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

Oleksandra Vereschak, Gilles Bailly, Baptiste Caramiaux

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:
Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. On the Way to Improving Experimental Protocols to Evaluate Users' Trust in AI-Assisted Decision Making. In *CHI'21 Workshop: Towards Explainable and Trustworthy Autonomous Physical Systems, May 7, 2021 Online Virtual Conference.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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

*Both authors contributed equally to this research.

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 -

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

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

students (RO3). 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-0004-02.

REFERENCES

[1] Ighoyota Ben. Ajenaghughrure, Sonia C. Sousa, Ilkka Johannes Kosunen, and David Lamas. 2019. Predictive Model to Assess User Trust: A Psycho-Physiological Approach. In *Proceedings of the 10th Indian Conference on Human-Computer Interaction* (Hyderabad, India) *(IndiaHCI ’19)*. Association for Computing Machinery, New York, NY, USA, Article 4, 10 pages. https://doi.org/10.1145/3364183.3364195

[2] Ban Al-Ani, Matthew J. Bietz, Yi Wang, Erik Trainer, Benjamin Koehne, Sabrina Marczak, David Redmiles, and Rafael Prikladnicki. 2013. Globally Distributed System Developers: Their Trust Expectations and Processes. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, Texas, USA) *(CSCW ’13)*. Association for Computing Machinery, New York, NY, USA, 563–574. https://doi.org/10.1145/2441776.2441840

[3] Defense Innovation Board. 2019. *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense.* Technical Report. United States Department of Defense, Virginia, United States. 11 pages. https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB$_$AI$_$PRINCIPLES$_$PRIMARY$_$DOCUMENT.PDF

[4] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. 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 (CHI ’19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300234

[5] European Commission. 2020. *On Artificial Intelligence - A European approach to excellence and trust.* Technical Report. European Commission, Brussels, Belgium. 27 pages. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020$_$en.pdf

[6] AXA Research Fund. 2019. *Artificial Intelligence: Fostering Trust.* Technical Report. AXA. 45 pages. https://www.axa-research.org/en/news/AI-research-guide

[7] G20. 2019. *G20 Ministerial Statement on Trade and Digital Economy.* Technical Report. G20, Brussels, Belgium. 14 pages. http://trade.ec.europa.eu/doclib/press/index.cfm?id=2027

[8] IBM Watson Health. 2020. Artificial Intelligence in medicine. https://www.ibm.com/watson-health/learn/artificial-intelligence-medicine

[9] Rebecca Heilweil. 2019. Artificial intelligence will help determine if you get your next job. https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen

[10] Brett W. Israelsen and Nisar R. Ahmed. 2019. “Dave...I Can Assure You ...That It’s Going to Be All Right ...” A Definition, Case for, and Survey of Algorithmic Assurances in Human-Autonomy Trust Relationships. *ACM Comput. Surv.* 51, 6, Article 113 (Jan. 2019), 37 pages. https://doi.org/10.1145/3267338

[11] Bran Knowles, Mark Rouncefield, Mike Harding, Nigel Davies, Lynne Blair, James Hannon, John Walden, and Ding Wang. 2015. Models and Patterns of Trust. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW ’15)*. Association for Computing Machinery, New York, NY, USA, 328–338. https://doi.org/10.1145/2675133.2675154

[12] KPMG. 2019. *Controlling AI: The imperative for transparency and explainability.* Technical Report. KPMG. 28 pages. https://advisory.kpmg.us/articles/2019/controlling-ai.html

[13] Roy Lewicki and Chad Brinsfield. 2011. Measuring trust beliefs and behaviours. In *Handbook of Research Methods on Trust*, Fergus Lyon, Guido Möllering, and Mark Saunders (Eds.). Edward Elgar, Cheltenham, UK; Northampton, MA, USA, Chapter 3, 29–39. https://doi.org/10.4337/9781781009246.00013

[14] Fergus Lyon, Guido Möllering, and Mark Saunders. 2015. *Handbook of Research Methods on Trust: Second Edition.* Edward Elgar Publishing, Cheltenham, United Kingdom. 1–343 pages. https://doi.org/10.4337/9781782547419

[15] Rob Matheson. 2019. Automating artificial intelligence for medical decision-making. http://news.mit.edu/2019/automating-ai-medical-decisions-0806

[16] White House Office of Science and Technology Policy. 2020. *American AI Initiative: Year One Annual Report.* Technical Report. White House Office of Science and Technology Policy, Brussels, Belgium. 36 pages. https://www.whitehouse.gov/ai/

[17] Cathy O’Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing Group, USA.

[18] Samir Passi and Steven J. Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 136 (Nov. 2018), 28 pages. https://doi.org/10.1145/3274405

[19] Lionel P. Robert. 2016. Monitoring and Trust in Virtual Teams. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 245–259. https://doi.org/10.1145/2818048.2820076

[20] Accenture Federal Services. 2019. *Responsible AI: A Framework for Building Trust in your AI Solutions*. Technical Report. Accenture. 13 pages. https://www.accenture.com/us-en/insights/us-federal-government/ai-is-ready-are-we

[21] Cassie Solomon, Mark Schneider, and Gregory P. Shea. 2018. How AI-based Systems Can Improve Medical Outcomes. https://knowledge.wharton.upenn.edu/article/ai-based-systems-can-improve-medical-outcomes/

[22] Jason Tashea. 2017. Courts Are Using AI to Sentence Criminals. https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/

[23] AI Taskforce. 2019. *Report of Estonia's AI Taskforce.* Technical Report. Republic of Estonia Government Office and Republic of Estonia Ministry of Economic Affairs and Communications, Estonia. 47 pages. https://ec.europa.eu/knowledge4policy/ai-watch/estonia-ai-strategy-report

[24] Cédric Villani, Yann Bonnet, Bertrand Rondepierre, et al. 2018. *For a meaningful artificial intelligence: Towards a French and European strategy.* Conseil national du numérique, France.

[25] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User Trust Dynamics: An Investigation Driven by Differences in System Performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI '17)*. Association for Computing Machinery, New York, NY, USA, 307–317. https://doi.org/10.1145/3025171.3025219

[26] 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