Building Appropriate Trust in Human-AI Interactions

Fatemeh Alizadeh, Oleksandra Vereschak, Dominik Pins, Gunnar Stevens, Gilles Bailly, Baptiste Caramiaux

Fatemeh Alizadeh, Oleksandra Vereschank, Dominik Pins, Gunnar Stevens, Gilles Bailly, Baptiste Caramiaux (2022): Building Appropriate in Human-AI Interactions. In: Proceedings of the 20th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing on the Design of Cooperation Technologies - Workshops, Reports of the European Society for Socially Embedded Technologies (ISSN 2510-2591), DOI: 10.48340/ecscw2022_ws04

Building Appropriate Trust in Human-AI Interactions

Fatemeh Alizadeh, Gunnar Stevens
University of Siegen
Fatemeh.alizadeh@uni-siegen.de

Oleksandra Vereschak, Gilles Bailly, Baptiste Caramiaux
Sorbonne University
vereschak@isir.upmc.fr

Dominik Pins
Fraunhofer Institute for Applied Information Technology
dominik.pins@fit.fraunhofer.de

Abstract. AI (artificial intelligence) systems are increasingly being used in all aspects of our lives, from mundane routines to sensitive decision-making and even creative tasks. Therefore, an appropriate level of trust is required so that users know when to rely on the system and when to override it. While research has looked extensively at fostering trust in human-AI interactions, the lack of standardized procedures for human-AI trust makes it difficult to interpret results and compare across studies. As a result, the fundamental understanding of trust between humans and AI remains fragmented. This workshop invites researchers to revisit existing approaches and work toward a standardized framework for studying AI trust to answer the open questions: (1) What does trust mean between humans and AI in different contexts? (2) How can we create and convey the calibrated level of trust in interactions with AI? And (3) How can we develop a standardized framework to address new challenges?

Introduction

Artificial intelligence (AI) plays an important role in helping people make sensitive decisions with uncertain outcomes. Yet the inner workings of AI-powered systems are often hidden from users. These opaque processes have been criticized as biased, discriminatory, and misleading, and users cannot be assured that their interests are respected (Eslami et al., 2019). However, building a collaborative partnership between human decision makers and AI-powered systems depends primarily on users' trust in the systems (Vereschak et al., 2021). In general, Human-machine trust can be defined as, *"An attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability"*(Lee & See, 2004).

Since AI is a broad term that has never represented a single technology in a specific time period (Alizadeh et al., 2021), the question arises whether this general definition of trust between humans and machines is still applicable to all types of systems under this umbrella term. Especially because trust in AI-enabled systems has been shown to be context-dependent. In the context of voice assistants, for example, trust has been shown to evolve around user privacy concerns (Završnik, 2021), while in medical systems, trustworthiness is equated with the accuracy of the system and its outcomes (Ghassemi et al., 2018). Moreover, previous approaches to building and assessing trust tend to be binary. That is to say, there is a lack of research on the multidimensional nuances that must be considered in long-term interactions with AI-enabled systems (Hoffman, 2017).

In this workshop, we aim to explore these challenges by enabling researchers and practitioners in the field to move toward a more flexible and standardized framework that accounts for these differences and promotes a shared understanding of the notion of human-AI trust across different contexts and applications of AI.

Background

In this section, we describe trust in the context of human interactions with AI-powered systems and address the challenges of establishing and evaluating trust. The questions we raise are not necessarily new, but are nonetheless relevant because they have not been satisfactorily answered for emerging cases. While we do not wish to limit the workshop to these challenges, we believe they are and will be important in past, current, and future research.

Human-AI trust

AI is being used to develop algorithms that increasingly make decisions about our daily lives. They decide for us what we read, what we watch, what we buy, and even who we date (Fry, 2018). However, AI algorithms are becoming increasingly opaque. Such a black box makes it difficult for users to understand, verify, or trust these potentially biased systems (Eslami et al., 2019). The demand for transparency and the need for users to trust AI-embedded systems has not only led to the European Commission issuing detailed guidance on the requirements for trustworthy AI models (Smuha, 2019), but has also led HCI researchers to investigate how to develop and ensure trustworthy AI. As a result, previous work has examined the factors that influence user trust (e.g., Cai et al., 2019; Robert Jr, 2016), how trust is established (e.g., Al-Ani et al., 2013; Passi & Jackson, 2018), and how it can be modeled (e.g., Ajenaghughrure et al., 2019; Knowles et al., 2015). Jacovi et al. have leveraged these requirements and combined them with standard research documents and explanatory methods to specify a set of useful contracts, namely (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination, fairness, (6) societal and environmental well-being, and (7) accountability (Jacovi et al., 2021). According to the authors, the European Commission's guideline is based on the premise that trust is the ability to anticipate intended behavior through the belief that a contract will be upheld. Therefore, an AI model is trustworthy with respect to a contract if it is able to honor that contract (Jacovi et al., 2021). But how can the guidelines for trustworthy AI be used to establish appropriate trust in AI-embedded technologies, and what are the challenges in this process?

Designing appropriate Human-AI trust

Researchers have argued that trust and trustworthiness are completely decoupled. For example, Ghassemi et al. have shown that physicians' trust in a tool can be increased by making changes to the tool's user interface without changing the tool's trustworthiness (Ghassemi et al., 2018). To clarify this disentanglement, Jacovi et al. distinguished between warranted trust and unwarranted trust. In this context, they defined warranted trust when trust is calibrated with trustworthiness and users do not feel betrayed because they trusted a model that was not trustworthy (Jacovi et al., 2021). Calibrating trust for trustworthiness is critical to avoid the risk of misuse, abuse or disuse of technology (Parasuraman & Riley, 1997).

However, there are several challenges to establishing appropriate Human-AI trust in practice. First, while AI is a broad umbrella term(Alizadeh et al., 2020), trust in AI is context-dependent (Vereschak et al., 2021). People can trust one thing in one context, but not in another (Hoffman, 2017). This is particularly important because different requirements are assigned different value in different contexts. For example, while privacy and data governance are the main important requirements for adoption of personal assistant systems (Liao et al., 2019), fairness and non-discrimination are much more important for AI decision-making systems for criminal justice (Završnik, 2021). This has led to different research focuses and approaches to trust in different applications of AI-embedded systems, making it difficult to interpret results and compare across studies.

Another challenge is that trust cannot be viewed in binary terms, but is multidimensional and changes over time and throughout the course of an interaction. Hoffman elaborates: "*In my own relation to my word processing software, I am positive that it will perform well in the crafting of simple documents, but I am simultaneously confident it will crash when the document gets long, or when it has multiple high-resolution images. And every time that there is a software upgrade, the trusting of many of the functions becomes tentative and skeptical. [So,] trust is not a single state*"(Hoffman, 2017). This suggests that even within the same context, we need models that account for the nuances of trust throughout the interaction process, rather than relying on single states.

Moreover, previous research has defined the boundary between interpersonal trust and human-machine trust in terms of reparability (Hoffman, 2017; Jacovi et al., 2021). That is, unlike interpersonal trust, which can be restored after a mistake, users lose their trust in the machine completely when it makes a mistake, with no opportunity to forgive it (Hoffman, 2017). However, further research shows that in some cases, users are able to forgive and accept the mistakes of AI-enabled technologies. For example, users of voice assistants have been shown to develop a sense of tolerance for miscommunication with their devices and to forgive their mistakes (Lahoual & Frejus, 2019). Thus, there is a need to explore useful mechanisms to restore trust in case of errors and loss of trust. Having said all this, the question remains how we can overcome these challenges to build and restore appropriate trust in human-AI interactions.

Workshop Goal

As approaches to experiences with building trust differ, we aim to find a common ground, based on the shared experiences from the field. In addition to finding possible solutions, we want to give participants the opportunity to connect and collaboratively work further on the discussed topics. Together, we want to rethink existing binary approaches and start working on a nuanced model, that better serves the needs of specific circumstances.

Organizers

Fatemeh Alizadeh (main contact) is a PhD student and research associate at the Institute for Information Systems and New Media, University of Siegen. In her research, she combines her knowledge in HCI with her computer engineering and AI background to study unexpected situations with intelligent systems. Her main research interest is to improve the understandability, explainability and trustworthiness of AI-embedded technologies.

Oleksandra Vereschak is a PhD student at ISIR, Sorbonne Université. Her main focus of interest is users' trust in AI, which situates her work in the interdisciplinary domain of Human-AI interaction. She predominantly focuses on the AI-based systems assisting human decision making in the high-risk contexts such as medical, recruiting, and credit decision making. She studies not only what influences human trust, but also how to improve experimental protocols to evaluate it drawing from her social sciences background.

Dominik Pins is a PhD student and a research associate at Fraunhofer Institute for Applied Information Technology (FIT) in the department of Human-Centered Engineering and Design. As a usability engineer and research associate with sociological background he focuses in his research on user needs and practices regarding trust and privacy in the home environment and the design of trustworthy technologies, specifically AI systems.

Gunnar Stevens is a Professor of Information Systems at the University of Siegen and Co-Director of the Institute for Consumer Informatics, Bonn-Rhein-Sieg University of Applied Sciences. He has been researching and publishing in the fields of HCI, CSCW, Usable Security and Digital Consumer Protection for years. For his research he received the IBM Eclipse-Innovation Award in 2005 and the PhD Award of the IHK Siegen-Wittgenstein in 2010.

Gilles Bailly is a CNRS researcher at ISIR, Sorbonne Université. His research is at the crossroad of human-computer interaction (HCI), skill acquisition, decision making, artificial intelligence (AI) and robotics. He designs novel interaction techniques (desktop interaction, mobile interaction, gestural interaction, etc.) and builds predictive models of performance and knowledge with a focus on the transition from novice to expert behavior.

Baptiste Caramiaux is a CNRS researcher at ISIR, Sorbonne Université. He conducts research in human-computer interaction (HCI), examining how machine learning (or artificial intelligence) algorithms can be used in various fields such as performing arts, health or pedagogy. He is particularly interested in learning technologies when they are integrated with communities of practice. In particular, he sees technology as a reflective tool that allows people to question their practice, learn, and express themselves.

Each of the organizers has a research background in transparency, explainability and trust of AI-embedded systems, and has in particular experienced the challenges and struggles of building and exploring trust in human-AI teams. It was through the sharing of these experiences among the co-organizers that this workshop was initiated. Each organizer will present their own position and research in the introduction of the workshop to start the discussion and open the floor for the participants.

Pre-Workshop Plans

The workshop will be promoted through a new website that will communicate the aims and structure of the upcoming event, and subsequently present its outcomes. By spreading the websites through a broad variety of mailing lists as well as personal contacts, the workshop will reach researchers, activists and practitioners. Candidates will be required to submit a position paper discussing their current, previous or planned work. These papers can be in immediate relation to trust in voice interaction design or they can be an example of work which was challenging with regard to the mentioned topics. We envisage a maximum of 10 participants (excluding the organizers), who will be selected based on the relevance and potential contribution of their position paper to the workshop topic and activities. The quite small number of participants will ensure a relaxed and safe environment to talk about sensitive topics.

Workshop Plan

We plan to hold an interactive workshop, during which the participants will mostly work on different tasks and questions instead of just presenting their previous and current work. The workshop will begin with an ice-breaker and short introductions before the morning coffee break. Following the morning coffee and lunch breaks, participants will work in small groups, formed based on their position papers and research interests. The aim is to share experiences and identify common aspects and workarounds of designing trust in voice interactions. Participants are invited to critique and rethink current concepts, methods and frameworks building trust that do not address the arising challenges. The outcome from the group sessions will be shared in a plenary after the afternoon coffee break, with a view to formulating more viable and practical approaches for designing trust with a focus on long-term voice interactions. The workshop will conclude with a plenary discussion of future plans for a collaboration on the further development of these guidelines.

Timetable

Timeslot	Activity
09:00-09:15	Welcome
09:15-10:00	Icebreaker and short presentation of participants
10:00-10:30	Coffee break
10:30-12:00	Identifying and discussing challenges of building and evaluating appropriate trust in human-AI interaction and the existing approaches
12:00-13:30	Lunch
13:30-15:00	Formulating possible solutions in small groups
15:00-15:30	Coffee break
15:30-17:00	Presentation and discussion of the formulated approaches
17:00-17.15	Closing of the day and future plans

Post-Workshop Plan

All the notes, documentation and other materials that are created during the discussions will be shared amongst the workshop participants and revised, prior to being uploaded to the workshop website. Follow-up workshops on other conferences will help this newly formed collaboration to continue, through discussions and new initiatives, thereby encouraging more researchers to reflect upon their own challenged they come across when building trust in voice interactions. In addition, the workshop participants should be become part of exchange group which should serve as support line when help is needed dealing with an uncommon situation.

Call for Participation

This one-day workshop aims to provide a forum for researchers as well as practitioners and activists to discuss challenges in building trust and to start working on solutions that are more practical and viable to adapt in the AI interaction context. The topics include but are not limited to:

- Definitions of trust and reliance.
- Interpersonal trust and lessons from social sciences.
- Qualitative and quantitative methods for building and evaluating trust.
- Challenges of designing appropriate trust and tradeoffs with other objectives.

- Solutions (and their limitations) for promoting appropriate trust (e.g., XAI, control mechanisms, human agency, communicating uncertainty etc).
- Safety mechanisms for when trust is broken.

We invite anyone interested in participating to submit a two to four-page position paper. Papers should critically reflect upon the authors' experiences from the field or research area related to challenges they face when building trust in AI interactions. Authors' prior experience does not have to be specifically concerned with these challenges, but the position papers will be expected to demonstrate how their experience is relevant to the workshop's topic and can be applied within the workshops' context.

Submissions should be sent to Fatemeh.alizadeh@uni-siegen.de in .pdf format. Position papers will be reviewed based on relevance and potential for contribution to the workshop. At least one co-author of each accepted paper must register to the ECSCW 2022 conference to attend the workshop.

References

Ajenaghughrure, I. B., Sousa, S. C., Kosunen, I. J., & Lamas, D. (2019). Predictive model to assess user trust: A psycho-physiological approach. *Proceedings of the 10th Indian Conference on Human-Computer Interaction*, 1–10.

Al-Ani, B., Bietz, M. J., Wang, Y., Trainer, E., Koehne, B., Marczak, S., Redmiles, D., & Prikladnicki, R. (2013). Globally distributed system developers: Their trust expectations and processes. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 563–574.

Alizadeh, F., Esau, M., Stevens, G., & Cassens, L. (2020). *eXplainable AI: Take one Step Back, Move two Steps forward*. https://doi.org/10.18420/muc2020-ws111-369

Alizadeh, F., Stevens, G., & Esau, M. (2021). I don't know, is AI also used in airbags? *I-Com*, *20*(1), 3–17.

Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F., Corrado, G. S., & Stumpe, M. C. (2019). Human-centered tools for coping with imperfect algorithms during medical decision-making. *Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems*, 1–14.

Eslami, M., Vaccaro, K., Lee, M. K., Elazari Bar On, A., Gilbert, E., & Karahalios, K. (2019). User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300724

Fry, H. (2018). *Hello World: How to be Human in the Age of the Machine*. Random House.

Ghassemi, M., Pushkarna, M., Wexler, J., Johnson, J., & Varghese, P. (2018). Clinicalvis: Supporting clinical task-focused design evaluation. *ArXiv Preprint ArXiv:1810.05798*.

Hoffman, R. R. (2017). A taxonomy of emergent trusting in the human–machine relationship. *Cognitive Systems Engineering: The Future for a Changing World*, 137–164.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 624–635. https://doi.org/10.1145/3442188.3445923

Knowles, B., Rouncefield, M., Harding, M., Davies, N., Blair, L., Hannon, J., Walden, J., & Wang, D. (2015). Models and patterns of trust. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 328–338.

Lahoual, D., & Frejus, M. (2019). When users assist the voice assistants: From supervision to failure resolution. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–8.

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Hum. Factors*. https://doi.org/10.1518/hfes.46.1.50.30392

Liao, Y., Vitak, J., Kumar, P., Zimmer, M., & Kritikos, K. (2019). Understanding the Role of Privacy and Trust in Intelligent Personal Assistant Adoption. In N. G. Taylor, C. Christian-Lamb, M. H. Martin, & B. Nardi (Eds.), *Information in Contemporary Society* (Vol. 11420, pp. 102–113). Springer International Publishing. https://doi.org/10.1007/978-3-030-15742-5_9

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253.

Passi, S., & Jackson, S. J. (2018). Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–28.

Robert Jr, L. P. (2016). Monitoring and trust in virtual teams. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 245–259.

Smuha, N. A. (2019). The EU approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International*, *20*(4), 97–106.

Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–39. https://doi.org/10.1145/3476068

Završnik, A. (2021). Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of Criminology*, *18*(5), 623–642.