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

Oleksandra Vereschak, Gilles Bailly, Baptiste Caramiaux

On the Way to Improving Experimental Protocols to Evaluate Users' Trust in AI-Assisted Clinical 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 medical decision making makes it critical to build these systems according to trustworthiness standards. 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 ameliorate the methodology of studying Human-AI trust in medical decision-making contexts. In this workshop, we would like to discuss how these insights could improve the quality of data about users' trust and, thus, lead to new steps towards closing the "last mile" between AI and healthcare workers.

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 Clinical Decision Making. In *CHI'21 Workshop: Realizing AI in Healthcare: Challenges Appearing in the Wild, May 8–9, 2021 Online Virtual Conference*. ACM, New York, NY, USA, 8 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 medical ones [34]. In such situations, medical personnel 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 **clinical AI-assisted decision making**. On the one hand, AI-assisted decision making has been shown to improve medical assistance [35, 50] and reduce costs of services. On the other hand, it may also lead to compromising safety and health of individuals, discrimination, and harming human dignity [10, 43]. Building a collaborative partnership between medical deciders and AI-embedded systems is therefore a challenge and most critically relies on **trust** from the users towards the systems [20].

Designing trustworthy AI has been reported by international institutions (European Commission [10], G20 [15]) and governments (USA [5, 41], Estonia [51], or France [52]) have highlighted the need for considering trust in the

*Both authors contributed equally to this research.

Manuscript submitted to ACM

design of AI. In the private sectors, companies such as AXA [14], Accenture [48], or KPMG [24] 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 [7, 46, 58], how trust is established and developed [2, 44, 57], and how it can be modeled [1, 23]. However, little focus is directed towards medical scenarios. In addition, trust remains a highly challenging theoretical concept to study due to its multidisciplinary and multifacet nature [28, 33]. 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

In this workshop, we focus on how to appropriately assess trust between human users and AI-embedded systems in decision making. We believe that this would allow for better quality of data about users' trust, avoiding overclaiming. This, in turn, would enhance understanding of healthcare workers' trust in AI. Therefore, we investigate 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 conducted 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, 5 of which are directly linked to clinical decision making. We 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. It is a submission to CSCW, currently under revision, and in this workshop, we present a synthesis of some of the results. More specifically, during the workshop, we explain and highlight why trust definition plays an important role for design of experimental protocols. Additionally, we share our research opportunities exclusively related to clinical AI-assisted decision making scenarios. These contributions are an opportunity to discuss during the workshop how they could be applied for studying trust in AI in the healthcare context and what other research opportunities can be emerged from them.

3 TRUST DEFINITION AND IMPLICATIONS

In almost half of the reviewed studies that provided a trust definition, trust is defined in the following manner: *"An attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability"* [26] (n = 10, 45.5% of the 22 papers with definitions). While comparing this definition with the remaining 10 trust definitions encountered in the reviewed papers, we identified three key elements of trust common among all the definitions: (1) **vulnerability**, (2) **positive expectations**, and (3) **attitude**. All of them have implications for an experimental protocol for studying trust.

3.1 Vulnerability and Task Outcomes

Let's imagine a situation where a patient has a serious illness, and their doctor proposes a treatment. The patient is in a situation of *vulnerability*, the first key element of trust, as this situation involves uncertainty of the outcomes of

a decision, with potential negative or undesirable consequences [17, 32]. For instance, following a treatment might just not work or might provoke severe side effects. Uncertainty might be due to the unpredictable nature of the world as well as the lack of human knowledge and capabilities [8]. However, it is necessary to distinguish two natures of uncertainty (sometimes referred as risk vs. ambiguity [22]): the possibility of outcomes can sometimes be estimated (e.g. the treatment has 30% of success with full recovery) or not (e.g. the percentage of success or the side effects of the treatment are not known). Here, the notion of vulnerability relates to both types of uncertainty. Without vulnerability, there is no need for trust to emerge [8, 16, 25, 42].

Therefore, if an experimental task is not immersive enough and the consequences of participants' decisions are not realistic, the data obtained is likely to be about **confidence** rather than trust. To avoid this problem, the task outcomes should be controlled either with real (e.g., money bonus or malus based on the decision quality) or virtual (e.g., virtual points bonus based on the decision quality) consequences.

3.2 Positive Expectations and Initial Interaction

Continuing with the previous example, trust would not emerge either if the doctor does not have *positive expectations* about the system. Even if the doctor decides to follow the AI's recommendation, we cannot claim that the doctors trust it [17, 32]. Trust has grounds to form only when one thinks as if negative outcomes associated with trusting do not exist or are very unlikely [30]. Without positive expectations, it is more appropriate to discuss about **distrust**. This construct is often confounded with low levels of trust [37]. While there are some researchers that deem trust and distrust as the opposite ends of one construct [47], recently the community views them as two separate ones [29, 49]. This means that they can both reach high and low levels and exist simultaneously.

It is thus important to help participants establish positive expectations about the system in the beginning of the experiment; otherwise, the data collected might be related more to distrust than trust. It can be done through mentioning that the AI system was trained for this task, stating its accuracy or ensuring that during an experiment the first few recommendations are error-free.

3.3 Trust as Attitude

Saying that trust is an *attitude* implies that trust does not systematically translate in a behavior. For example, the doctors's level of trust might be sufficient enough to follow the AI's recommendation, but they decide not to do so, because none of their colleagues are using this system. A socio-cognitive approach to defining trust suggests that trust is rather an attitude [8], i.e a certain way of feeling about the object [6]. Trust then cannot be always fully observable to the third parties (unless clearly and objectively communicated in a verbal or written form), which has an important impact on the choice of the methods to study trust.

For instance, it means that observational studies are not enough to draw conclusions about healthcare workers' levels of trust, and should be paired up with other methods. Supplementary qualitative methods to evaluate trust could be *retrospective* (about the past experiences) and *non-retrospective* (during the interaction). Retrospective methods include interviews, and while there are many types of procedures, **critical incident technique** [12] is especially useful for capturing changes (both positive and negative) in levels of trust. It is a set of procedures used to collect data from narrated past experiences (or observations) to identify and brainstorm about important events related to a pre-defined problem [3]. When applied to trust, it is especially useful to study real life cases in which trust was established, destroyed or repaired [40]. Researchers directly ask participants what aspects of others' behavior was important for trust weakening or strengthening. Just like this information can be applied, for example, towards improving patients' experience during

a medical visit [53, 54] or enhancing intercultural business negotiations [40], it can also be used for understanding breakdown moments for healthcare workers with AI.

Non-retrospective methods, underused in the corpus, include **think-aloud protocols**, which can generate authentic and spontaneous reactions of the participants as these ones are not given any prompts to speak up. Moreover, this method avoids memory distortion effects, which sometimes happened with methods used post experiment.

Lastly, trust can also be measured using quantitative measures, for example, through **questionnaires**. Among 32 papers with multi-question questionnaires in our corpus, we identified 21 different questionnaires used to measure participants' trust in an AI-embedded system. As there is an abundance of choice, it is challenging to understand which questionnaire to choose. We identify at least two questionnaires that equally focus on vulnerability *and* positive expectations [39] - by Mayer [36] ($n = 3$) and by McKnight [38]. Otherwise, there is a risk of obtaining the data about confidence or distrust.

Questionnaire data can be completed with the additional one calculated from behavioral logs. This additional data should be called *trust-related behavioral measures* [37], instead of "behavioral trust measures", as it is usually referred to assuming that trust can be directly observed. These measures are reliance (how many times a doctor decided to use the system), compliance (how many times a doctor decided to follow the AI's recommendation), and switch ration (how many times a doctor changed his opinion after seeing the AI's recommendation). The later one can be mostly calculated in the experimental laboratory settings, where participants are asked to explicitly state their opinion.

Therefore, trust definition has a direct link with the design of experimental protocols. If the key elements of trust are not incorporated in a study, the data yielded might be linked to other theoretical concepts related to trust (see Figure 1).

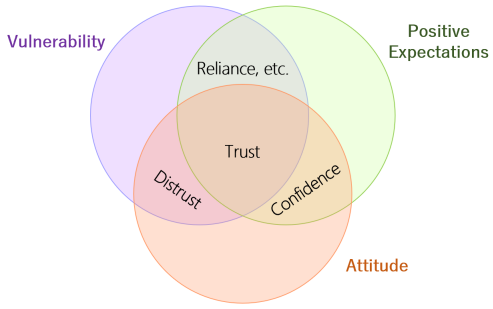

Fig. 1. A simplified representation of some constructs related to trust and how they are connected with the key elements of trust.

4 RESEARCH OPPORTUNITIES FOR CLINICAL AI-ASSISTED DECISION MAKING

Here we would like to share our observations on the unexplored areas in the empirical studies on trust in the context of clinical AI-assisted decision making.

4.1 Group Decisions

The predominant trend in Human-AI trust with decision-making is to investigate trust of an individual. However, a line of literature in social sciences suggests the importance of considering *trust of a group* (e.g., [11, 13, 18]). Indeed, group decisions with an AI-embedded system are part of real-life cases for the medical field (e.g., [55]). Moreover,

group decision-making and trust processes have been shown to be different from the individual ones [21]. For example, repairing trust has been found to be more difficult for groups than for individuals [21].

4.2 Subjective Expertise

Another common question to participants if asking for years of their experience with the task, but it might be not enough. Another interesting aspect to investigate is their *subjective expertise* (also called self-confidence or self-efficacy [45]). Subjective expertise is how well participants think they can achieve their goal (e.g., solving a problem). Research suggests that people are generally overconfident in their abilities, which leads to biased judgement [31, 56] and in turn might affect trust-related perceptions and decisions [27]. It is believed that its magnitude depends on the gender [4], the age or the culture [19].

4.3 Indirect Stakeholders

In the most reviewed studies, the participant has the role of the user *directly* interacting with the system. However, there are other stakeholders who do not interact with the system directly, yet can be impacted by the decisions made with AI-embedded systems, and it could be insightful investigate their trust, too. For example, would patients still trust and listen to the doctor if they had known beforehand the doctor is assisted by an AI for diagnosis assessment [9]?

4.4 Dynamic Trust

Trust can be increased, decreased, repaired, and maintained [28], that is it changes over time and can be *dynamic*. However, most of the studies we have reviewed allowed for one-time under-20-minute interactions. As this amount of time might be not enough for capturing all the stages of trust development, it would insightful to incorporate more of longitudinal studies to be able to investigate these changes. Would a heath worker be as skeptical or enthusiastic about the AI-embedded system, for example, after 1 month?

5 CONCLUSION

In this workshop, we summarized some of the results of the systematic review submitted to CSCW. We would like to open the discussion about how to improve experimental protocols evaluating users' trust in AI-assissted clinical decision making. We share our findings about three key elements of trust - vulnerability, positive expectations, and attitude - and how they affect the design of experimental protocols for investigating trust. Specifically, task outcomes, the way the system is introduced, and the choice of measures are affected by trust definition. We would also like to highlight unexplored research opportunities in clinical AI-assisted decision making. We believe these two points will lead to better quality of data on healthcare workers' trust, avoiding overclaiming and improving across-study results comparability. These could be the new steps towards closing the "last mile" between AI and healthcare workers.

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 (IndiaHCI '19)*. Association for Computing

Machinery, New York, NY, USA, 10. 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] Bengt-Erik Andersson and Stig-Göran Nilsson. 1964. Studies in the reliability and validity of the critical incident technique. *Journal of Applied Psychology* 48 (1964), 398–403. https://doi.org/10.1037/h0042025

[4] Brad M. Barber and Terrance Odean. 2001. Boys Will be Boys: Gender, Overconfidence, and Common Stock Investment. *The Quarterly Journal of Economics* 116, 1 (2001), 261–292. http://www.jstor.org/stable/2696449

[5] 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

[6] Gerd Bohner and Nina Dickel. 2011. Attitudes and Attitude Change. *Annual Review of Psychology* 62, 1 (2011), 391–417. https://doi.org/10.1146/annurev.psych.121208.131609 PMID: 20809791.

[7] 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

[8] Cristiano Castelfranchi and Rino Falcone. 2010. *Socio-Cognitive Model of Trust: Basic Ingredients.* John Wiley & Sons, Ltd, Chichester, United Kingdom, Chapter 2, 35–94. https://doi.org/10.1002/9780470519851.ch2

[9] I. Glenn Cohen. 2020. Informed Consent and Medical Artificial Intelligence: What to Tell the Patient? *Georgetown Law Journal* 108 (2020), 1425–1469. https://doi.org/10.2139/ssrn.3529576

[10] 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

[11] Graham Dietz and Deanne N. Den Hartog. 2006. Measuring trust inside organisations. *Personnel Review* 35 (2006), 557–588. https://doi.org/10.1108/00483480610682299

[12] J. C. Flanagan. 1954. The critical incident technique. *The Psychological Bulletin* 51, 4 (1954), 327–358.

[13] C. Ashley Fulmer and Michele J. Gelfand. 2012. At What Level (and in Whom) We Trust: Trust Across Multiple Organizational Levels. *Journal of Management* 38, 4 (2012), 1167–1230. https://doi.org/10.1177/0149206312439327 arXiv:https://doi.org/10.1177/0149206312439327

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

[15] 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

[16] Diego Gambetta. [n.d.]. *Can We Trust Trust?* Department of Sociology,University of Oxford, Oxford, United Kingdom.

[17] Larue Tone Hosmer. 1995. Trust: The Connecting Link between Organizational Theory and Philosophical Ethics. *The Academy of Management Review* 20, 2 (1995), 379–403. http://www.jstor.org/stable/258851

[18] Lenard Huff and Lane Kelley. 2003. Levels of Organizational Trust in Individualist versus Collectivist Societies: A Seven-Nation Study. *Organization Science* 14, 1 (2003), 81–90. http://www.jstor.org/stable/3086035

[19] J. S. Hyde. 2005. The gender similarities hypothesis. *Am Psychol* 60, 6 (Sep 2005), 581–592.

[20] 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

[21] Peter H. Kim, Cecily D. Cooper, Kurt T. Dirks, and Donald L. Ferrin. 2013. Repairing trust with individuals vs. groups. *Organizational Behavior and Human Decision Processes* 120, 1 (2013), 1–14. https://doi.org/10.1016/j.obhdp.2012.08.0

[22] F. H. Knight. 1921. *Risk, Uncertainty, and Profit.* Houghton Mifflin, New York, USA. https://fraser.stlouisfed.org/files/docs/publications/books/risk/riskuncertaintyprofit.pdf

[23] 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

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

[25] Alexander Lascaux. 2008. Trust and uncertainty: a critical re-assessment. *International Review of Sociology* 18 (03 2008), 1–18. https://doi.org/10.1080/03906700701823613

[26] John Lee and Katrina See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human factors* 46 (February 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

[27] John D. Lee and Neville Moray. 1994. Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies* 40, 1 (1994), 153 – 184. https://doi.org/10.1006/ijhc.1994.1007

[28] 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

[29] Roy J. Lewicki, Daniel J. McAllister, and Robert J. Bies. 1998. Trust and Distrust: New Relationships and Realities. *The Academy of Management Review* 23, 3 (1998), 438–458. http://www.jstor.org/stable/259288

[30] J. David Lewis and Andrew Weigert. 1985. Trust as a Social Reality. *Social Forces* 63, 4 (1985), 967–985. http://www.jstor.org/stable/2578601

[31] Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D. Phillips. 1977. Calibration of Probabilities: The State of the Art. In *Decision Making and Change in Human Affairs*. Springer Netherlands, Netherlands, 275–324. https://doi.org/10.1007/978-94-010-1276-8_19

[32] Niklas Luhmann. 1979. *Trust and Power* (1 ed.). Wiley, Chichester, Toronto.

[33] 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

[34] Tamra Lysaght, Hannah Yeefen Lim, Vicki Xafis, and Kee Yuan Ngiam. 2019. AI-Assisted Decision-making in Healthcare. *Asian Bioethics Review* 11, 3 (01 Sep 2019), 299–314. https://doi.org/10.1007/s41649-019-00096-0

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

[36] James H. Mayer, Roger C.;Davis. 1999. The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Business and Industrial Personnel* 84, 1 (1999), 123–136. https://doi.org/10.1037/0021-9010.84.1.123

[37] D. Mcknight and Norman Chervany. 2001. Trust and Distrust Definitions: One Bite at a Time. In *Trust in Cyber-societies: Integrating the Human and Artificial Perspectives*, R. Falcone, M. Singh, and Y. H. Tan (Eds.). Springer, Heidelberg, Germany, 27–54. https://doi.org/10.1007/3-540-45547-7_3

[38] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research* 13, 3 (2002), 334–359. https://doi.org/10.1287/isre.13.3.334.81

[39] D. Harrison McKnight, Larry L. Cummings, and Norman L. Chervany. 1998. Initial Trust Formation in New Organizational Relationships. *Academy of Management Review* 23, 3 (1998), 473–490. https://doi.org/10.5465/amr.1998.926622

[40] Robert Münscher and Torsten M. Kühlmann. 2011. Using critical incident technique in trust research. In *Handbook of Research Methods on Trust*, Fergus Lyon, Guido Möllering, and Mark Saunders (Eds.). Edward Elgar, Cheltenham, UK; Northampton, MA, USA, Chapter 14, 161–172.

[41] 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/

[42] Claus Offe. [n.d.]. *How can we trust our fellow citizens?* Cambridge UP, Cambridge, United Kingdom.

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

[44] 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

[45] Patricia Perry. 2011. Concept Analysis: Confidence/Self-confidence. *Nursing Forum* 46, 4 (2011), 218–230. https://doi.org/10.1111/j.1744-6198.2011.00230.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1744-6198.2011.00230.x

[46] 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

[47] Julian B. Rotter. 1980. Interpersonal trust, trustworthiness, and gullibility. *American Psychologist* 35, 1 (1980), 1–7. https://doi.org/10.1037/0003-066X.35.1.1

[48] 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

[49] Sim B. Sitkin and Nancy L. Roth. 1993. Explaining the Limited Effectiveness of Legalistic "Remedies" for Trust/ Distrust. *Organization Science* 4, 3 (1993), 367–392. http://www.jstor.org/stable/2634950

[50] 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/

[51] 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

[52] 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.

[53] Eva K. Wendt, Bengt Fridlund, and Evy Lidell. 2004. Trust and confirmation in a gynecologic examination situation: a critical incident technique analysis. *Acta obstetricia et gynecologica Scandinavica* 83 12 (2004), 1208–1215.

[54] Rodrigo Ya Apmez-Gallardo and Sandra Valenzuela-Suazo. 2012. Critical incidents of trust erosion in leadership of head nurses. *Revista Latino-Americana de Enfermagem* 20 (02 2012), 143 – 150. http://www.scielo.br/scielo.php?script=sci$_$arttext&pid=S0104-11692012000100019&nrm=iso

[55] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F. Antaki. 2016. Investigating the Heart Pump Implant Decision Process: Opportunities for Decision Support Tools to Help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4477–4488. https://doi.org/10.1145/2858036.2858373

[56] J. Frank Yates. 1990. *Judgment and decision making.* Prentice-Hall, Inc, Englewood Cliffs, NJ, US. xvi, 430–xvi, 430 pages.

[57] 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

[58] 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 (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852