical environments, even in the case of seemingly

simple deployments [90, 91]. Observing the AI using live clinical data, in context, with clinicians can help surface these socio-environmental problems [90].

Levy et al. emphasize the importance of long-term investigations of AI's impact on trust. In introducing automation in annotating clinical texts, they found that some users tended to accept improper results, lose engagement in the task, and take less initiative in making their own annotations [92]. How does partial automation and decision support affect the task, what unintended consequences come of this automation [15, 93]? Alberdi et al. note how a CAD for radiologists resulted in automation bias from an absence of prompts on a mammogram, while Nishikawa et al. note how radiologists ignore correct prompts [5, 94]. Maintaining vigilance as the clinician becomes habituated to the workflow is key [95, 96].

Given appropriate trust is built upon having knowledge of the AI's performance, it is important to inform clinicians on the AI's past performance at their local hospital [25, 85, 97]. Such metrics need to be relevant and understandable to clinicians, while respecting the limited time clinicians have and not getting in the way of their work [30, 51, 56]. 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 [25]. Further, specific environmental contexts can cause the AI to not perform well and reduce trust [8, 25, 33, 98]. 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 [90]. How well can the AI explain why this happened, how could the user accommodate, and is it worth accommodating? Could the user have prevented errors from happening by learning how the AI works? How the system responds to such failures can have an effect whether trust is repaired (e.g., how the AI expresses regret or apologizes) [99, 100]. As an example, BoneXpert will automatically reject radiographs not suitable for evaluation and 48% radiologists pointed to that being important to building trust [70].

How are trust dampening and trust repairing mechanisms released as needed, and do they even work as intended [46]? Trust repair would repair trust after a trust violation, while trust dampening would lower expectations as needed [46]. How people respond to different trust repair strategies varies by person and context [45, 101, 102]. 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 [45]. McDermott have referred to calibration points, points in time where AI performance is degraded or improved, and trust needs to be increased or dampened [103]. Through the human-centered design process, designers could uncover what information the system needs to show at different calibration points and prototype these different scenarios [104, 105].

XAI may play an important role in modulating trust during actual usage by explaining global model processes, limitations, and instance/local specific explanations when appropriate [25, 86, 106]. For instance, confidence scores can be used for different marks in a computer-aided diagnostic system for radiologists [25, 107, 108]. Jacobs et al. mention how the AI could bring up a recommendation that contrasts with existing clinical knowledge due to the AI finding a nuanced relationship not known in clinical practice [30]. Understanding this divergence and why the recommendation was made would be important to maintain trust [106]. However, different tasks and contexts will need different levels of XAI and transparency (e.g., low stakes v. high stakes scenarios). Jacobs et al. note how clinicians in high stakes scenarios or with limited time won't check the explanation of a decision, and instead more emphasis should be placed early in the

development process to determine when the system errs or when it should not be used [30, 49].

Trust during actual usage will also be influenced by other stakeholders in the clinical team. The clinical team as a complex social system with incredible prowess in awareness, coordination, articulation, and collaboration is well studied [109, 110, 111, 112, 113, 114, 115, 116, 117]. Clinicians often do not follow rigid, perfect protocols; instead, clinical workflows are notably chaotic, rife with biases and communication breakdowns, dependent on specific local norms [8, 118, 119, 120]. Clinicians will have varying degrees of experience with the AI: some holding positive expectations, others dismissing the AI. 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" [33]. These networks of trust relations will be impacted by each other [44]. Whether or not a trusted, senior clinician trusts an AI will influence the rest of the team's trust in the AI [91]. Similarly, clinicians may be required to interact with an EHR system more, increasing situational trust [54]. The implementation of AI will have a broader effect on trust between healthcare professionals, and in turn, will affect trust in the AI [121].

3.5. Post-Deployment

3.5.1. Trust Through Continual Monitoring and Auditing

The development team needs to continually monitor performance to ensure the clinical AI is performing effectively and safely [56]. 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 [14]. Medical procedures and best practices are similarly dynamic. Does the AI reflect up to date knowledge of clinical practice [89]? 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 [89, 92]. Feedback from clinicians after deployment will be critical to maintaining relationships and maintaining trust [122].

4. Conclusion and Future Work

> *"Choosing the right boundaries for a unit of analysis is a central problem in every science and the basic approach to this problem has been in place for 2,000 years. Plato advised that one should ''carve nature at its joints'' (Phaedrus 265d–266a). By this, Plato meant that we should place the boundaries of our units where connectivity is relatively low."* [123]

Through this paper, we have shown how the unit of analysis of trust in clinical AI should be expanded beyond current foci of explainable AI and usage in situ. The process of trusting extends throughout the clinical AI deployment process, from the construction of the development team to how it is audited post-deployment. Each of these points have various interventions to calibrate trust and we will need future work to investigate how to best calibrate trust at each point, as it makes sense in each context. There may also be other significant points that are not surfaced in the current clinical AI literature.

We need empirical research investigating trust building in clinical AI as it occurs in actual clinical practice [14, 24]. As Chiou and Lee mention, we should go beyond "necessary but insufficient guidelines" that argue for transparency (e.g., "make clear what the system can do") and focus on guidelines for how trust is developed across interactions and situations [36, 88]. Institutions will differ in how they trust clinical AI. Okolo emphasizes this further in pointing out differences in trusting processes between clinicians in the Global North and the Global South: there does not seem to be a one size fits all solution [124]. We share the emphasis on the need for a participatory, iterative design process to successfully integrate into clinical workflows [124].

Future work will investigate each point in the clinical AI deployment cycle to better understand what trust calibration methods are most useful, how to measure and investigate trust at each stage, and how these factors vary across different clinical contexts. Trust cannot be an afterthought, but rather it should be a central design aim of the project [125].

5. Acknowledgements

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] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94–8. Available from: http://dx.doi.org/10.7861/futurehosp.6-2-94

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

[3] Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider. Ann Fam Med [Internet]. 2014;12(6):573–6. Available from: http://dx.doi.org/10.1370/afm.1713

[4] Sikka R, Morath JM, Leape L. The Quadruple Aim: care, health, cost and meaning in work. BMJ Qual Saf [Internet]. 2015;24(10):608–10. Available from: http://dx.doi.org/10.1136/bmjqs-2015-004160

[5] Alberdi E, Povykalo A, Strigini L, Ayton P. Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. Acad Radiol [Internet]. 2004;11(8):909–18. Available from: http://dx.doi.org/10.1016/j.acra.2004.05.012

[6] Topol E. Deep medicine: How artificial intelligence can make healthcare human again. London, England: Basic Books; 2019.

[7] van Beek EJR, Mullan B, Thompson B. Evaluation of a real-time interactive pulmonary nodule analysis system on chest digital radiographic images. Acad Radiol [Internet]. 2008;15(5):571–5. Available from: http://dx.doi.org/10.1016/j.acra.2008.01.018

[8] Wang D, Wang L, Zhang Z, Wang D, Zhu H, Gao Y, et al. "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. New York, NY, USA: ACM; 2021.

[9] Gichoya JW, Nuthakki S, Maity PG, Purkayastha S. Phronesis of AI in radiology: Superhuman meets natural stupidity. arXiv [csCY] [Internet]. 2018 [cited 2022 Mar 9]; Available from: http://arxiv.org/abs/1803.11244

[10] Gong B, Nugent JP, Guest W, Parker W, Chang PJ, Khosa F, et al. Influence of artificial intelligence on Canadian medical students' preference for radiology specialty: ANational survey study. Acad Radiol [Internet]. 2019;26(4):566–77. Available from: http://dx.doi.org/10.1016/j.acra.2018.10.007

[11] Langlotz CP. Will artificial intelligence replace radiologists? Radiology: Artificial Intelligence [Internet]. 2019;1(3):e190058. Available from: http://dx.doi.org/10.1148/ryai.2019190058

[12] 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. BMJ [Internet]. 2020;368:m689. Available from: http://dx.doi.org/10.1136/bmj.m689

[13] Shen J, Zhang CJP, Jiang B, Chen J, Song J, Liu Z, et al. Artificial intelligence versus clinicians in disease diagnosis: Systematic review. JMIR Med Inform [Internet]. 2019;7(3):e10010. Available from: http://dx.doi.org/10.2196/10010
[14] Sendak M, Elish MC, Gao M, Futoma J, Ratliff W, Nichols M, et al. 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. New York, NY, USA: ACM; 2020.
[15] Sujan M, Furniss D, Grundy K, Grundy H, Nelson D, Elliott M, et al. Human factors challenges for the safe use of artificial intelligence in patient care. BMJ Health Care Inform [Internet]. 2019;26(1):e100081. Available from: http://dx.doi.org/10.1136/bmjhci-2019-100081
[16] Coiera E. The last mile: Where artificial intelligence meets reality. J Med Internet Res [Internet]. 2019;21(11):e16323. Available from: http://dx.doi.org/10.2196/16323
[17] Elwyn G, Scholl I, Tietbohl C, Mann M, Edwards AGK, Clay C, et al. "Many miles to go …": a systematic review of the implementation of patient decision support interventions into routine clinical practice. BMC Med Inform Decis Mak [Internet]. 2013;13 Suppl 2(S2):S14. Available from: http://dx.doi.org/10.1186/1472-6947-13-S2-S14
[18] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med [Internet]. 2019;17(1):195. Available from: http://dx.doi.org/10.1186/s12916-019-1426-2
[19] Yang Q, Steinfeld A, Zimmerman J. 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. New York, NY, USA: ACM; 2019.
[20] Sandhu S, Lin AL, Brajer N, Sperling J, Ratliff W, Bedoya AD, et al. Integrating a machine learning system into clinical workflows: Qualitative study. J Med Internet Res [Internet]. 2020;22(11):e22421. Available from: http://dx.doi.org/10.2196/22421
[21] Benda NC, Das LT, Abramson EL, Blackburn K, Thoman A, Kaushal R, et al. "How did you get to this number?" Stakeholder needs for implementing predictive analytics: a pre-implementation qualitative study. J Am Med Inform Assoc [Internet]. 2020;27(5):709–16. Available from: http://dx.doi.org/10.1093/jamia/ocaa021
[22] Strohm L, Hehakaya C, Ranschaert ER, Boon WPC, Moors EHM. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol [Internet]. 2020;30(10):5525–32. Available from: http://dx.doi.org/10.1007/s00330-020-06946-y
[23] Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons for physicians not adopting clinical decision support systems: Critical analysis. JMIR Med Inform [Internet]. 2018;6(2):e24. Available from: http://dx.doi.org/10.2196/medinform.8912
[24] Yang Q, Zimmerman J, Steinfeld A, Carey L, Antaki JF. 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. New York, NY, USA: ACM; 2016.
[25] Jorritsma W, Cnossen F, van Ooijen PMA. Improving the radiologist-CAD interaction: designing for appropriate trust. Clin Radiol [Internet]. 2015;70(2):115–22. Available from: http://dx.doi.org/10.1016/j.crad.2014.09.017
[26] Cabitza F, Campagner A, Balsano C. Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters. Ann Transl Med [Internet]. 2020;8(7):501–501. Available from: http://dx.doi.org/10.21037/atm.2020.03.63
[27] Hoff KA, Bashir M. Trust in automation: integrating empirical evidence on factors that influence trust: Integrating empirical evidence on factors that influence trust. Hum Factors [Internet]. 2015;57(3):407–34. Available from: http://dx.doi.org/10.1177/0018720814547570
[28] European Commission. On Artificial Intelligence - A European approach to excellence and trust. Technical Report. European Commission, Brussels, Belgium. 27 pages. 2022. https://ec.europa.eu/info/sites/info/files/commission-whitepaper-artificial-intelligence-feb2020$_$en.pdf
[29] European Commission. Building Trust in Human-Centric Artificial Intelligence - JRC Science Hub Communities - European Commission [Internet]. JRC Science Hub Communities - European Commission. 2022 [cited 9 March 2022]. Available from: https://ec.europa.eu/jrc/communities/en/community/digitranscope/document/building-trust-human-centric-artificial-intelligence
[30] Jacobs M, He J, F. Pradier M, Lam B, Ahn AC, McCoy TH, et al. 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. New York, NY, USA: ACM; 2021.
[31] Vereschak O, Bailly G, Caramiaux B. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. Proc ACM Hum Comput Interact [Internet]. 2021;5(CSCW2):1–39. Available from: http://dx.doi.org/10.1145/3476068

[32] Lai V, Chen C, Liao QV, Smith-Renner A, Tan C. Towards a science of human-AI decision making: A survey of empirical studies. 2021; Available from: http://dx.doi.org/10.48550/ARXIV.2112.11471
[33] Lee JD, See KA. Trust in automation: Designing for appropriate reliance. Hum Factors [Internet]. 2004;46(1):50–80. Available from: http://dx.doi.org/10.1518/hfes.46.1.50.30392
[34] Lyon F, Möllering G, Saunders M. Handbook of research methods on trust. Edward Elgar Publishing; 2015.
[35] Rousseau DM, Sitkin SB, Burt RS, Camerer C. Not so different after all: A cross-discipline view of trust. Acad Manage Rev [Internet]. 1998;23(3):393–404. Available from: http://dx.doi.org/10.5465/amr.1998.926617
[36] Chiou EK, Lee JD. Trusting automation: Designing for responsivity and resilience. Hum Factors [Internet]. 2021;187208211009995. Available from: http://dx.doi.org/10.1177/00187208211009995
[37] Cummings M. Automation bias in intelligent time critical decision support systems. In: AIAA 1st Intelligent Systems Technical Conference. Reston, Virigina: American Institute of Aeronautics and Astronautics; 2004.
[38] Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen [Internet]. 2015;144(1):114–26. Available from: http://dx.doi.org/10.1037/xge0000033
[39] Khodyakov D. Trust as a process: A three-dimensional approach. Sociology [Internet]. 2007;41(1):115–32. Available from: http://dx.doi.org/10.1177/0038038507072285
[40] Elish MC. The stakes of uncertainty: Developing and integrating machine learning in clinical care. Conf Proc Ethnogr Prax Ind Conf [Internet]. 2018;2018(1):364–80. Available from: http://dx.doi.org/10.1111/1559-8918.2018.01213
[41] Marsh S, Dibben MR. The role of trust in information science and technology. Annu rev inf sci technol [Internet]. 2005;37(1):465–98. Available from: http://dx.doi.org/10.1002/aris.1440370111
[42] Sato T, Yamani Y, Liechty M, Chancey ET. Automation trust increases under high-workload multitasking scenarios involving risk. Cogn Technol Work [Internet]. 2020;22(2):399–407. Available from: http://dx.doi.org/10.1007/s10111-019-00580-5
[43] Meyerson D, Weick KE, Kramer RM. Swift Trust and Temporary Groups. In: Trust in Organizations: Frontiers of Theory and Research. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc.; 2012. p. 166–95.
[44] Huang L, Cooke NJ, Gutzwiller RS, Berman S, Chiou EK, Demir M, et al. Distributed dynamic team trust in human, artificial intelligence, and robot teaming. In: Trust in Human-Robot Interaction. Elsevier; 2021. p. 301–19.
[45] de Visser EJ, Pak R, Shaw TH. From "automation" to "autonomy": the importance of trust repair in human-machine interaction. Ergonomics [Internet]. 2018;61(10):1409–27. Available from: http://dx.doi.org/10.1080/00140139.2018.1457725
[46] de Visser EJ, Peeters MMM, Jung MF, Kohn S, Shaw TH, Pak R, et al. Towards a theory of longitudinal trust calibration in human–robot teams. Int J Soc Robot [Internet]. 2020;12(2):459–78. Available from: http://dx.doi.org/10.1007/s12369-019-00596-x
[47] Möllering G. Process views of trusting and crises [Internet]. Handbook of Advances in Trust Research. Edward Elgar Publishing; 2014. p. 285–306. Available from: http://dx.doi.org/10.4337/9780857931382.00024
[48] He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med [Internet]. 2019;25(1):30–6. Available from: http://dx.doi.org/10.1038/s41591-018-0307-0
[49] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell [Internet]. 2019;1(5):206–15. Available from: http://dx.doi.org/10.1038/s42256-019-0048-x
[50] Fitzpatrick G, Ellingsen G. A review of 25 years of CSCW research in healthcare: Contributions, challenges and future agendas. Comput Support Coop Work [Internet]. 2013;22(4–6):609–65. Available from: http://dx.doi.org/10.1007/s10606-012-9168-0
[51] Benda NC, Novak LL, Reale C, Ancker JS. Trust in AI: why we should be designing for APPROPRIATE reliance. J Am Med Inform Assoc [Internet]. 2021;29(1):207–12. Available from: http://dx.doi.org/10.1093/jamia/ocab238
[52] Tolmeijer S, Gadiraju U, Ghantasala R, Gupta A, Bernstein A. Second chance for a first impression? Trust development in intelligent system interaction. In: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization. New York, NY, USA: ACM; 2021.
[53] Gillath O, Ai T, Branicky MS, Keshmiri S, Davison RB, Spaulding R. Attachment and trust in artificial intelligence. Comput Human Behav [Internet]. 2021;115(106607):106607. Available from: http://dx.doi.org/10.1016/j.chb.2020.106607

[54] Henry KE, Adams R, Parent C, Sridharan A, Johnson L, Hager DN, et al. Evaluating adoption, impact, and factors driving adoption for TREWS, a machine learning-based sepsis alerting system [Internet]. bioRxiv. 2021. Available from: http://dx.doi.org/10.1101/2021.07.02.21259941
[55] Chien S-Y, Lewis M, Sycara K, Kumru A, Liu J-S. Influence of culture, transparency, trust, and degree of automation on automation use. IEEE Trans Hum Mach Syst [Internet]. 2020;50(3):205–14. Available from: http://dx.doi.org/10.1109/thms.2019.2931755
[56] de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med [Internet]. 2022;5(1):2. Available from: http://dx.doi.org/10.1038/s41746-021-00549-7
[57] Parasuraman R, Riley V. Humans and automation: Use, misuse, disuse, abuse. Hum Factors [Internet]. 1997;39(2):230–53. Available from: http://dx.doi.org/10.1518/001872097778543886
[58] Wilson A, Saeed H, Pringle C, Eleftheriou I, Bromiley PA, Brass A. Artificial intelligence projects in healthcare: 10 practical tips for success in a clinical environment. BMJ Health Care Inform [Internet]. 2021;28(1):e100323. Available from: http://dx.doi.org/10.1136/bmjhci-2021-100323
[59] Cai CJ, Winter S, Steiner D, Wilcox L, Terry M. "hello AI": Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. Proc ACM Hum Comput Interact [Internet]. 2019;3(CSCW):1–24. Available from: http://dx.doi.org/10.1145/3359206
[60] Engström E, Strand F, Strimling P. Human-AI interactions in a trial of AI breast cancer diagnostics in a real-world clinical setting. ACM CHI Workshop on Realizing AI in Healthcare: Challenges Appearing in the Wild, May 08–09, 2021, Virtual. ACM, New York, NY, USA, 6 pages. Available from: http://francisconunes.me/RealizingAIinHealthcareWS/papers/Engstrom2021.pdf
[61] Lee MH, Siewiorek DP, Smailagic A, Bernardino A, Bermúdez i Badia S. Co-design and evaluation of an intelligent decision support system for stroke rehabilitation assessment. Proc ACM Hum Comput Interact [Internet]. 2020;4(CSCW2):1–27. Available from: http://dx.doi.org/10.1145/3415227
[62] Cai CJ, Reif E, Hegde N, Hipp J, Kim B, Smilkov D, et al. 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. New York, NY, USA: ACM; 2019.
[63] Bansal G, Nushi B, Kamar E, Lasecki WS, Weld DS, and Horvitz E. Beyond accuracy: the role of mental models in human-AI team performance. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 2019. pp. 2–11.
[64] Gero KI, Ashktorab Z, Dugan C, Pan Q, Johnson J, Geyer W, et al. Mental models of AI agents in a cooperative game setting. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM; 2020.
[65] Holliday D, Wilson S, Stumpf S. User trust in intelligent systems: A journey over time. In: Proceedings of the 21st International Conference on Intelligent User Interfaces. New York, NY, USA: ACM; 2016.
[66] Muir BM. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics [Internet]. 1994;37(11):1905–22. Available from: http://dx.doi.org/10.1080/00140139408964957
[67] Lyell D, Coiera E, Chen J, Shah P, Magrabi F. How machine learning is embedded to support clinician decision making: an analysis of FDA-approved medical devices. BMJ Health Care Inform [Internet]. 2021;28(1). Available from: http://dx.doi.org/10.1136/bmjhci-2020-100301
[68] Barda AJ, Horvat CM, Hochheiser H. 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 [Internet]. 2020;20(1):257. Available from: http://dx.doi.org/10.1186/s12911-020-01276-x
[69] Lo Piano S. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. Humanit Soc Sci Commun [Internet]. 2020;7(1). Available from: http://dx.doi.org/10.1057/s41599-020-0501-9
[70] Thodberg HH, Thodberg B, Ahlkvist J, Offiah AC. Autonomous artificial intelligence in pediatric radiology: the use and perception of BoneXpert for bone age assessment. Pediatr Radiol [Internet]. 2022; Available from: http://dx.doi.org/10.1007/s00247-022-05295-w
[71] Kiani A, Uyumazturk B, Rajpurkar P, Wang A, Gao R, Jones E, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPJ Digit Med [Internet]. 2020;3(1):23. Available from: http://dx.doi.org/10.1038/s41746-020-0232-8
[72] Lee MH, Siewiorek DPP, Smailagic A, Bernardino A, Bermúdez i Badia SB. 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. New York, NY, USA: ACM; 2021.
[73] De Boo DW, Prokop M, Uffmann M, van Ginneken B, Schaefer-Prokop CM. Computer-aided detection (CAD) of lung nodules and small tumours on chest radiographs. Eur J Radiol [Internet]. 2009;72(2):218–25. http://dx.doi.org/10.1016/j.ejrad.2009.05.062

[74] Casner SM, Hutchins EL. What do we tell the drivers? Toward minimum driver training standards for partially automated cars. J Cogn Eng Decis Mak [Internet]. 2019;13(2):55–66. Available from: http://dx.doi.org/10.1177/1555343419830901
[75] Neri E, Coppola F, Miele V, Bibbolino C, Grassi R. Artificial intelligence: Who is responsible for the diagnosis? Radiol Med [Internet]. 2020;125(6):517–21. Available from: http://dx.doi.org/10.1007/s11547-020-01135-9
[76] Lu C, Chang K, Singh P, Pomerantz S, Doyle S, Kakarmath S, et al. Deploying clinical machine learning? Consider the following. arXiv [csLG] [Internet]. 2021 [cited 2022 Mar 9]; Available from: http://arxiv.org/abs/2109.06919
[77] Cosgriff CV, Stone DJ, Weissman G, Pirracchio R, Celi LA. The clinical artificial intelligence department: a prerequisite for success. BMJ Health Care Inform [Internet]. 2020;27(1):e100183. Available from: http://dx.doi.org/10.1136/bmjhci-2020-100183
[78] Center for Devices, Radiological Health. Clinical performance assessment: Considerations for CAD devices [Internet]. U.S. Food and Drug Administration. 2020 [cited 2022 Mar 10]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-performance-assessment-considerations-computer-assisted-detection-devices-applied-radiology
[79] Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med [Internet]. 2020;3(1):118. Available from: http://dx.doi.org/10.1038/s41746-020-00324-0
[80] Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. New York, NY, USA: ACM; 2019.
[81] Widder DG, Dabbish L, Herbsleb JD, Holloway A, Davidoff S. Trust in collaborative automation in high stakes software engineering work: A case study at NASA. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM; 2021.
[82] Johnson CJ, Demir M, McNeese NJ, Gorman JC, Wolff AT, Cooke NJ. The impact of training on human-autonomy team communications and trust calibration. Hum Factors [Internet]. 2021;187208211047323. Available from: http://dx.doi.org/10.1177/00187208211047323
[83] Kocielnik R, Amershi S, Bennett PN. Will you accept an imperfect AI?: Exploring designs for adjusting end-user expectations of AI systems. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM; 2019.
[84] Schelble BG, Flathmann C, McNeese NJ, Freeman G, Mallick R. Let's think together! Assessing shared mental models, performance, and trust in human-agent teams. Proc ACM Hum Comput Interact [Internet]. 2022;6(GROUP):1–29. Available from: http://dx.doi.org/10.1145/34928322
[85] Yin M, Wortman Vaughan J, Wallach H. Understanding the effect of accuracy on trust in machine learning models. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM; 2019.
[86] Bussone A, Stumpf S, O'Sullivan D. The role of explanations on trust and reliance in clinical decision support systems. In: 2015 International Conference on Healthcare Informatics. IEEE; 2015.
[87] Nourani M, King J, Ragan E. The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. HCOMP [Internet]. 2020Oct.1 [cited 2022Mar.10];8(1):112-21. Available from: https://ojs.aaai.org/index.php/HCOMP/article/view/7469
[88] Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P, et al. Guidelines for Human-AI Interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM; 2019.
[89] Magrabi F, Ammenwerth E, McNair JB, De Keizer NF, Hyppönen H, Nykänen P, et al. Artificial intelligence in clinical decision support: Challenges for evaluating AI and practical implications: A position paper from the IMIA Technology Assessment & Quality Development in Health Informatics Working Group and the EFMI Working Group for Assessment of Health Information Systems. Yearb Med Inform [Internet]. 2019;28(1):128–34. Available from: http://dx.doi.org/10.1055/s-0039-1677903
[90] Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM; 2020.
[91] Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O. Diffusion of innovations in service organizations: systematic review and recommendations. Milbank Q [Internet]. 2004;82(4):581–629. Available from: http://dx.doi.org/10.1111/j.0887-378X.2004.00325.x
[92] Levy A, Agrawal M, Satyanarayan A, Sontag D. 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. New York, NY, USA: ACM; 2021.

[93] Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA [Internet]. 2017;318(6):517–8. Available from: http://dx.doi.org/10.1001/jama.2017.7797
[94] Nishikawa RM, Edwards A, Schmidt RA, Papaioannou J, Linver MN. Can radiologists recognize that a computer has identified cancers that they have overlooked? In: Jiang Y, Eckstein MP, editors. SPIE Proceedings. SPIE; 2006.
[95] Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc [Internet]. 2017;24(2):423–31. Available from: http://dx.doi.org/10.1093/jamia/ocw105
[96] Warm JS, Dember WN, Hancock PA. Vigilance and Workload in Automated Systems. In: Raja Parasuraman MM, editor. Automation and Human Performance: Theory and Applications. Boca Raton, FL: CRC Press; 1996. p. 18.
[97] Frison A-K, Wintersberger P, Riener A, Schartmüller C, Boyle LN, Miller E, et al. 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. New York, NY, USA: ACM; 2019.
[98] Pryor M, Ebert D, Byrne V, Richardson K, Jones Q, Cole R, et al. Diagnosis behaviors of physicians and non-physicians when supported by an electronic differential diagnosis aid. Proc Hum Factors Ergon Soc Annu Meet [Internet]. 2019;63(1):68–72. Available from: http://dx.doi.org/10.1177/1071181319631420
[99] Baker AL, Phillips EK, Ullman D, Keebler JR. Toward an understanding of trust repair in human-robot interaction: Current research and future directions. ACM Trans Interact Intell Syst [Internet]. 2018;8(4):1–30. Available from: http://dx.doi.org/10.1145/3181671
[100] Kox ES, Kerstholt JH, Hueting TF, de Vries PW. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. Auton Agent Multi Agent Syst [Internet]. 2021;35(2). Available from: http://dx.doi.org/10.1007/s10458-021-09515-9
[101] Buçinca Z, Malaya MB, Gajos KZ. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc ACM Hum Comput Interact [Internet]. 2021;5(CSCW1):1–21. Available from: http://dx.doi.org/10.1145/3449287
[102] Fahim MAA, Khan MMH, Jensen T, Albayram Y, Coman E. Do integral emotions affect trust? The mediating effect of emotions on trust in the context of human-agent interaction. In: Designing Interactive Systems Conference 2021. New York, NY, USA: ACM; 2021.
[103] McDermott PL, Brink RNT. Practical guidance for evaluating calibrated trust. Proc Hum Factors Ergon Soc Annu Meet [Internet]. 2019;63(1):362–6. Available from: http://dx.doi.org/10.1177/1071181319631379
[104] Browne JT. Wizard of oz prototyping for machine learning experiences. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems - CHI EA '19. New York, New York, USA: ACM Press; 2019.
[105] Jensen T. Disentangling trust and anthropomorphism toward the design of human-centered AI systems. In: Artificial Intelligence in HCI. Cham: Springer International Publishing; 2021. p. 41–58.
[106] Xie Y, Chen M, Kao D, Gao G, Chen X "anthony." CheXplain: Enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM; 2020.
[107] Ghesu FC, Georgescu B, Gibson E, Guendel S, Kalra MK, Singh R, et al. Quantifying and leveraging classification uncertainty for chest radiograph assessment. In: Lecture Notes in Computer Science. Cham: Springer International Publishing; 2019. p. 676–84.
[108] Zhang Y, Liao QV, Bellamy RKE. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York, NY, USA: ACM; 2020.
[109] Abraham J, Reddy MC. Re-coordinating activities: An investigation of articulation work in patient transfers. In: Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13. New York, New York, USA: ACM Press; 2013.
[110] Bardram JE, Hansen TR. Context-based workplace awareness: Concepts and technologies for supporting distributed awareness in a hospital environment. Comput Support Coop Work [Internet]. 2010;19(2):105–38. Available from: http://dx.doi.org/10.1007/s10606-010-9110-2
[111] Bjørn P, Hertzum M. Artefactual multiplicity: A study of emergency-department whiteboards. Comput Support Coop Work [Internet]. 2011;20(1–2):93–121. Available from: http://dx.doi.org/10.1007/s10606-010-9126-7
[112] Büyüktür AG, Ackerman MS. Information work in bone marrow transplant: Reducing misalignment of perspectives. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. New York, NY, USA: ACM; 2017.
[113] Mentis HM. Collocated use of imaging systems in coordinated surgical practice. Proc ACM Hum Comput Interact [Internet]. 2017;1(CSCW):1–17. Available from: http://dx.doi.org/10.1145/3134713

[114] Reddy M, Dourish P. A finger on the pulse: Temporal rhythms and information seeking in medical work. In: Proceedings of the 2002 ACM conference on Computer supported cooperative work - CSCW '02. New York, New York, USA: ACM Press; 2002.
[115] Reddy MC, Dourish P, Pratt W. Temporality in Medical Work: Time also Matters. Comput Support Coop Work [Internet]. 2006;15(1):29–53. Available from: http://dx.doi.org/10.1007/s10606-005-9010-z
[116] Scupelli PG, Xiao Y, Fussell SR, Kiesler S, Gross MD. Supporting coordination in surgical suites: Physical aspects of common information spaces. In: Proceedings of the 28th international conference on Human factors in computing systems - CHI '10. New York, New York, USA: ACM Press; 2010.
[117] Zhang Z, Sarcevic A. Coordination mechanisms for self-organized work in an emergency communication center. Proc ACM Hum Comput Interact [Internet]. 2018;2(CSCW):1–21. Available from: http://dx.doi.org/10.1145/3274468
[118] Ely JW, Graber ML, Croskerry P. Checklists to reduce diagnostic errors. Acad Med [Internet]. 2011;86(3):307–13. Available from: http://dx.doi.org/10.1097/ACM.0b013e31820824cd
[119] Niazkhani Z, Pirnejad H, Berg M, Aarts J. The impact of computerized provider order entry systems on inpatient clinical workflow: a literature review. J Am Med Inform Assoc [Internet]. 2009;16(4):539–49. Available from: http://dx.doi.org/10.1197/jamia.M2419
[120] Tu S, Peleg M. Section 5: Decision support, knowledge representation and management: Decision support, knowledge representation and management in medicine. Yearb Med Inform [Internet]. 2006;15(01):72–80. Available from: http://dx.doi.org/10.1055/s-0038-1638482
[121] Raj M, Wilk AS, Platt JE. Dynamics of physicians' trust in fellow health care providers and the role of health information technology. Med Care Res Rev [Internet]. 2021;78(4):338–49. Available from: http://dx.doi.org/10.1177/1077558719892349
[122] Filice RW, Ratwani RM. The case for user-centered artificial intelligence in radiology. Radiol Artif Intell [Internet]. 2020;2(3):e190095. Available from: http://dx.doi.org/10.1148/ryai.2020190095
[123] Hutchins E. Cognitive ecology. Top Cogn Sci [Internet]. 2010;2(4):705–15. Available from: http://dx.doi.org/10.1111/j.1756-8765.2010.01089.x
[124] Okolo CT. Optimizing human-centered AI for healthcare in the Global South. Patterns (N Y) [Internet]. 2022;3(2):100421. Available from: https://www.sciencedirect.com/science/article/pii/S2666389921003044
[125] Knowles B, Harding M, Blair L, Davies N, Hannon J, Rouncefield M, et al. Trustworthy by design. In: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. New York, NY, USA: ACM; 2014.