On the Way to Improving Experimental Protocols to Evaluate Users’ Trust in AI-Assisted Decision Making

Oleksandra Vereschak, Gilles Bailly, Baptiste Caramiaux

To cite this version:


HAL Id: hal-03418712
https://hal.sorbonne-universite.fr/hal-03418712
Submitted on 8 Nov 2021

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.
On the Way to Improving Experimental Protocols to Evaluate Users' Trust in AI-Assisted Decision Making

OLEKSANDRA VERESCHAK, Sorbonne Université, CNRS, ISIR, France
GILLES BAILLY∗, Sorbonne Université, CNRS, ISIR, France
BAPTISTE CARAMIAUX∗, Sorbonne Université, CNRS, ISIR, France

The spread of AI-embedded systems involved in human decision making makes it critical to build these systems according to trustworthiness standards. To understand whether this goal was achieved, users’ trust in these systems must be studied. However, empirically investigating trust is challenging. One reason is the lack of standard protocols to design trust experiments. To get an overview of the current practices in the experimental protocols for studying trust in the context of AI-assisted decision making, we conducted a systematic review of such papers. We annotated, categorized, and analyzed them along the constitutive elements of an experimental protocol (i.e., participants, task). Drawing from empirical practices in social and cognitive studies on human-human trust, we provide practical guidelines and research opportunities to improve the methodology of studying Human-AI trust in decision-making contexts. In this workshop, we would like to start the discussion about how these guidelines and research questions can be used in the laboratory and in the wild.

CCS Concepts: • Human-centered computing → HCI theory, concepts and models.

Additional Key Words and Phrases: trust, artificial intelligence, decision making, methodology

ACM Reference Format:

1 INTRODUCTION

Artificial Intelligence (AI) has acquired a critical role in assisting humans in making sensitive decisions such as hiring [9], treatment assignment [8], or criminal investigation processes [22], to name a few. In such situations, humans make decisions based on their own expertise and on recommendations provided by an AI-based algorithm (e.g. data-driven models, knowledge-based models, etc.), which we call AI-assisted decision-making. On the one hand, AI-assisted decision making has been shown to improve medical assistance [15, 21], reduce costs of public and business services, and enhance security. On the other hand, it may also lead to compromising safety and health of individuals, discrimination, and harming human dignity [5, 17]. Building a collaborative partnership between human deciders and AI-embedded system is therefore a challenge and most critically relies on trust from the users towards the systems [10].

Designing trustworthy AI has been reported by international institutions (European Commission [5], G20 [7]) and governments (USA [3, 16], Estonia [23], or France [24]) have highlighted the need for considering trust in the design of...
AI. In the private sectors, companies such as AXA [6], Accenture [20], or KPMG [12] are also taking this path of research in order to foster trust by going beyond system’s accuracy, promoting privacy, security, algorithm accountability and transparency. To evaluate whether one succeeded in building a trustworthy system, one has to measure study users’ trust in it. Thus, designing and ensuring trustworthy AI has raised interest in HCI. For instance, previous work has looked at what factors influence users’ trust and how [4, 19, 26], how trust is established and developed [2, 18, 25], and how it can be modeled [1, 11]. However, trust remains a highly challenging theoretical concept to study due to its multidisciplinary and multifacet nature [13, 14]. To address this, the literature does not yet provide guidelines that support the empirical study of human trust in AI-based decision support systems.

2 SYSTEMATIC REVIEW AND FINDINGS

In this article, we focused on how to appropriately assess trust between human-users and AI-embedded systems in decision making. Therefore, we investigated questions such as: Which measures to use to study trust? What kind of task to give to users to correctly measure trust? How to include the key elements of trust in an experimental protocol?

To tackle them, we present a comprehensive survey of the experimental methodologies set to investigate trust in AI-assisted decision making. In ACM Digital Library, we searched full papers that had empirical studies of trust with human participants who made final decisions based on the recommendations of AI-based systems. We found 83 papers (mostly published in the past 15 years), and annotated, categorized, and summarized their definitions and methods (both quantitative and qualitative) of trust. Through this literature review, we identified good practices in the current theoretical and experimental approaches, as well as potential caveats, allowing us to draw guidelines and research opportunities in the experimental study of trust in AI-assisted decision making.

Our three main findings are: 1) the three theoretical elements of trust, vulnerability, positive expectations, and attitude are not fully integrated in the reviewed papers’ experimental protocols and qualitative measurements. There is therefore a risk that some empirical studies capture constructs other than trust (confidence, distrust, and reliance); 2) a large variability among the designs and measurements used to assess trust which can impair validity and replicability; and 3) the challenge of investigating the dynamics of trusts considering the constraints of laboratory experiments and the applicability of existing methods.

Based on these findings, we proposed a set of 16 guidelines (G) to help researchers in the design of experimental protocols that would prevent the identified caveats in the study of trust in the specific context of AI-assisted decision making. In complement to guidelines, we identify 9 research opportunities (RO) regarding the elaboration of practical methods to studying trust and its dynamics in laboratory experiments or the investigation of relevant factors (e.g. individual differences, task outcomes) on Human-AI trust (see Table 1 for the full list).

3 USE OF THE TRUST EVALUATION GUIDELINES IN CONTROLLED SETTINGS

While we derived the guidelines, they are only as useful as much they are employed. We illustrate a case example to show how to apply our guidelines (and more generally this review) in practice. Consider designers who have been working on an AI-embedded system for college recruitment following principles of trustworthy AI and would like to evaluate users’ trust in it. First, they are familiarized with what trust is (G1). They can avoid confusing terminology in their literature review search and write-up (G2). Reminded that individual differences such as age, gender, cultural background can contribute to trust variance (G3, G4), designers make sure to explore this in their analysis (RO1). Additionally, we can bring attention of the system’s developers to the fact that college decisions might be made in group, rather than individually, (RO2) and to the fact that university using AI for candidates selection can affect indirect stakeholders -

<table>
<thead>
<tr>
<th>Sections</th>
<th>Guidelines</th>
<th>Research Opportunities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>(G1) Provide a clear definition of trust</td>
<td>(RO1) Investigate individual differences</td>
</tr>
<tr>
<td></td>
<td>(G2) Prevent any confusion between trust and related constructs</td>
<td>(RO2) Investigate how groups of users trust an AI-embedded system</td>
</tr>
<tr>
<td></td>
<td>(G3) Assess the expertise and prior experience of users</td>
<td>(RO3) Investigate how AI-embedded systems are perceived by indirectly impacted stakeholders</td>
</tr>
<tr>
<td></td>
<td>(G4) Consider users’ self-confidence</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(G5) Favour a higher number of participants</td>
<td></td>
</tr>
<tr>
<td>Participants</td>
<td>(G6) Consider alternative interaction flows</td>
<td>(RO4) Investigate the impact of the interaction flows, as factors, on trust</td>
</tr>
<tr>
<td></td>
<td>(G7) Ensure to involve vulnerability</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(G8) Assess participants’ likelihood to exhibit realistic behaviors</td>
<td>(RO5) Investigate the impact of delayed feedback on the dynamics of trust</td>
</tr>
<tr>
<td>Task</td>
<td>(G9) Ensure to control initial participants’ expectations</td>
<td>(RO6) Investigate to what extent virtual outcomes might replace real ones</td>
</tr>
<tr>
<td>Procedure and Design</td>
<td>(G10) Favour interactions over a long period of time</td>
<td>(RO7) Investigate new methodologies to assess dynamic trust in practice</td>
</tr>
<tr>
<td>Quantitative measures</td>
<td>(G11) Favour the use of well-established questionnaires that comprise the key elements of trust</td>
<td>(RO8) Investigate whether single-item questionnaires capture trust as well as other measures</td>
</tr>
<tr>
<td></td>
<td>(G12) Report psychometric statistics</td>
<td>(RO9) Explore more fundamental correlates between physiological sensing and trust</td>
</tr>
<tr>
<td></td>
<td>(G13) Use the term &quot;trust-related behavioral measure&quot; to avoid theoretical confusion</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(G14) FavOUR measures relative to the system’s precision</td>
<td></td>
</tr>
<tr>
<td>Qualitative measures</td>
<td>(G15) Increase empirical rigor when reporting on qualitative methods</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(G16) Adopt under-used qualitative methods for studying trust (Critical Incident Technique, Repertory Grid, Hermeneutics)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Summary of the main guidelines and research opportunities organized according to the constructive elements of an experimental protocol.

While developing an experimental protocol, designers are reminded that their participants have to have something at stake while doing the task (G7, G8), so introducing real or virtual consequences immersive enough is important. Designers also learn about the importance of the first impressions for participants’ trust formation (G9), which can be facilitated with describing how the system was trained, announcing system’s accuracy or making the first recommendations always correct. Because of the dynamic, that is changing, nature of trust, they can understand that their study should allow for an interaction long enough to record multiple stages of trust development (G10). This would also encourage them to explore which trust measures are more suitable for this (RO7, RO8, RO9). Lastly, this paper will help developers select an appropriate trust questionnaire (G11) and remind them what questionnaire-related statistics should be reported (G12), and will familiarize them with other trust-related measures, which do not measure trust directly (G13). If developers decide to conduct qualitative studies with their participants, we provide them with some examples of appropriate tools (Critical Incident Technique, Repertory Grid, Hermeneutics) to run, analyze and report one (G15, G16).

4 USE OF THE TRUST EVALUATION GUIDELINES IN REAL-WORLD SETTINGS

Our guidelines have been predominantly based on the studies done in the laboratory settings, whereas little is known about how trust is evaluated with the real AI-embedded systems for decision making in the field. How often do practitioners evaluate these systems with potential users, and is trust one of the aspects discussed? When talking about "trustworthy", do practitioners consider what users feel while using the system or is "trustworthiness" just
one of the metrics for the systems’ technical performance? If the latter is the case, one should not assume that if a system is considered trustworthy based on technical criteria, users will automatically trust it. Therefore, we believe that practitioners in the field would benefit from these guidelines, too, and are looking forward the discussion on how to adapt them in the real life context.

ACKNOWLEDGMENTS

This work was performed within the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR within the Investissements d’Avenir programme under reference ANR-11-IDEX-004-02.

REFERENCES


