What AI Practitioners Say about Human-AI Trust: Its Role, Importance, and Factors That Affect It
Oleksandra Vereschak, Gilles Bailly, Baptiste Caramiaux

To cite this version:
Oleksandra Vereschak, Gilles Bailly, Baptiste Caramiaux. What AI Practitioners Say about Human-AI Trust: Its Role, Importance, and Factors That Affect It. International Conference on Hybrid Human-Artificial Intelligence, Jun 2022, Amsterdam, Netherlands. hal-03679043

HAL Id: hal-03679043
https://hal.sorbonne-universite.fr/hal-03679043
Submitted on 25 May 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
What AI Practitioners Say about Human-AI Trust: Its Role, Importance, and Factors That Affect It

Oleksandra VERESCHAK

Gilles BAILLY and Baptiste CARAMIAUX

Sorbonne Université, CNRS, ISIR, Paris, France

Trust has become a priority when designing and deploying AI-embedded systems alongside with other Human-Centered AI values, i.e. explainability, transparency, and fairness. However, due to their multifaceted and multidisciplinary nature, these terms can have various context-dependent meanings. Thus, translating these values into design can be a challenge [6]. Trust is not an exception.

Understanding what Human-AI trust is and what factors affect it comes largely from controlled lab experiments or studies with prototypes of AI-embedded systems [3, 7]. However, little is known about how Human-AI trust is addressed in development and deployment of real-world AI products and services. AI practitioners, people involved in different aspects of system design and deployment in the field, with roles ranging from AI developers to project managers and policy makers, can shed light on the role of Human-AI trust and what Human-AI trust factors are considered in real organizational settings. Their insights can better detail the needs, challenges, and experiences of different stakeholders when it comes to Human-AI trust.

In this work-in-progress paper, we study how Human-AI trust is addressed in development and deployment of real AI systems. We conduct a series of interviews with AI practitioners who develop and deploy AI-embedded decision support systems in various risk-sensitive contexts (finance, law, management). We specifically focus on these systems, because human trust in AI is especially pertinent for them due to their potential societal impact. The interviews are part of a bigger project around AI practitioners’ experiences with Human-AI trust, but in this working paper we report the preliminary findings from the first 5 interviewees (see Table 1). Specifically, we present our preliminary analysis of participants’ replies to the questions regarding the role of Human-AI trust in their practices and what factors are considered when establishing it in the context of AI-assisted decision making.

For the results’ analysis, two independent reviewers read all the interviews at least two times and independently identified phrases of interests and codes for them, following the thematic analysis approach [1]. Together, they compared and finalized the list of selected phrases, and fine-tuned codes’ formulation. By grouping the codes, the reviewers identified three major themes: 1) the role of Human-AI trust in developing and designing AI-embedded decision support systems, 2) importance of Human-AI trust in AI practitioners’ work, and 3) what factors AI practitioners believe contribute to establishing trust in their systems.

1 Corresponding Author: vereschak@isir.upmc.fr
June 13-17 2022

Table 1. Characterization of participants, their companies, and AI they work with.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Role</th>
<th>Background</th>
<th>Organization Size</th>
<th>Type of AI</th>
<th>AI Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Explainable AI implementation and research</td>
<td>Computer Science and Mathematics</td>
<td>Large</td>
<td>Convolutional neural networks</td>
<td>Transport management, paleontology classification</td>
</tr>
<tr>
<td>P2</td>
<td>Explainable AI implementation and research</td>
<td>Engineering and Mathematics</td>
<td>Small</td>
<td>Operations research</td>
<td>Task planning</td>
</tr>
<tr>
<td>P3</td>
<td>President of the company</td>
<td>Mathematics</td>
<td>Small</td>
<td>Supervised learning (classic and homebacked)</td>
<td>Evaluation of law cases</td>
</tr>
<tr>
<td>P4</td>
<td>Senior research project manager</td>
<td>Human-Computer Interaction</td>
<td>Large</td>
<td>Operations research, supervised and unsupervised learning</td>
<td>Project-based</td>
</tr>
<tr>
<td>P5</td>
<td>Senior research project manager</td>
<td>Psychology and Human Factors</td>
<td>Large</td>
<td>No data due to interview time constraints</td>
<td>Project-based</td>
</tr>
</tbody>
</table>

Our preliminary findings firstly indicate that while Human-AI trust is viewed as a commercial advantage to retain users (P2, P4), it is left behind as a research topic for later with obtaining better system’s performance and AI certification as the main focus (P1, P4). Secondly, Human-AI trust plays an important role when there is risk associated with a decision and when the task is complex. However, the definition of risk remains an open question (P2, P4, P5).

Lastly, AI practitioners consider the following factors when establishing Human-AI trust: AI performance and error, explainability of AI, Human-Human trust, and interaction with AI over time. The effect of AI errors on users’ trust varies depend on three nuances – context (testing phase vs scaled deployment, P1), the errors’ frequency rather than their existence (P1, P2), and the extent to which AI recommendation is surprising (P1, P2). The importance of AI explanations for trust depends on context and users (P2, P3). For example, explanations truly matter for Human-AI trust when users have doubts about the decision (P2). P1 sees explanations as a tool for developers (but not direct users) to understand their AI models better and, as a consequence, calibrate developers’ trust in the AI recommendations. Furthermore, Human-Human trust, for example, trust between users and developers (P2, P3, P4) and between users and model subjects (those affected by the Human-AI decision making, P2) also affect establishment of Human-AI trust. Lastly, Human-AI trust is not static, it evolves over time as users map out a more robust mental model (P1, P4).

These preliminary results have implications for future research directions around Human-AI trust and for design and deployment of AI-embedded systems. Firstly, as risk is one of the prerequisites of Human-AI trust [7, 4], we have to clearly define it as well as understand what other elements could trigger establishment of trust in Human-AI interaction. Nuanced specifications around AI performance and errors (e.g. frequency, surprise) and AI explanations (context- and user-dependent) promise interesting research opportunities and echoes Lai’s et al. [5] call for detailing context in experimental studies with AI in a standardized manner. The importance of Human-Human trust for Human-AI one can be further investigated through design concepts like social transparency elaborated by Ehsan et al. [2], where users become aware of the experiences and actions of other users with the same AI system. Additionally, since Human-AI trust evolves over time, it is beneficial to monitor how users’ trust in AI changes after deployment on a continuous basis. To further advance our fundamental understanding of Human-AI trust, we plan on interviewing additional AI practitioners as well as other stakeholders of related AI-embedded systems in the context of decision making and beyond.
June 13-17 2022

References