Practising Appropriate Trust in Human-Centred AI Design

Degachi, Chadha; Mehrotra, Siddharth; Yurrita, Mireia; Niforatos, Evangelos; Tielman, Myrthe

DOI
10.1145/3613905.3650825

Publication date
2024

Document Version
Final published version

Published in
CHI EA '24

Practising Appropriate Trust in Human-Centred AI Design

Chadha Degachi*
Siddharth Mehrotra*
c.degachi@tudelft.nl
s.mehrotra@tudelft.nl
Delft Univeristy Of Technology
Delft, The Netherlands

Mireia Yurrita
m.yurritasemperena@tudelft.nl
Delft Univeristy Of Technology
Delft, The Netherlands

Evangelos Niforatos
e.niforatos@tudelft.nl
Delft Univeristy Of Technology
Delft, The Netherlands

Myrthe Tielman
m.l.tielman@tudelft.nl
Delft Univeristy Of Technology
Delft, The Netherlands

ABSTRACT

Appropriate trust, trust which aligns with system trustworthiness, in Artificial Intelligence (AI) systems has become an important area of research. However, there remains debate in the community about how to design for appropriate trust. This debate is a result of the complex nature of trust in AI, which can be difficult to understand and evaluate, as well as the lack of holistic approaches to trust. In this paper, we aim to clarify some of this debate by operationalising appropriate trust within the context of the Human-Centred AI Design (HCD) process. To do so, we organised three workshops with 13 participants total from design and development backgrounds. We carried out design activities to stimulate discussion on appropriate trust in the HCD process. This paper aims to help researchers and practitioners understand appropriate trust in AI through a design lens by illustrating how it interacts with the HCD process.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**.

KEYWORDS

appropriate trust, human-centered design, AI design

ACM Reference Format:
Chadha Degachi, Siddharth Mehrotra, Mireia Yurrita, Evangelos Niforatos, and Myrthe Tielman. 2024. Practising Appropriate Trust in Human-Centred AI Design. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3613905.3650825

1 INTRODUCTION

While AI systems have vastly improved in accuracy and ability in recent years, they remain error-prone, *i.e.*, creating misuse caused by over-trust or disuse caused by under-trust in AI [31]. For effective collaboration between humans and AI systems, human trust in the system must be appropriate, enabling users to manage system risk [17]. For example, over-trusting an AI-based credit scoring system may lead a loan officer to accept an applicant when they should not, resulting in said applicant defaulting. Similarly, under-trusting it would involve dismissing the scores generated by the system as unreliable without proper investigation, and unjustly denying some applicants. Without appropriate trust in AI, people may not recognize the potential risks and limitations of AI, or be unable to understand and interpret the results of these systems.

Trust in AI is itself a complex topic, and understanding "appropriate trust" as well further complicates matters. No one definition of appropriate trust exists within the literature, with differing perspectives arising from the different backgrounds influencing the field. However, one commonly referenced *theoretical* definition is by Yang et al., who state that appropriate trust is the alignment of perceived and actual system performance [41]. Furthermore, they describe appropriate trust as related to users' ability to rely on the system when it is correct and to recognize when the system is incorrect. Meanwhile, Okamura and Yamada [29] and Ososky et al. [30], focus on the importance of balancing the benefits and risks of AI.

As a consequence of the theoretical complexities of defining appropriate trust, there is little consensus on how to *empirically* evaluate appropriate trust [39], e.g., there are currently few methodological contributions that shed light on how to measure (appropriate) trust [28]. Furthermore, prior studies have mostly evaluated the effect that different factors (e.g., explanations [3, 15, 38]) have on appropriate trust and related constructs through lab-based experiments with crowdworkers simulating end users. While these experimental settings generate a granular understanding of the antecedents of appropriate trust, they fail to generate a holistic view of designing for appropriate trust in AI [23]. Implications drawn from such focused studies might be difficult to translate into practice such that they advance Human-Centred AI Design (HCD) processes, which are uniquely difficult to craft [34, 42].

In this work, we opt for conducting generative co-design workshops [32] with current and future AI practitioners. A qualitative approach provides a deeper understanding and detail of our context, enabling us to explore subtle nuances of appropriate trust that

*Both authors contributed equally to this research.

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3650825

quantitative methods might overlook [14]. Along with our participants, we explore how to conceptualize appropriate trust and the way in which appropriateness of trust can influence HCD practices. Thus, our **RQ** is *What considerations do practitioners make when working with appropriate trust focused Human-Centred AI design processes?*

To this end, we conducted 3 workshops with 13 current and future (i.e., master and PhD students already involved in AI projects) AI practitioners with a background in either computer science or design. In those workshops, we explored the needs and priorities of stakeholders when designing for appropriate trust and examined the way in which accounting for those needs affects HCD processes.

Our work contributes to the CHI community by providing a holistic view of the impacts that designing for appropriate trust has on HCD processes. We draw implications for practice from the preliminary insights gained. We identified three lenses through which participants understood and used appropriate trust when working with the HCD process, **Development Ethicality, Communication of Trustworthiness, and Interactivity**. Using these lenses, we understood designing for appropriate trust to affect various aspects of the HCD process, from data concerns to stakeholder values.

2 BACKGROUND

2.1 Appropriate trust and its related concepts

Appropriate trust is a complex topic as it requires consideration of context's influence, the AI system's goal-related characteristics, and the cognitive processes that govern the development and erosion of trust [8, 20]. The current landscape of AI research predominantly emphasises building trust, without delving into the subtleties of appropriateness [18, 27]. This research gap represents a critical juncture where we aspire to bridge the divide and offer insights into the nuanced world of designing for appropriate trust, ensuring that AI systems align harmoniously with human needs and aspirations.

Concepts like "appropriate trust", "calibrated trust" and "appropriate reliance" are often used interchangeably in prior research [36]. There have been debates in the community about what appropriate trust is and how different concepts related to appropriate trust are different or similar [11]. There are more than ten concepts resonating with the concept of appropriate trust [26]. For example, calibrated trust is similar to the appropriate trust in that a human's trust belief about the agent corresponds to their actual trustworthiness. However, calibrated trust necessarily involves a process of trust calibration or trust alignment that corrects for over/under — trust through repeated interactions [40]. Other reoccurring concepts and principles within the literature include; aligning reliability and user trust [19], usage of (in)correct model prediction [40], reducing over/under — trust [10], and summing appropriate agreement and appropriate disagreement [25]. In summary, making clear distinctions can help reduce the discord among the community on approaching the concept of appropriate trust.

2.2 Designing for appropriate trust

Different approaches have been taken towards designing for appropriate trust, for example, providing explanations [12, 44]. Often, explanations are combined with confidence scores to help users align their perceived trustworthiness with the actual trustworthiness of the system [28]. Researchers have also explored uncertainty communication [16] and verbal assurances [2]. However, when and how these methods could be incorporated during the design process remains an unexplored area. We argue that, before understanding the impact which the use of these communication methods will cause, it is important to operationalise key requirements of appropriate trust by following a human-centred design framework.

Liao and Sundar [24] proposed the MATCH model to design for responsible trust[1], which describes how trustworthiness is communicated in AI systems through trustworthiness cues and how those cues are processed by people to make trust judgments. In the MATCH model, the authors describe users' cognitive processes in making trust judgments and their potential limitations based on human factors. Building up on the work of [24] by following a human-centred design framework, Sousa et al. [34] provide guidelines for mapping and defining user trust on the socio-ethical and organisational needs of AI system design. However, designing for appropriate trust is not investigated. Filling this literature gap, Jorritsma et al. [21] identified four ways (confidence ratings, performance level, global and local rationale) for improving the output of computer-aided diagnosis so that it enables more informed radiologist trust judgements. Similarly, Benda et al. [4] argue that appropriate reliance[2] can be fostered by knowing the purpose of a tool, its process for making recommendations, and its performance in the given context. Overall, we see researchers developing various frameworks for building appropriate trust in AI. However, a focus on incorporating those frameworks into the human-centred design cycle remains missing.

3 METHODOLOGY

In this section we summarize the procedure and materials used in our study (section 3.1), participant recruitment (section 3.2), and the steps we followed for data collection and analysis (section 3.3).

3.1 Procedure and Materials

Method and Use Case. For our study, we opted for a generative workshop within co-design methodologies [32]. Co-design methods are especially appropriate to engage stakeholders (in our case, current and future AI designers and developers) in technology design processes, and favour human-centred development approaches [7]. Previous work in AI co-design has successfully applied generative workshops to address their research questions [7, 33, 37]. We followed prior work [7] and designed a generative workshop to explore future design scenarios [22] where AI development processes were modified to design for appropriate trust. To this end, we focused on the context of AI for healthcare, as this is a highly sensitive context where mis- and dis-use can be very costly, and thus one where designing for appropriate trust is key. We designed a scenario where a Local Medical Practice contacted our Research Institution. The Local Medical Practice was interested in augmenting their expertise in first-point-of-contact patient support with

[1]Here, the concept "responsible trust" has been used in the place of appropriate trust. The authors use the framework of responsible trust to explore methods of empowering end users in making more accurate trust judgments.
[2]Tolmeijer et al. [36] inform us that although both trust and reliance are related, they should be treated and measured as independent concepts.

AI capabilities. The goal of this AI system was to support general practitioners and patients by providing AI supported long-term diabetes management plans based on user profiles in the form of recommendations and question-answering. See appendix A for the full use case description.

Workshop Structure and Materials. Before attending the workshop, our participants were encouraged to complete a home workbook where they would be introduced to some sensitizing activities [32] to get familiar with the concept of *appropriate trust*. They were also asked to reflect on everyday interactions with AI systems where appropriate trust played a role. The workshops were two-part, the first half focusing on appropriate trust, and the latter half focusing on more granular facets of appropriate trust (e.g., calibrated trust, contractual trust). In this work, we report on and analyse the first half of each workshop. During the first half of these workshops, we first presented a few slides to further help participants to get familiar with the concept of appropriate trust and the selected use case. We then divided participants into small groups where developers and designers would get mixed. These groups completed two activities. The first activity consisted in mapping primary, secondary, and tertiary stakeholders that should be involved in the design of the presented AI system. Participants also reflected on the aims of each stakeholder when designing for appropriate trust. The results of the activity were then shared with the rest of the groups. In the second activity, participants were asked to map the HCD process (from the project ideation stage to the maintenance stage) and to build a storyline involving the discussed stakeholders. The results of the second activity were again shared with the rest of the groups. During these discussions, we encouraged participants to reflect on the way designing for appropriate trust affected the design process. See Appendix B for examples of workshop materials in use. This procedure was approved by the Human Research Ethics Committee of Delft University of Technology (no. 3491).

3.2 Participants

We conducted 3 workshops with 4 to 5 participants each in December 2023. In total, 13 AI design and development graduate students participated in our workshops. For the recruitment of participants, we used purposive sampling. We were specifically interested in having participants with various backgrounds. Some that were more familiar with the technical design of AI artefacts (i.e., developers), and others that were more familiar with the design of appropriate human-AI interactions (i.e., designers). We posted recruitment advertising elements around our institution, shared the call for participation in AI emailing lists of our institution, and reached out to our personal contacts. We then selected participants to ensure diversity in their backgrounds. All participants were involved in AI projects, 6 of them being specifically involved in medical AI projects. Of our participant pool, 5 had a background in computer science or in technical fields, while the remaining participants had an industrial design background.

3.3 Data Collection and Analysis

Data Collection. The data we generated consisted of 1) results of the sensitizing activities in the home workbook, 2) mappings generated by participants in the first activity, 3) sketches of the design process generated in the second activity, and 4) transcripts of the discussions and final reflection.

Data analysis. We analysed data using *reflexive thematic analysis* [6, 9] with a combination of inductive and deductive orientation to data. We used reflexive thematic analysis because it is a flexible method that adequately adapts to the analysis of multi-modal data [6] that our research question required (e.g., oral discussions, workbook, design process visuals). Workbooks were analysed by identifying the definitions given by participants for appropriate trust. We also analysed the elements that participants deemed important when designing for appropriate trust and their prioritization. Stakeholder mappings and design processes were analysed by identifying the presence or absence of stakeholders and features in the Human-AI interaction design. Discussion data was analysed using the following workflow: 1) transcription of audio recordings, 2) familiarization with the material, 3) selective coding, generation of codes and code groups, 4) generation of themes, 5) review and refinement of codes. The main researchers analysed the data. Having multiple people analysing the data allowed us to develop rich insights into the data.

Statement of Positionality. The main researchers are personally in favour of designing human-AI interactions based on a more nuanced understanding of trust. The two first authors have previously argued in favour of understanding trust as a multi-faceted concept when designing AI systems [26].

4 RESULTS

We identified three lenses for understanding appropriate trust employed by participants throughout our study: **Development Ethicality, Communication of Trustworthiness, and Interactivity**, illustrated in Figure 1. Participants are quoted here with a numeral designation, e.g., P3.

4.1 Development Ethicality

Using this lens, participants understand appropriate trust as a user-system relationship enabled by the incorporation of ethical considerations during development. In this cluster, concepts around **autonomy, privacy, competence, and fitness to task** are discussed. A system development process for fostering appropriate trust as seen through this lens shows several properties. Firstly, it understands stakeholder needs around user empowerment and creates space for user autonomy and choice in the system. Secondly, it prioritizes protecting stakeholder privacy and takes into account the variance in what data stakeholders may wish to share. Third, it adopts a twofold understanding of system competence wherein accuracy is highly important, but moreover, system behaviour is in line with user expectations. Lastly, it is a process which very clearly understands user needs and values and works from there to detail technical requirements, such that system design is *appropriate* to its task.

4.1.1 Motivation for Development Ethicality. Participants gave many reasons as to why this is an important lens through which to interpret appropriate trust. These reasons revolved around removing triggers for under-trust and creating opportunities to foster trust. In the first grouping, the participants cited practices such as constant

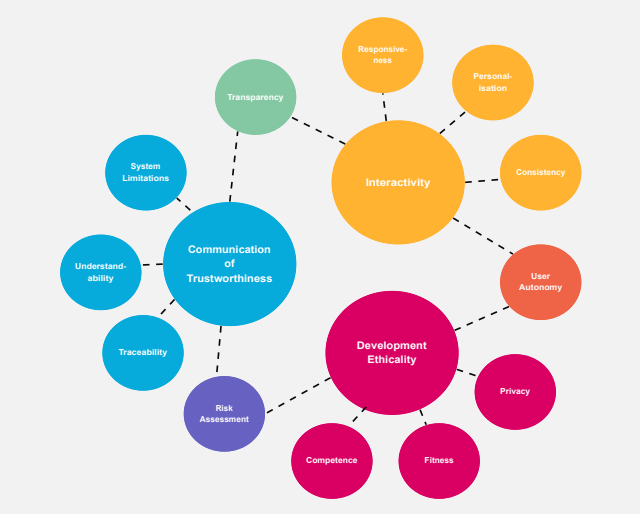

Figure 1: Lenses Identified For Understanding Appropriate Trust and Associated Concepts. Lenses Were Identified Through Reflective Thematic Analysis of Workshop Transcripts and Materials.

monitoring and data hoarding (*"[...] what are we going to solve here? So that if we know actually where we're going to depend on, [in the] AI [model], we precisely grab that data. We don't just grab whatever the data that the patient has."* [P1]). In the second group, participants discussed data transparency and user empowerment as trust-fostering (*"Maybe also having some patient group organisations and to make sure there's lobbying for not just what the General Practitioner (GP) needs, but that you also have the needs of the patient in this"* [P2]).

4.1.2 Impact on the Human-Centred Design Process. Many recommendations and best practices arose under this umbrella that affect HCD practice. Many of those recommendations are data concerns, such as improving data collection and sharing transparency and controllability, while reducing its invasiveness (*"[...] if you don't take your insulin [...] it'll be very clear. [...] and [the system] could have a warning thing of saying this patient you need to call them and say, hey you, we need to talk. Maybe they don't even see the data, but they'll just get that alarm notification"* [P2] *"Yeah, and it can start with like a notification to the patient itself. Like, hey, tak[e] your insulin and if it's not checked off or something, the GP just calls to check."* [P3] *"Yeah, that could be super nice that it's something where you know you're being watched, but only in the moment where you're clearly showing you're not well."* [P2]). However, we also see recommendations for increased user empowerment and ownership within the design process, and a need for a more critical understanding of power dynamics within participatory design, highlighting, in our use case, the possible imbalance between entities such as health insurers or food companies and individual patients as opposed to that between those same organisations and patient advocacy groups, (*"[...] the people with diabetes actually have ownership, so they get a say in saying, OK, what is it you need in your everyday life? What are the challenges that we could be solved [...]."* [P2]).

4.2 Communication of Trustworthiness

Here, appropriate trust is understood as a user-system relationship enabled by the trustworthiness signalling from the system to the user. In this cluster, concepts around **understandability, system limitations, and traceability** are examined. A system development process for fostering appropriate trust as seen through this lens shows several properties. Firstly, it discovers appropriate communication channels with users such that communication style adapts to user preference, background knowledge, and cognitive load. Second, it develops and communicates a very clear system scope such that users are aware of risks and limitations. Lastly, it provides traceable data processing pipelines in relation to where user data goes once collected, but also in relation to where model training data comes from.

4.2.1 Motivation for Communication of Trustworthiness. Proponents of this lens were motivated by its enabling of trust judgements, both through the communication of key information about the system and the recognition of the sense-making processes which decipher this information and their complexities. For example, user AI literacy and AI mastery were seen as supporting the learning process on which "system-use-judgements are made" (*"For me, appropriate trust is close to the notion that in some of HCI books you find it being called mastery of an AI system. So it's the point where you know how and when to use a particular technique. You know, there's sort of like learning and other understanding processes that might lead people to that one."* [P4]). Moreover, participants cited communicating system limitations as creating "well-informed" users with more accurate "belief-judgments" (*"What would be an ideal scenario that we have a patient that is very well-informed on how much to depend on AI and very well [...] able to assess when to believe and when not to believe. [...]. So having this continuous assessments like having this information on AI, what it can do and what it cannot do [...]."* [P1]). Meanwhile, the lack of communication on data traceability is discussed as a trigger for under-trust (*"If you're just putting in your data somewhere. It's very frustrating not knowing what happens to that data, or where it goes, or what that means, and not having an overview of whether or not you actually then put in your data correctly means that might also in general limit your trust."* [P3]).

4.2.2 Impact on the Human-Centred Design Process. The changes to the HCD process, put forth using this lens, were mostly process concerns. For example, training and education was recommended for improving user understanding of the system (*"But we think AI literacy is one of the important parts that can actually fill the [well-informed user] part."* [P1]), uncertainty communication for managing the user-system trust relationship (*"A reinforcement learning as a part of [mitigating overtrust]. But also just as a way of making it clear that the system doesn't know everything yet."* [P5]), and creating system investigation affordances for empowering users in tracing the use of their data (*"We all know, like you know the steps to check [T-shirts][3], you're trained on that, and it would be interesting if the app had the same ability of explainability of saying, you know the steps you can dive into further, but you also know you can stop at any point and still buy the T-shirt."* [P2]). Furthermore, understanding and designing for how human factors, besides cognitive processing

[3]The participant used clothing as a metaphor for AI systems here.

preferences and ability, impact the reception of system communications was highlighted as key to the modified HCD process (*"Like for me, maybe an embodied agent is more like more convincing than a website or something and also some human factors about, like how believable and also how natural the agents interact with users, makes it trustworthy."* [P6]).

4.3 Interactivity

Using this final lens, appropriate trust is understood as a user-system relationship enabled by the direct interaction with, and control over, the system. Under this umbrella, concepts such as **responsiveness, personalisation, and autonomy** are discussed. A system development process for fostering appropriate trust through this lens shows several properties. Feedback channels and human-in-the-loop workflows are a core part of such process. In this process, pluralism, flexibility, and adaptivity inform small scale design choices like flexible communication styles as well as large scale design choices like multiplicity of intervention pathways. Lastly, such a process enables control on the user-system interaction level, over both machine-generated insights and product-owner-generated data policies.

4.3.1 Motivation for Interactivity and Modifiability. Reasoning for adoption of this lens on appropriate trust mostly revolved around fostering trust in users. Controllability, for one, was noted as a method for improving user autonomy and fostering trust (*"Maybe the patient is particular, like wants to know how their diet is affecting their disease, then they are OK to share their own data about their diet to the platform because they care about it. But if I don't, then I don't share that data. So it can be more nuanced."*[P7]). Micro and macro level system adaptations and continuous interaction with monitoring bodies were similarly highlighted as fostering trust through personalisation and greater system oversight (*"Some people in the administration in hospitals that will have an additional role of "hey, we have a lot of complaints coming from patients because this part is being fed by AI and they maybe they are complaining that the referral system is not working""* [P7]). However, participants also noted that users may want to invest in systems' behaviour regardless of their trust relationship (*"The users and the AI they need to build up a positive feedback loop so that they kind of like can [be] constantly engaging, and like I think that's the way people are as well."* [P8]).

4.3.2 Impact on the Human-Centred Design Process. Design practices recommended through this lens also revolved around process concerns but also stakeholder values. On the process level, participants recommended designing accessible control options (*"So I think there is a different variation in type of people, and it could be nice if you had a slider. Literally to say like where — what do I want to know?"* [P2]), creating channels and a responsive to user feedback (*"Make me able to give feedback on things I perceive as wrong analysis, so I have more trust it will do right the next time"* [P9]), and creating opportunities for proactive context-seeking artificial agents to gain deeper knowledge of user needs (*"We ask the question, we get an answer, but it should be more. Something else? Why can't we iterate on not just getting an answer, with like a follow-up question? [...] Because basically, these models have so much data that's stored within them, and they want to understand another layer of your response that can make their answer more suitable."* [P7]). Stakeholder value considerations in this modified HCD process included pluralism, technology acceptability, interpretability, and user autonomy (*"so that's why we feel it's very important that we have it inclusive and diverse in the group, not just one, one particular country or not just one particular type of users with the lifestyle we need to have multiple types of which involved in the testing process and also the review process, not just the doctors, but also having the medical groups."* [P1]).

5 DISCUSSION

The thematic analysis of our sensitizing workbooks and workshop activities (stakeholder mapping, HCD process narrative building, and MoSCoW ranking) revealed three lenses through which appropriate trust can be viewed in the context of HCD. These lenses were **Development Ethicality, Communication of Trustworthiness, and Interactivity**.

Much of the focus of our implementing these lenses was on enabling user trust judgements. Designing and developing methods, such as explanations [12], confidence scores [44], and verbal assurances [2], which furnish the user with the information needed to make a trust decision is key to fostering appropriate trust. This finding aligns with the recommendations of Liao and Sundar [24], who propose the use of trustworthiness cues as key to fostering responsible trust in AI. However, similarly key, was designing these methods in ways which empower user sensemaking, whether that is through personalisation, accessible design, or increased user autonomy. We are now beginning to investigate how human factors affect the perception of certain trustworthiness cues like explanations [12, 35, 43] and much space remains for developing comprehensive frameworks of sensemaking in trust judgements.

Aside from enabling trust judgements, participants also conceptualised the appropriate trust process as reducing the triggers for over/under — trust. This aim reintroduces the issues we have previously discussed with adequately measuring trust. In their work, Sousa et al. [34] advocate for the measurement of trust a pillar of the design process, advocating for the use of the Gulati et al. [13] human-computer trust scale and the Assessment List for Trustworthy AI[4] to do so. However, the issues inherent to relying on self-reported data and expert assessments to the exclusion of real-time dynamic observed trust insights, though they may be non-trivial to obtain, remain. Looking to the work of Parasuraman and Riley [31] on the mis/dis — use of automation, we may direct practitioners to look for instances of inappropriate and unexpected use of AI tools, in order to identify these pitfalls.

Interestingly, much of the discussion around the **Development Ethicality** lens on appropriate trust aligns with research by social scientists on the same topic, such that the concept is defined in terms of alignment with human values and ethics, fairness, and impact on society [5]. Such convergence indicates moving outwards from the focus on understanding appropriate trust through system competence [29] may be necessary for successful operationalisation.

[4]Link:https://altai.insight-centre.org/

The limitations of this study stem from our small participant pool, which allowed us to explore our research question in depth, but not breadth. Further, we are limited by our use cases focus on the healthcare domain. We thus present a limited perspective on designing for appropriate trust, embedded within the cultural and social norms of our participants.

Future work will analyse the second half of our workshops where we disentangle some concepts around appropriate trust such as warranted trust [18], responsible trust [24], justified trust [1], and so on. We hope future work will shed light on how the multiplicity of terms in the study of appropriate trust is understood and used. Further, through analysing these associated terms in relation to the design process, we hope to extend and enrich the framework proposed here.

6 CONCLUSION

Based on this study, we highlight three lenses through which practitioners may understand and use appropriate trust in their HCD processes: **Development Ethicality, Communication of Trustworthiness, and Interactivity**. We discuss how each lens impacts the HCD process.

REFERENCES

[1] Arjun Akula, Shuai Wang, and Song-Chun Zhu. 2020. Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 2594–2601.

[2] Basel Alhaji, Michael Prilla, and Andreas Rausch. 2021. Trust dynamics and verbal assurances in human robot physical collaboration. *Frontiers in artificial intelligence* 4 (2021), 703504.

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[4] 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* 29, 1 (2022), 207–212.

[5] Nick Bostrom and Eliezer Yudkowsky. 2018. The ethics of artificial intelligence. In *Artificial intelligence safety and security*. Chapman and Hall/CRC, 57–69.

[6] V Braun and V Clarke. 2021. Thematic analysis: a practical guide [eBook version]. *SAGE moradi H, vaezi A. lessons learned from Korea: COVID-19 pandemic* 41 (2021), 873–4.

[7] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[8] Jessie YC Chen and Michael J Barnes. 2014. Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems* 44, 1 (2014), 13–29.

[9] Victoria Clarke and Virginia Braun. 2013. Successful qualitative research: A practical guide for beginners. *Successful qualitative research* (2013), 1–400.

[10] Sven Coppers, Davy Vanacken, and Kris Luyten. 2020. Fortniot: Intelligible predictions to improve user understanding of smart home behavior. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–24.

[11] Andrea Ferrario and Michele Loi. 2022. How Explainability Contributes to Trust in AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1457–1466. https://doi.org/10.1145/3531146.3533202

[12] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable active learning (XAL) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.

[13] Siddharth Gulati, Sonia Sousa, and David Lamas. 2019. Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology* 38, 10 (Oct. 2019), 1004–1015. https://doi.org/10.1080/0144929x.2019.1656779

[14] Beverley Hancock, Elizabeth Ockleford, and Kate Windridge. 2001. *An introduction to qualitative research*. Trent focus group London.

[15] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. 7, CSCW2, Article 276 (Oct. 2023), 29 pages. https://doi.org/10.1145/3610067

[16] Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. 2013. Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In *Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications*. 210–217.

[17] Robert R Hoffman. 2017. A taxonomy of emergent trusting in the human–machine relationship. *Cognitive Systems Engineering* (2017), 137–164.

[18] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 624–635.

[19] Theodore Jensen, Mohammad Maifi Hasan Khan, Md Abdullah Al Fahim, and Yusuf Albayram. 2021. Trust and anthropomorphism in tandem: the interrelated nature of automated agent appearance and reliability in trustworthiness perceptions. In *Designing interactive systems conference 2021*. 1470–1480.

[20] Carolina Centeio Jorge, Emma M van Zoelen, Ruben Verhagen, Siddharth Mehrotra, Catholijn M Jonker, and Myrthe L Tielman. 2024. Appropriate context-dependent artificial trust in human-machine teamwork. In *Putting AI in the Critical Loop*. Elsevier, 41–60. https://doi.org/10.1016/B978-0-443-15988-6.00007-8

[21] Wiard Jorritsma, Fokie Cnossen, and Peter MA van Ooijen. 2015. Improving the radiologist–CAD interaction: designing for appropriate trust. *Clinical radiology* 70, 2 (2015), 115–122.

[22] Robert Jungk and Norbert Müllert. 1987. *Future Workshops: How to create desirable futures*. Inst. for Social Inventions.

[23] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (CHIcago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 77–88. https://doi.org/10.1145/3593013.3593978

[24] Q Vera Liao and S Shyam Sundar. 2022. Designing for responsible trust in AI systems: A communication perspective. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1257–1268.

[25] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–45.

[26] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M Jonker, and Myrthe L Tielman. 2023. A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction. *arXiv preprint arXiv:2311.06305* (2023).

[27] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M Jonker, and Myrthe L Tielman. 2023. Building Appropriate Trust in AI: The Significance of Integrity-Centered Explanations.. In *HHAI*. 436–439.

[28] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M Jonker, and Myrthe L Tielman. 2023. Integrity Based Explanations for Fostering Appropriate Trust in AI Agents. *ACM Transactions on Interactive Intelligent Systems* (2023).

[29] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *Plos one* 15, 2 (2020), e0229132.

[30] Scott Ososky, David Schuster, Elizabeth Phillips, and Florian G Jentsch. 2013. Building appropriate trust in human-robot teams. In *2013 AAAI spring symposium series*.

[31] Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors* 39, 2 (June 1997), 230–253. https://doi.org/10.1518/001872097778543886

[32] Elizabeth B-N Sanders and Pieter Jan Stappers. 2012. *Convivial toolbox: Generative research for the front end of design*. BIS.

[33] Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices*. CRC Press.

[34] Sonia Sousa, Jose Cravino, Paulo Martins, and David Lamas. 2023. Human-centered trust framework: An HCI perspective. *arXiv preprint arXiv:2305.03306* (2023).

[35] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(Iui '21)*. Association for Computing Machinery, New York, NY, USA, 109–119. https://doi.org/10.1145/3397481.3450662

[36] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. Capable but amoral? Comparing AI and human expert collaboration in ethical decision making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[37] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability for content moderation. *Proceedings of the ACM on human-computer interaction* 5, CSCW2 (2021), 1–28.

[38] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (April 2023), 38 pages. https://doi.org/10.1145/3579605

[39] 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 (Oct. 2021), 39 pages. https://doi.org/10.1145/3476068

[40] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*. 318–328.

[41] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.

[42] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301

[43] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effects of Explanations, Human Oversight, and Contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 134, 21 pages. https://doi.org/10.1145/3544548.3581161

[44] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. 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* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

A USE CASE DESCRIPTION

This appendix documents the use case description shared with our participants before the workshop.

A.1 Description

The SGZ-Studentengezondheidszorg in Delft have contacted TU Delft because they are interested in augmenting their expertise in first-point-of-contact patient support with AI capabilities. They are hoping to support GPs and patients by providing AI supported long term diabetes management plans based on user profiles.

Diabetes is one of the most important chronic diseases that threatens public health. Since 2000, the prevalence of diabetes has more than tripled, and by 2021, more than 530 million people worldwide will have diabetes. The main goal of diabetes management is to maintain glycemic control within the target range, which is often accomplished through lifestyle modification and the self-monitoring of blood glucose (SMBG) in patients with type 2 diabetes mellitus (T2DM). However, maintaining glycemic control is challenging for both patients and health care providers (HCPs) because it is difficult to encourage or motivate patients to make long-term lifestyle changes, interpret their SMBG data, and provide immediate feedback and understand the patients' lifestyle due to brief clinic visit times and long visit intervals.

Together, TU Delft and SGZ will develop a high effective digital intervention for personalized diabetes management support over the course of one year. This application will be accessible from mobile phones and personal computers with features such as diet tracking and recommendation, exercise tracking and recommendation, medication tracking, and medical question answering. Moreover, they are interested in developing a system which doctors and patients are happy to use and adopt into their workflow, so they are adopting a human-centred approach to development.

Figure 2: A Stakeholder Mapping from Two Participants in Our Second Workshop

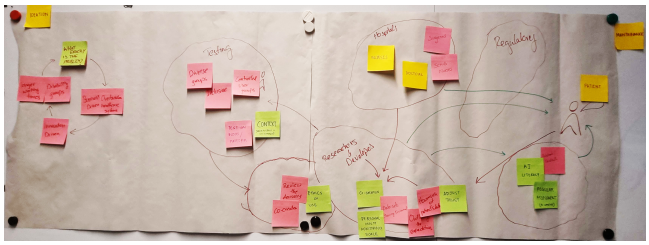

Figure 3: A Design Process From Two Participants in Our First Workshop

As designers and developers, you are responsible for the implementation of this system. You have access to the [Local Health Organisation] staff and resources, as well as [Research Institution] staff and resources.

B WORKSHOP MATERIALS

This appendix showcases the materials and templates employed in this study's workshop.

B.1 Stakeholder Mapping

A template with three concentric circles was provided for the first workshop activity to allow participants to think about the stakeholders involved in our use case (See Figure 2).

B.2 Design Process

A template with a vaguely defined design process was provided for design process storyline activity (See Figure 4). Many participating groups, however, decided to sketch the design process from scratch (See Figure 3).

B.3 MOSCOW Ranking

A template with four columns marked Must-Have, Should-Have, Can-Have, and Won't-Have. This design activity is designed to help participants rank and prioritize properties and elements of their design processes. (See Figure 5).

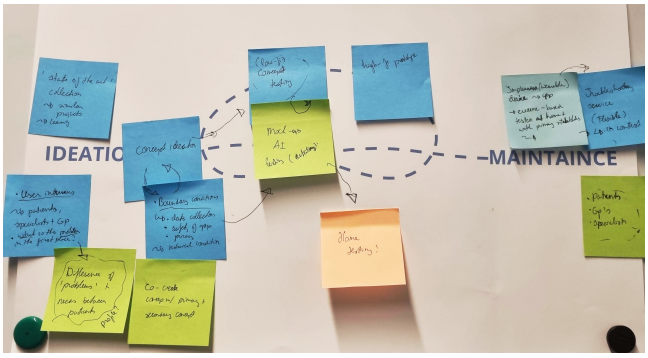

Figure 4: A Design Process From Two Participants in Our Third Workshop

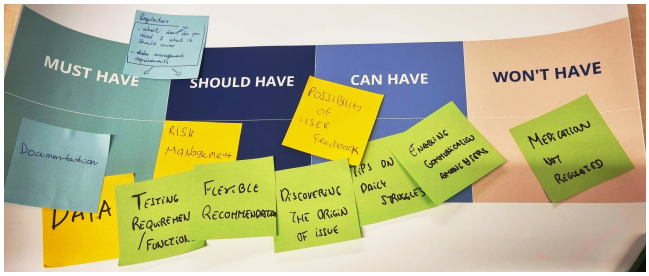

Figure 5: A MoSCoW mapping From our Second Workshop