How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies

Oleksandra Vereschak, Gilles Bailly, Baptiste Caramiaux

How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies

OLEKSANDRA VERESCHAK, Sorbonne Université, 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 studying human trust in these systems critical. However, empirically investigating trust is challenging. One reason is the lack of standard protocols to design trust experiments. In this paper, we present a survey of existing methods to empirically investigate trust in AI-assisted decision making and analyse the corpus along the constitutive elements of an experimental protocol. We find that the definition of trust is not commonly integrated in experimental protocols, which can lead to findings that are overclaimed or are hard to interpret and compare across studies. Drawing from empirical practices in social and cognitive studies on human-human trust, we provide practical guidelines to improve the methodology of studying Human-AI trust in decision-making contexts. In addition, we bring forward research opportunities of two types: one focusing on further investigation regarding trust methodologies and the other on factors that impact Human-AI trust.

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. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 327 (October 2021), 39 pages. https://doi.org/10.1145/3476068

1 INTRODUCTION

Artificial Intelligence (AI) has acquired a critical role in assisting humans in making sensitive decisions with uncertain outcomes such as hiring [83], treatment assignment [82], or criminal investigation processes [212], 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 [135, 205], 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 [36, 164]. Building a collaborative partnership between human deciders and AI-embedded systems is therefore a challenge and most critically relies on **trust** from the users towards the systems [94].

Designing trustworthy AI has been reported by international institutions (European Commission [36], G20 [65]) and governments (USA [15, 159], Estonia [213], or France [222]) have highlighted the need for considering trust in the design of AI. In private sectors, companies such as AXA [64], Accenture [199] or KPMG [108] are also taking this path of research in order to foster trust

*Both authors contributed equally to this research.

Authors' addresses: Oleksandra Vereschak, vereschak@isir.upmc.fr, Sorbonne Université, ISIR, Paris, France; Gilles Bailly, Sorbonne Université, CNRS, ISIR, Paris, France, gilles.bailly@sorbonne-universite.fr; Baptiste Caramiaux, Sorbonne Université, CNRS, ISIR, Paris, France, baptiste.caramiaux@sorbonne-universite.fr.

by going beyond system's accuracy, promoting privacy, security, algorithm accountability and transparency. 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 [26, 183, 248], how trust is established and developed [3, 167, 245], and how it can be modeled [2, 105]. However, trust remains a highly challenging theoretical concept to study due to its multidisciplinary and multifaceted nature [119, 129]. To address this, the literature does not yet provide **guidelines** that support the empirical study of human trust in AI-based decision support systems.

In this article, we focus on how to assess that trust exists between human-users and AI-embedded systems in decision making. Therefore, we are interested in questions such as: Which measures can be used to measure trust? What kind of task should be given 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. Through this literature review, we aim to identify 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; 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 trust considering the constraints of laboratory experiments and the applicability of existing methods.

Based on these findings, we propose a set of 16 guidelines 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 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.

Our main contributions are:

(1) An exhaustive presentation of the variety and complexity of the methods to study Human-AI trust in the decision making context;
(2) A structured discussion of the current Human-AI trust protocols highlighting flaws in methodologies with a stronger link to the Human-Human Trust community;
(3) A set of guidelines and research opportunities to improve research quality in Human-AI Trust, highlighting the need for a greater empirical rigor in the community.

2 RESEARCH IN HUMAN-AI INTERACTION

The notion of trust in the fields of CSCW and HCI is transverse to several research lines of inquiry. In this section, we describe how the systematic review presented further on in this article, contributes to three lines of inquiry in the Human-AI literature: empirical research on trust in AI, Human-AI guidelines, and multidisciplinary constructs in AI (including explainable and interpretable AI).

2.1 Empirical Research on Trust in AI

Many empirical studies (e.g., [10, 80, 86, 116, 241]) investigate the impact of factors related to user, system and task on trust while interacting with an AI. For instance, [241] explores the impact of stated and actual system accuracy on users' trust, and [56] studies how experts and novices of

the given task react to Machine Learning recommendations. Recently, Glikson and Woolley [72] reviewed, synthesised and discussed these empirical findings. While their focus was on factors that affect users' trust, they also remarked the need to address the great variance of measures used to study trust in AI. The authors urged to refer to other disciplines in an effort to improve the current research methodology on trust in AI with human subjects. Our work does so by drawing from social and cognitive sciences, and henceforth, opens a cross-disciplinary dialogue about the practices suitable for studying trust. Our main focus is, thus, investigating *how to study* trust rather than *factors that affect* trust. Readers can refer to [72] for a general overview of the latter, to [32] for a review of advances in visualization techniques related to trust, and to [31] for a discussion of the role of explainable AI in the context of trust.

2.2 Human-AI Guidelines

An increasing number of Human-AI guidelines provide both high and low level suggestions on how to build systems that users can trust. They focus on different aspects such as transparency [146, 158, 162, 218], understandability [204, 218] or explainability [157]. Trust plays an important role in these guidelines. For instance, a recent review [98] states that at least 30% of the ethical guidelines for AI name *trust* as one of the main ethical principles. Amershi et al. [6] present guidelines to help in designing and evaluating AI-embedded systems that users can *trust* and work with efficiently. A framework for building trust in AI proposed by Accenture [199] names Human Centered Design as one of the main tools to instill trust in users.

These Human-AI guidelines are often built on practitioners' experience and existing empirical literature. While trust is often mentioned, it is challenging to assess how exactly and which of the recommendations might contribute to users' trust development. The future Human-AI Guidelines can benefit from our review through further understanding of trust and through being able to assess rigorousness of the empirical studies their guidelines are based on.

2.3 Multidisciplinary Constructs in AI

Working on Human-AI Interaction (HAI) requires to manipulate several multidisciplinary and complex theoretical constructs such as fairness [151], explainability [224], interpretability [70] or trust. Loose use of definitions and conflicting terminology, inherent to multidisciplinary terms, cause misunderstandings in the community. Consequently, multiple projects have been developed with the aim of disentangling these constructs in the HAI community by providing more theoretical foundations. For instance, Mulligan et al. [151] examine fairness from the perspectives of various fields from law to computer science and create a heuristic tool for more structured interdisciplinary discussions and research collaborations around fairness. Wang et al. [224] examine explainability as another construct which usually lacks thorough comprehension. Leveraging research on human reasoning and biases, they identify gaps in the existing Explainable AI techniques and propose and validate new ways to facilitate decision-making with AI explanations. Gilpin et al. [70] discuss the theoretical differences between interpretability and explanability when interacting with an AI.

Our work contributes to the line of research of multidisciplinary constructs in Human-AI Interaction in two ways. First, we examine trust and its main theoretical components, and this construct has not been the major focus of this line of research yet. Second, we go beyond theoretical notions of trust and explore how current empirical practices of studying trust in Human-AI Interaction can be improved though drawing from other sciences.

3 SYSTEMATIC REVIEW METHODOLOGY

We propose a systematic review of previous research in Human-AI trust literature in order to understand how human trust has been empirically investigated in AI-assisted decision-making. In

this section, we describe the method used to perform the systematic review. It encompasses three phases: keywords identification, and 2 steps of papers selection (see Figure 1).

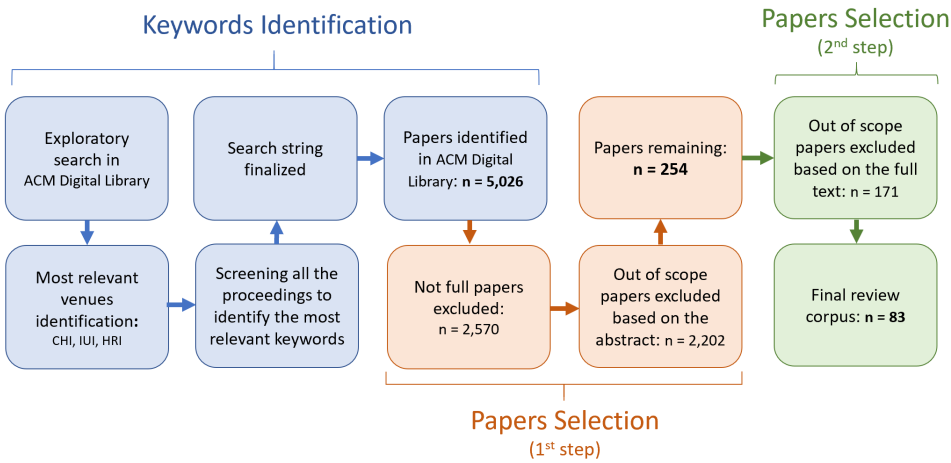

Fig. 1. Papers search and selection process.

3.1 Keywords Identification

We first conducted an explanatory search to identify the search keywords. In the ACM Digital Library, we searched for papers likely to include an empirical study about human trust in an AI-embedded system, where participants have to make a decision. For that, we required the abstract to include either *artificial intelligence* or *machine learning* together with either *trust* or *decision*. The full text should have included either *trust* or *reliance* with *participant* to filter out purely theoretical, technological or modelling papers.

This exploratory search produced 386 results[1]. Out of them, we chose 48 relevant papers (using the same selection procedure as for the final selection of the papers, see *Paper Selection*) to identify the most reoccurring and relevant venues, which are CHI, IUI, and HRI. To find new keywords to be included in our final research string, we manually reviewed every publication in all of their proceedings from 2005 to 2020[2] (8108 papers in total) to find additional relevant papers.

This step resulted in 17 more relevant papers, and from their abstracts we identified additional keywords used to describe the systems (e.g., algorithm, agent). We had iterated different combinations of the keywords up until the moment all papers, deemed relevant in the exploratory step, appeared among the search results of ACM Digital Library. The final search string is the following:

```
(("Abstract": "artificial intelligence" OR " ML " OR " AI " OR "machine learning"
OR "systems" OR agent OR algorithm* OR automat*) AND ("Abstract": trust OR
decision* OR user*)) AND ("Full Text Only": "trust" AND "participants")
```

[1]This phase was conducted in April 2020. We set no time restriction, the earliest papers found dated 2005. Such year range coincides with the recent rise of interest in Human-AI research [76]
[2]This phase was conducted in April 2020

3.2 Selection Criteria

The refined search led to 5026[3] papers, and we manually selected the relevant ones for our scope based on five criteria:

(1) **Trust.** A paper to be selected should have results discussing Human-AI trust. If there are no results on trust reported in the paper or if there are results on Human-Human trust instead of Human-AI trust, the paper is not included.
(2) **Experiments with human participants.** We excluded all the papers that did not have an experiment (e.g., theoretical, guidelines). We also excluded the papers that ran experiments without human participants, for instance, experiments using simulated cognitive models.
(3) **AI technology.** We considered the papers involving AI technologies[4]. As AI is a broad term, this criterion was an important selection challenge. To address it, we followed the methodology presented in [72] and included the following systems if a paper did not explicitly mention the system is AI-embedded: robots, virtual agents, and automated vehicles.
(4) **Human decision making.** There is a full spectrum of ways in which humans and machines can collaborate to make a final decision with uncertain outcomes, from AI-assisted human decision making (human-centered) to AI system assisted by a human (machine-centered). A paper would be deemed relevant if it is a participant who makes the final decision(s) based on the system's output(s). For example, a hiring system could suggest to accept candidate A, but it is up to a user to take the final decision.
(5) **Format.** We included only full papers, so that all the reviewed papers could contain similar level of details about a study. Therefore, posters, late-breaking works, workshops etc., were excluded. We also excluded papers in a language we could not read.

3.3 Papers Selection

The paper selection consisted of two steps, both manual. In the first one, we focused on papers' formats and their abstracts. We excluded 2570 papers due to their format and 2202 papers due to the main goal of a study and type of system, which left us with 254 papers.

In the second selection step, the principal investigator read the full texts of these 254 papers. All the papers were read twice with a time gap of 2 weeks in a randomly reshuffled order without seeing the previous annotations during these weeks to ensure selection stability. 171 papers were deemed irrelevant because they did not have studies with human participants (n=73) or decision making (n = 66), they used irrelevant systems (n=13), did not focus on trust (n=15), or instead on Human-Human trust (n=4).

3.4 Corpus Overview

The final corpus consisted of **83 papers**. We did a first analysis on the publication venues and the year of publication, depicted in Figure 2. It shows that 79 were published in the conference proceedings and 4 in journals. Papers published at *CHI* (n=16), *IUI*, and *HRI* (n=11 each) account for 45.7%[5] of the selected corpus (Figure 2a). The other venues were centered around socio-technical systems (e.g., CSCW, FAccT), interfaces (e.g., AutomotiveUI, DPPI), and autonomous and intelligent agents (e.g., AAMAS, HAI). 66% of the corpus have been published in the past 5 years (see Figure 2b). Such a trend reflects the increasing number of publications in the venues as well as establishments of new venues (e.g., FAccT, HAI).

[3]All the numbers reported are as of beginning of January 2021.
[4]We found several keywords including automated decision aid (e.g., [91]), AI-based decision support system (e.g., [24]), intelligent assistant (e.g., [2]), intelligent agent (e.g., [214]), classifier (e.g., [237]), etc.
[5]Henceforth, all reported percentages are rounded up to the nearest tenth.

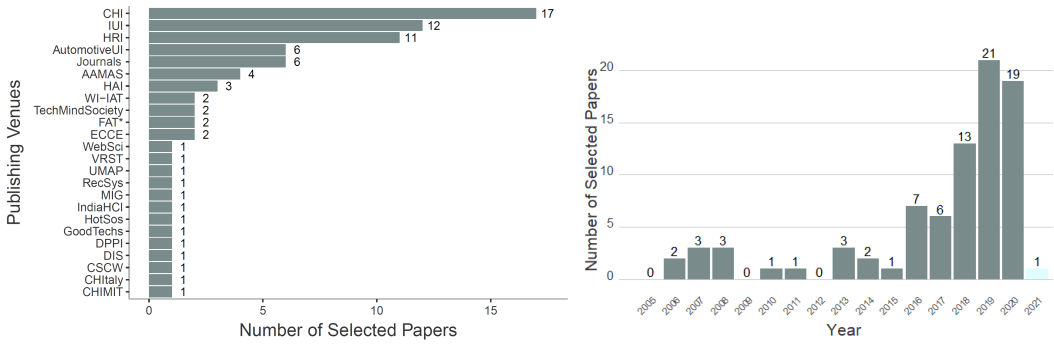

Fig. 2. a) Number of the selected papers per publishing venue. b) Number of the selected papers per year from 2005 to 2021. Please note that the data for 2021 is incomplete since the data collection for this survey was conducted in the beginning of January 2021.

3.5 Corpus Analysis

3.5.1 Papers Annotation. To be able to analyze and discuss the information present in every empirical study in a structured and systematic way, we elaborated a grid of analysis of the corpus of papers. First, we extracted the **definitions of trust** used and their origins. Then, we extracted the information corresponding to each element of an experimental protocol:

- **Participants**: experience, expertise, and number;
- **Task**: the process of decision-making, task feedback and outcomes;
- **Procedure and experimental design** focusing on instructions and the order of questionnaires;
- **Data collection methods and analysis**: types of measures used, how they are implemented and analyzed.

3.5.2 Papers Analysis. Once we annotated each paper based on the grid above, we identified similarities in the methods used in each section and grouped them. This resulted in a categorized and complete overview of methods used for studying trust in AI-embedded systems assisting human decisions. This allowed us to identify the common practices in the community and compare them to the ones of Human-Human trust research. To do so, principal investigator, with a background in social and cognitive sciences, relied on handbooks and reviews (e.g., [30, 69, 129]) to identify Human-Human trust community discussions on methodology, raised issues, and proposed solutions pertinent to each element of an experimental protocol. If the common practices between Human-AI trust and Human-Human trust differed or if no common trend was spotted for the former, we explained the limitations of the current approaches stemming from our corpus. Additionally, drawing from social and cognitive science literature on Human-Human trust, we provided guidelines (**G**) aiming at overcoming these methodological limitations. We also proposed research opportunities (**RO**) for investigating further trust factors and trust methodology in the context of AI-assisted decision making.

3.6 Review Structure and Summary

Each section of this review is dedicated to each element of an experimental protocol we used as an annotation grid for our corpus. We start each subsection with a categorized summary of

the methods used in the corpus, which we discuss in the light of social and cognitive sciences literature on Human-Human trust, highlighting strengths and limitations. Based on this, we provide guidelines and research opportunities stemming from our discussion. Table 1 reports a summary of the guidelines and research opportunities per section.

3.7 A Practical Example

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, thanks to Sections 4.1 and to Sections 4.2 of this review, they are familiarized with what trust is (G1). Using section 4.3, 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, Section 5.2 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 - 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), and Section 6.3.1 can provide multiple examples of how to do it. Designers also learn about the importance of the first impressions for participants' trust formation (G9), and can find several examples of how to introduce the system in Section 7.1. From Section 7.3, 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, Section 8.1 will help developers select an appropriate trust questionnaire (G11) and remind them what questionnaire-related statistics should be reported (G12). Section 8.2 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 Sections 9.2 and 9.3 will provide them with some examples of appropriate tools to run, analyze and report one (G15, G16).

In the following sections, we present the analysis of the reviewed papers and detail the guidelines and research opportunities mentioned above.

4 TRUST DEFINITIONS

In this section, we review existing definitions used in the Human-AI literature, highlighting the differences with the ones used in Human-Human trust literature. We identify the components of trust. We then suggest what should be considered while defining trust in a paper as it influences the choice of the experimental set-up and empirical methods to study it.

4.1 Definitions in Human-AI Trust

Only 26.5% ($n = 22$) papers of our corpus provide a definition of trust resulting in 11 different definitions. 50% of these definitions are adopted directly from the social sciences literature on Human-Human trust [18, 138, 242] or adapted to Human-Machine trust, based on the grounds of social sciences [51, 115, 116, 130][6]. Other papers provide definitions based on a review of existing definitions of trust in Human-Machine trust [179] or propose their own based on Human-Machine and Human-Human trust definitions [2]. Finally, three definitions' origins were not provided

[6]We consider [116] and [115] as one definition source, because both of them have the same first author and are almost identical.

Sections	Guidelines	Research Opportunities
Definition *Section 4*	**(G1)** Provide a clear definition of trust **(G2)** Prevent any confusion between trust and related constructs	
Participants *Section 5*	**(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 *Section 6*	**(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 *Section 7*	**(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 *Section 8*	**(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 *Section 9*	**(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.

[21, 193, 235]. We thus observe a variety of definitions in the few reviewed papers with definitions. Three of them are most reoccurring[7] (appearing in 15 out of 22 papers; 68%):

(1) *"An attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability"* [116] (n = 10, 45.5% of the 22 papers with definitions);
(2) *"The extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid"* [130] (adapted from [139]) (n = 3, 13.6%);
(3) *"The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party"* [138] (n = 3, 13.6%).

4.2 Elements of Trust

By looking at the common terms in all the trust definitions in the corpus, we identify three most common types of phrases characterizing trust (summarized in Table 2). They mirror three key elements of trust which arise in economics, psychology, and sociology [192][8]: trust is linked to a situation of **vulnerability** and **positive expectations**, and is **an attitude**.

[7]See supplementary materials for the full list of trust definitions.
[8]It is not surprising as many definitions of our corpus rely on the one proposed by Mayer et al. [138], a slightly modified version of the most widely accepted trust definition in social sciences [52, 192]

All the reviewed definitions (11) define trust as an attitude, with one paper explicitly stating that it is an "unobservable variable" [235]. 8 definitions include phrases related to positive expectations [21, 51, 116, 130, 138, 179, 193, 242], but only 3 definitions [116, 138, 235] mention vulnerability Vulnerability and positive expectations emerge from these definitions as they are the condition for trust to exist [90, 127], and the idea that trust is an attitude dictates how it should be investigated and measured.

Table 2. List of Trust Definitions in Human-AI Papers

Definition	Origin	Vulnerability	Positive Expectations	Attitude	Papers
Lee and See [116] and Lee and Moray [115]	Automation, adapted from Human-Human Trust	✓	✓	✓	[53, 61, 111, 181, 209, 239, 244, 245]; [246]
Mayer et al. [138]	Human-Human Trust	✓	✓	✓	[62, 169]
Ekman [51], a combination of [116] and [138]	Automated Vehicles, adapted from Automation and Human-Human Trust	✗	✓	✓	[232]
Madsen [130] (adapted from McAllister [139])	HCI, adapted from Human-Human Trust	✗	✓	✓	[48, 109, 237]
Young and Albaum [242]	Human-Human Trust	✗	✓	✓	[247]
Bonn and Holmes [18]	Human-Human Trust	✗	✗	✓	[77]
Rajaonah et al. [179]	Review of Human-Automation and Human-Computer Trust	✗	✓	✓	[178]
Their own definition	Review of Human-Human and Human-Computer Trust	✗	✗	✓	[2]
Flawed source stated	-	✗	✓	✓	[21]
No source stated	-	✗	✓	✓	[193]
No source stated	-	✓	✗	✓	[235]
No definition: [1, 4, 8, 22, 24, 26, 27, 33, 34, 38, 42, 50, 56, 60, 67, 71, 75, 81, 84, 87, 89, 91, 97, 107, 112, 118, 122, 123, 126, 136, 150, 154, 156, 165, 166, 170, 176, 177, 180, 185, 196–198, 201, 207, 210, 214–216, 219–221, 225, 227, 228, 233, 234, 238, 241, 243, 248]					

To better illustrate these elements, 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 [90, 127]. 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 [30]. However, it is necessary to distinguish two natures of uncertainty (sometimes referred as risk vs. ambiguity [104]): 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). In this paper, the notion of vulnerability relates to both types of uncertainty. Without vulnerability, there is no need for trust to emerge [30, 66, 114, 160].

Similarly, trust will not emerge if the patient does not have *positive expectations*, the second key element, about the treatment the doctor assigned them. Even if the patient decides to follow it anyway, we cannot claim that the patient trusts the doctor [90, 127]. Trust has grounds to form only when one thinks that negative outcomes associated with trusting do not exist or are very unlikely [121].

The third key element is that trust does not systematically translate into a behavior. For example, the patient's level of trust might not be sufficient enough to follow the doctor's suggestion or the patient trusts the doctor's suggestion enough, but, lacks financial resources to follow it. It is also possible to have actions without trust if the patient has no other option, but to follow the doctor's suggestion. A socio-cognitive approach to defining trust suggests that trust is rather an attitude [30], i.e. a certain way of feeling about the object [16]. Trust then cannot always be 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 (see sections 8 and 9).

The definition of trust plays a role for an experimental set-up (*vulnerability* and *positive expectations*) and choice of trust measures (*attitude*). Therefore, the first guideline would be:

G1 **Provide a clear definition of trust** in a paper, which would guide researchers in their experimental protocol as well as readers in better understanding of the decisions behind it.

In the rest of our paper, we rely on the Lee and See [116] definition when referring to trust: "*An attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability.*" We favour this definition as it comprises the three key elements of trust and is the most used definition in the corpus as well as in the Human-Automation literature. We note, however, that mentioning both "vulnerability" and "uncertainty" in this definition appears to be redundant since vulnerability already comprises uncertainty as explained above. On the other hand, it is the only definition in the corpus that explicitly highlights the notion of uncertainty as impossibility to estimate the likelihood of outcomes, which characterizes the general theme of the scenarios employed in our corpus - decision making with uncertain outcomes.

4.3 Constructs Related to Trust

The three elements presented above are constitutive to trust. Concealing one element leads to consider a different concept, yet related to trust. For instance, without an element of vulnerability in a situation, it would be more appropriate to consider **confidence** instead of trust [52, 128], which is the case in 3 definitions from the corpus [21, 130, 193]. Continuing with the previous example, let's say the illness is not severe and the treatment is unlikely to have serious side effects. When the patient decides to follow the treatment without considering any alternatives and thinks that they will only be better off with it, this suggests that the patient is *confident* in the doctor's suggestion. When there is more of vulnerability for the patient's health, the patient might start looking into alternatives or into not accepting the treatment at all. If in the end they decide to follow the doctor's suggestion despite potential serious side effects of the treatment, this suggests that the patient *trusts* this suggestion [128].

Without positive expectations, it is more appropriate to discuss **distrust**. This construct is often confounded with low levels of trust [141]. While there are some researchers who deem trust and distrust as the opposite ends of one construct [191], recently the community views them as two separate ones [120, 203]. This means that they can both reach high and low levels and exist simultaneously. Just like for trust, too much of distrust can be harmful as it can lead

to inability to identify correct recommendations. Only calibrated levels of trust and distrust are beneficial for decision-making as under this condition users are less likely to blindly follow incorrect recommendations and to override correct ones [141].

Sometimes trust is confounded with behaviors such as **reliance** and **compliance**. The former is defined as the decision to follow someone's recommendation, and the latter as the decision to ask for a recommendation in the first place. As we have discussed before, trust does not always translate to a behavior, but there is definitely a relation between them [43, 44, 86, 88, 116, 145]. Figure 3 summarizes how these constructs are related to trust.

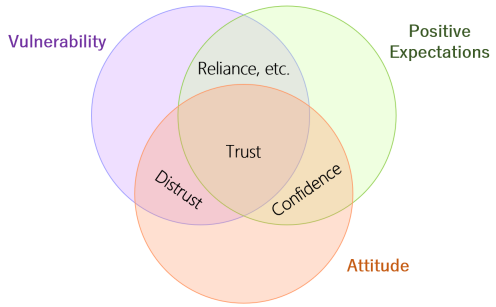

Fig. 3. A simplified representation of some constructs related to trust and how they are connected with the key elements of trust.

Finally, trust is also related to **perceived trustworthiness**. If the patient thinks that the doctor is trustworthy (e.g., has many diplomas, was recommended by someone), this does not mean the patient will trust them. Generally, perceived trustworthiness is not recommended to be used as a proxy for how much another person trusts the counterpart. In addition, having beliefs about the degree of someone's trustworthiness does not involve any vulnerability, a key element of trust [68].

As these constructs are related to trust, they are sometimes used interchangeably, leading to a further theoretical entanglement between these related terms.

G2 To **prevent any confusion between trust and related constructs**, such as confidence, reliance, distrust, and trustworthiness, one should put particular care on the choice of terminology.

5 PARTICIPANTS

We now discuss elements related to the participants' profiles and how they can impact trust formation and development.

5.1 Experience and Expertise

Prior Experience. Studies from the corpus generally involve participants with no prior experience with neither the considered AI-embedded system nor the task associated with it ($n = 49$, 59%). Some experiments recruit participants with prior experience with the task at hand ($n = 25$, 30.1%), for instance, with the expert domain or the AI-embedded system ($n = 6$, 7.5%). Additionally, six papers provide a training session before the actual experiment. In these papers, expertise is either assessed by asking about their educational and professional backgrounds or by testing their knowledge on the topic.

Our past experiences drive our expectations [168, 211], and can affect the way we update our beliefs. For example, if the past experience with a system was negative, a participant is more

likely to overreact to an error during an experiment, reconfirming their initial expectations [54]. Therefore, inquiring about participants' prior experience can help in analysing the study's data. It is thus recommended to:

G3 **Assess the expertise and prior experience of users** regarding both the AI-embedded systems and the task when running a study.

Subjective expertise. A small subset of papers (n = 10, 12%) also ask participants about how well they think they understand how to use the system or, in other words, measure their subjective expertise (also called self-confidence or self-efficacy [171]). 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 [124, 240] and in turn might affect trust-related perceptions and decisions [117].

It is believed that its magnitude depends on gender [13], age or culture [93]. In our corpus, only 3 studies consider these individual differences, but they do not link them to self-confidence. The first research opportunity would be:

RO1 **Investigate individual differences** related to self-confidence, gender, culture, and beyond to establish their precise impact on trust.

It is, therefore, important to:

G4 **Assess self-confidence, or subjective expertise,** of participants in studies on trust in AI-embedded systems alongside with other individual differences.

Subjective expertise can explain a variation in users' trust as well as objective skills and knowledge, but it is not entirely clear yet why and how exactly in the context of decision making with AI-embedded systems.

5.2 Groups and Stakeholders

Number of participants. The average number of participants per study is 134 (SD = 259.9, median = 48), which is high in comparison with standard HCI experiments. Such a high number can be explained by the fact that some crowd-sourcing studies (n = 29, 33.7%) recruited very large number of participants (i.e., 1994, 757, and 1042 in [241]). Consequently, the average number of participants per crowd-sourcing study is 340 and is much higher than the one of the rest - 51. This could be an indicator that in the corpus, trust has been mostly studied by recruiting a large number of participants in order to compute quantitative correlates (see Section 8). It could also be due to the fact that studies related to social sciences and psychology are recommended to recruit a larger number of participants [23, 187]. As trust is a psychological construct, one should:

G5 **Favour higher numbers of participants** than in standard HCI experiments [28].

Individual vs. group of participants. The predominant trend in Human-AI trust with decision-making is to investigate trust of an individual (n = 81, 97.76%). However, a line of literature in social sciences suggests the importance of considering trust of a group (e.g., [46, 63, 92]). Indeed, group decisions with an AI-embedded system are part of real-life cases, especially in the medical field (e.g., [238]). Moreover, group decision-making and trust processes have been shown to be different from the individual ones [103]. For example, repairing trust has been found to be more difficult for groups than for individuals [103]. In our corpus, we found only two papers that investigate trust of a group with decision-making, and they look into groups of 2 users [201, 225]. Thus, there needs to be more research on:

RO2 **Investigating how groups of users trust an AI-embedded system** and collectively make a decision similarly to real-life scenarios involving several users.

Direct vs. indirect interaction. In most experiments (n = 75, 93.75%), the participant has a 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 to investigate their trust, too. For example, would patients still listen to the doctor if they had known beforehand the doctor is assisted by an AI for diagnosis assessment [35]? Would citizens be upset to the same extent about a new bus schedule if it had been created manually instead of with the help of an AI [95]?

Discussions around this type of trust, referred to as indirect trust, is predominately found in the research community of reputation systems, mostly with a purpose of software optimization [78]. In the reviewed corpus, we found that automated vehicles research starts to focus on studying trust of indirect stakeholders such as pedestrians, because they are also affected by the decisions of direct users [1, 87, 180]. Besides automated vehicles, the attitudes towards AI-embedded systems of stakeholders, who are affected by the decisions of direct users, is also explored with algorithms [22, 233]. This promising line of research should be further expanded by:

RO3 **Investigating how AI-embedded systems are perceived by stakeholders indirectly impacted** by the decisions made with the help of such systems in various contexts.

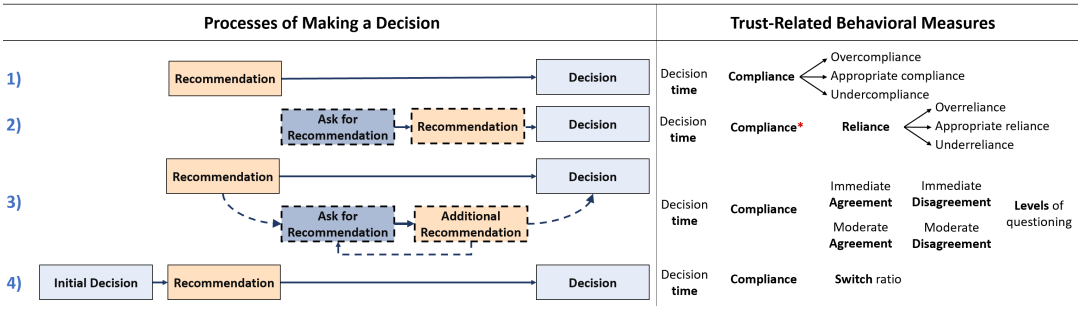

Fig. 4. A schematic representation of different types of decision making processes and behavioral measures associated with them (discussed further in Section 8.2). Dotted lines indicates that the step is optional. * indicates that the measure is possible only if optional steps are taken.

6 TASK

In this section, we examine different aspects to consider when elaborating a task, focusing on the process of decision-making and its outcomes.

6.1 Interaction flow in the Decision Making Process

We found a large variability in the process of making decision as illustrated in Figure 4. The most common and simplest pattern consists of presenting a recommendation and letting the participant follow, or not, this recommendation (case 1; n = 46, 60.8% of all the quantitative studies). This setup reflects, for instance, an autonomous system detecting a risk and suggesting an alternative way of working: the operator is then free to accept or reject this recommendation. An alternative process does not automatically provide a recommendation. It gives participants the choice of questing it if needed (case 2; n = 6, 9.2%). This process reflects, for instance, when a person is free to choose their source of recommendations if any at all. It also has the advantage to study participants' reliance on the system. These two processes can be combined as illustrated in Figure 4. After having provided

an initial recommendation, participants can ask for one or several recommendations or additional information (case 3, $n = 3$, 4.6%). Asking for several recommendations is common in Human-Human trust [100, 106, 113]. For instance, asking the opinion of several doctors before deciding for a specific treatment. Finally, case 4 is interesting ($n = 14$, 21.5%) because it is rare to get a recommendation *after* having made a decision in a real-world scenario. For example, in the 7 papers in our corpus that studied functioning prototypes, the recommendation was shown immediately, without a request from participants. The practicality of case 4 in experimental settings makes it possible, however, to compare the decision made before and after receiving a recommendation, and thus to capture the degree of participants' compliance with the system.

The choice of an interaction flow affects what logs researchers can record related to participants' decisions, how and if these decisions evolve. As a consequence, this choice also impacts what other measures researchers can calculate (right part of Figure 4, learn more about these measures in 8.2). Yet, usually studies in the corpus do not motivate their choice of the interaction flow. The only exceptions are the 14 papers that adopted case 4 to explore whether and when participants changed their decisions. In other cases, the link between an interaction flow and possible measures was not payed attention to. While developing a study, researchers should:

G6 **Consider alternative interaction flows** to derive measures related to trust (e.g. compliance, reliance, etc.) and highlight their variability with respect to the scenario at hand.

The choice of an interaction flow also changes the conditions under which recommendations appear: mandatory and immediate (case 1 and case 3), unique non-mandatory (case 2), and mandatory and non-immediate (case 4). It still remains unclear what impact this could have on participants' trust as they would receive additional advice from the system under different circumstances. It would be, hence, interesting to:

RO4 **Investigate the impact of the interaction flows, as factors, on trust** and to study to which extent the findings might be generalized from one case to another.

6.2 Feedback

Once participants made a decision, they might receive feedback. In the majority of the cases, it is about participants' performance, and rarely about the one of the system. It can be done through verbally stating that the participants' decision was either correct or wrong ($n = 30$,46%) as well as through updating participants' score based on the correctness of the decision ($n = 5$). This means that participants can infer the accuracy of the recommendations indirectly depending on whether they followed one or not and whether the decision turned out to be correct or wrong. For example, a participant that did not follow a recommendation and was told that that their decision was wrong can infer that the recommendation was correct. In the case when participants get feedback only after several decisions usually in the form of the percentage of correct decisions ($n = 3$), they could infer the general accuracy of the system's recommendations, but would not be able to know which recommendations were correct or wrong. Rarely, direct feedback is given about the system's accuracy such as number of errors the system made, updated after each mistake or percentage of correct recommendations after several decisions ($n = 1$ each). In this case, participants learn directly about the system's accuracy and indirectly about their performance.

The feedback can be **immediate** ($n = 39$, 60%) allowing participants to dynamically update their level of trust. For instance, you will immediately update your trust when you realize you followed a slower route with a lot of traffic due to AI's recommendation. However, in many scenarios, feedback can only be received after some **delay**, e.g. the consequence of the choice of a medical treatment. In our corpus, only 7.6% of the studies provide feedback after a block of decisions ($n = 4$), after one day ($n = 1$) or even within weeks ($n = 1$). This indicates there is a lack of studies with more

real-world scenarios. For example, while assigning a treatment or giving out a loan, the decision maker might learn whether they were wrong over a longer period of time - weeks, months or years. As it remains unclear,

RO5 **The effect of such delayed feedback on the way trust evolves needs to be investigated**.

6.3 Task Outcomes

Vulnerability is one of the pillars of trust (section 2). While (im)possibility to predict the likelihood of outcomes is introduced to the experiments through the nature of scenarios AI-embedded systems are used for in the studies (e.g., medical decisions, rescue operations), introducing undesired and regretful outcomes might require more thought. To immerse participants in the state of vulnerability, they must feel that their decisions matter, that is having something at stake. Therefore, task outcomes play an important part in triggering a sense of vulnerability in participants. Researchers have to:

G7 **Ensure they involve vulnerability** through task outcomes (e.g., monetary incentives), to avoid a mismatch with confidence in data collection.

In our corpus, 12 studies included no element of vulnerability. Vulnerabilities associated with decisions should be realistic enough, which can be introduced through real incentives (e.g., monetary bonuses and maluses, avoiding injuries). However, this option might not always be attainable in experimental settings (e.g., budgetary constrains, no life can be put in danger). Instead, virtual incentives (e.g., game points, lives of virtual teammates at risk) can be used as a replacement. When using virtual incentives, one should:

G8 **Assess participants' likeliness to exhibit realistic behaviors**, that is how immersive they are and to which extent participants are engaged.

We now discuss how exactly the studies in the corpus account for vulnerability.

6.3.1 Real Incentives. Only few studies introduce real outcomes ($n = 20$, 30.8% of the 65 quantitative studies). One of the strategies includes **temporal** incentives ($n = 3$, 4.6%). For every wrong decision, participants have to wait a couple of seconds which can quickly be annoying. Another one is **monetary** incentive ($n = 12$, 18.5%) where participants can receive only performance *bonus* ($n = 8$, 12.3%) or *bonus and malus* ($n = 4$, 6.2%). This strategy is widespread in economics as it has been proven that participants tend to give more optimal answers and avoid random guessing (with an exception for when the task is too complicated) [29, 85]. If the bonuses are too small, participants might disregard them and feel unmotivated to perform well [73]. If the bonuses are too high, this might put unnecessary pressure on the participants, hindering their motivation [14]. Another strategy is **cognitive effort** incentives ($n = 2$, 3.1% e.g. solving a puzzle for a long time and losing in the end [111]).

6.3.2 Virtual Incentives. The majority of studies ($n = 48$, 73.85%) use virtual incentives, presumably because they are easier to implement. Among them, we differentiate **virtual penalties** ($n = 11$, 16.9% e.g. game points [154, 244–246]), **negative virtual consequences for participants** ($n = 29$, 44.6%, e.g. car accident [136, 178]) or **negative virtual consequences for other stakeholders** ($n = 9$, 13.4%, e.g. injury or fatality [53, 220]).

However, it is unclear whether virtual outcomes (e.g. virtual car accident) might replace real ones and produce a sense of vulnerability. Recent findings in experimental economics suggest that if the virtual environment is immersive enough (through a presence questionnaire: e.g., [231]), participants might suppress the feeling of participating in an experiment and consequently demonstrate more realistic behaviors in decision-making tasks with risk [79]. It remains that participants know that the researchers are not allowed to hurt them. For instance, one study simulates an emergency

evacuation but participants rated it 1.5 out of 7 on credibility [185]. More exploration needs to be done to:

RO6 **Investigate to what extent virtual outcomes might replace real ones** and produce a sense of vulnerability.

7 PROCEDURE AND DESIGN

While the previous section focuses on task, this section discusses how the task is integrated in the whole experiment.

7.1 Introducing the System Performance

Positive expectations are a necessary component of trust (Section 4.2). If before or at the initial stages of interaction participants do not have positive expectations about the system, then trust will not start forming and developing. It is thus important to:

G9 **Control participants' expectations about the system** in the beginning of an experiment.

We note, however, it is more appropriate to do so in studies that explore various aspects of trust and its factors rather than studies directed at evaluating a system. In the latter case, the evaluation might be biased due to the deceiving priming effect. In the corpus, we identified three main strategies for establishing initial positive expectations.

The first one is **instructions**. Two studies [237, 241] directly signal to participants the systems' accuracy percentage (stated accuracy). Four studies [8, 225, 234, 247] follow a less direct approach and introduce their systems claiming they have appropriate expertise for the task, without going into many details. For instance, the systems are simply described to be "reliable" [247] to do the task. Three other studies [8, 225, 234] mention the system had relevant past experience.

The second is the **initial experience** when interacting with the system. Several studies make the system error-free for the first recommendation [237] or in the first group of recommendations [244–246] to evoke positive expectations. Indeed, the effect of a mistake occurring during the early stages of an interaction on trust is unlikely to diminish over time [116, 134, 184]. A mistake during the last stages of interaction can also negatively distort the trust reports due to a bias in memory [102].

The third one is the **behavior** of the system itself, by guarantying a minimum level of accuracy. Indeed, previous studies indicate that 60% - 70% accuracy is considered to be a threshold for investigating users' trust in AI-embedded decision-support systems [241, 244, 248]. Below this threshold, the study is more likely to study distrust rather than trust. We would expect this threshold, however, to be context-dependent. For example, 80% in the medical field would be too low.

7.2 Experimental Design

Between-subject design is especially appropriate when the investigated factors can introduce learning effects ($n = 43$) (e.g. the way system communicates) or are related to the profile of the participants (n=5) (age, nationality, etc.). However, it requires a large number of participants. In contrast, **within-subject design** requires less participants if it is compatible with the research question. For example, studies can adopt a within-subject design for investigating *accuracy of the system* if they are interested in how different levels of accuracy affect users' trust. If it is not the case and running a between-subject study is not possible, one can to increase the elapsed time between the different conditions[9] (e.g. 2-5 days apart) to reduce learning effects [237].

[9]Condition is a level of the independent variable that is manipulated by the researcher in order to assess the effect on a dependent variable (from American Psychological Association Dictionary). For example, system's accuracy as variable can have three conditions - low, average, high.

7.3 Assessing Pre-, Post-, or Dynamic Trust

It is common practice to use questionnaires during the experiment to capture different aspects of trust (see Section 8.1). Introducing a questionnaire **before** the interaction with a system (n = 3, 5.5% of the 55 studies that included questionnaires) captures participants' initial trust, based on their own beliefs and previous experiences if any. **After** the interaction (n = 22, 40%), it captures participants final trust in a system, affected by the recent interaction. However, these approaches do not capture changes in the participants' trust in the system. Trust is dynamic, it can be increased, decreased, repaired, and maintained [119]. To capture some of these changes, an alternative is to introduce the same questionnaire **before and after** the interaction (n = 5, 9.1%). In within-subject studies, it is common to introduce the trust questionnaire **after each condition** (n = 13, 24.1%) to avoid interference between conditions.

Another approach is to investigate trust at a smaller time-scale - at a scale of a trial. If we consider an experiment as a collection of repetitive events, one of these events is a trial. In the context of the studies in our corpus, a trial usually consists of participants making a decision following or not a recommendation. Questionnaires can be introduced **after each trial** (n = 13, 23.6%) or **after each block, or group, of trials** (n = 6, 10.9%). While this approach might better capture the dynamics of trust, e.g. if there were any spikes in the levels of trust, and what trial exactly caused such fluctuations, it increases the length of the experiment and/or requires short questionnaires.

In summary, one major practical challenge of *Procedure* is the length of the experiment, then:

G10 **Long interaction phases should be favored for capturing the dynamics of trust.**

Moreover, trust requires pre- inter- and/or post-treatments (e.g. questionnaires) which are as long as the interaction phase. Considering several conditions also increases the length of the experiment. However, more than a third of studies (n = 29, 35%) last 1 hour or less, the interaction time being limited to about 34.5 minutes (SD = 29.7). A main challenge for future work is to:

RO7 **Develop new methodologies to investigate dynamic trust in practical settings**.

Ideally, they should not be too long and intrusive in the course of an experiment, and should try to capture trust measures as continuously as possible (more about different types of measure see Section 8).

We will now discuss first quantitative and then qualitative methods used for studying trust, which are summarized in Figure 5.

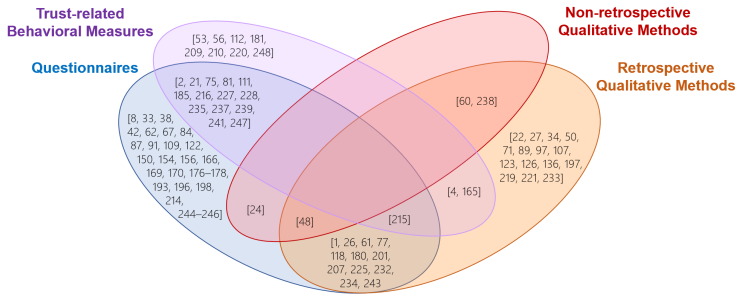

Fig. 5. Distribution of papers according to which types of quantitative and/or qualitative methods used. There are 66 quantitative trust studies and 34 qualitative ones. Questionnaires (n = 56) and Trust-related Behavioral Measures (n = 25) belong to quantitative methods. Non-retrospective (n = 4) and Retrospective (n = 34) belong to qualitative methods.

8 QUANTITATIVE MEASURES

We distinguish *questionnaires* (multi-question and single-item) and *behavioral logs* to quantitatively assess trust in Human-AI interaction.

8.1 Questionnaires

Questionnaires usually consist of a series of questions, a minimum of three or four [186]. They are a common method to measure Human-Human trust because [69]: (1) they allow to capture a person's attitude, what they feel and think [119]; (2) participants might feel more at ease reporting their psychological state as they are not facing another person, but just a screen or a sheet of paper; and (3) they are relatively easy and quick to implement, allowing to collect a bigger quantity of data in a shorter period of time (in comparison with interviews and observations).

8.1.1 Questionnaires Origins. 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 (refer to supplementary materials for their full text). Four of the questionnaires originate from Human-Automation literature [33, 96, 144, 149] and are cited by almost half (n = 15) of the papers with questionnaires. 2 questionnaires were taken from Human-Human trust literature [137, 161, 230] (n = 5, 15.6% of the 32 papers that used questionnaires). The remaining questionnaires originate from Human-Robot trust [190, 195] (n = 4, 12.5%), e-Commerce [142], Human-Agent Interaction [33], Automated Vehicles [40], and HCI [130] (n = 1, 3.1% each). Finally, 6 questionnaires were not explicitly attributed a source. All the papers, but one [33], use an already existing questionnaire. [33] introduced their own questionnaire, which accounts for cultural influences on trust. 3 papers [67, 150, 243] combine multiple existing and validated trust questionnaires to create a new one for their studies.

Such a variety of questionnaires in the corpus mirrors a general trend of a quite broad choice of questionnaires among the existing trust questionnaires in social sciences [140]. This could be explained by the efforts of both of the communities to make the questionnaires more context-specific. However, for now, it has led to the abundance of choice, inhibition understanding of how to appropriately choose and use a questionnaire.

8.1.2 Link Between Trust Definitions and Trust Questionnaires. The review of Human-Human trust questionnaires [140] suggests that the Mayer questionnaire [137] equally comprises all the theoretical concepts related to trust without focusing on the related constructs. It also named two other trust questionnaires that reflect well the trust definition: Boundary Role Persons (BRP)[10] [39], and Behavioral Trust Inventory [68].

However, theoretical discrepancies often appear both in many Human-Human trust questionnaires [47, 74, 140] and in the questionnaires from our corpus. Indeed, several questionnaires in our corpus mainly focus on *positive expectations* (n = 8, 38.1% of the 21 questionnaires) *or vulnerability* (n = 8, 38.1%). In contrast, the questionnaire by Mayer [137] (n = 3) and the one by McKnight [142] equally focus on vulnerability *and* positive expectations [143][11]. Finally, one of the most reoccurring trust questionnaires [96] (n = 11, 34.4% of the 32 papers with questionnaires) in the corpus also includes vulnerability and positive expectations but still faces some theoretical discrepancy. 5 out of 11 of its questions are reverse coded and thus related to distrust. However, the research community preferably regard trust and distrust as two separate constructs (see section 4.3).

[10] Boundary Role Persons is an umbrella term for employees who represent their company/group outside the organization and collect and rapport information from external sources to their employers.

[11] See supplementary materials for the questionnaires' items.

8.1.3 Questionnaires Modifications. More than half of the papers of our corpus using multi-question questionnaires (n = 19, 59.4%) introduce modifications to the original, validated questionnaires. It includes changing some words in the questions to better fit the case study (n = 4), e.g. replacing the word "system" with "decision aid" [156] or reducing the number of questions (n = 8).

However, most of these papers do not report what are the changes (n = 8, 42.1% of the papers with modifications). Additionally, none of the modifications have been validated in all the 16 papers. This can undermine the questionnaires reliability and the accuracy of the replications as even small changes such as replacing a word or inverting two questions can invalidate the investigation of complex constructs such as trust [194].

The abundance of possibilities between a plethora of validated trust questionnaires and of opportunities to modify them can make the choice of a trust questionnaire challenging. One should:

G11 **Favour well-established questionnaires that equally comprise all the theoretical concepts related to trust** without focusing on the related constructs [68].

The examples of such questionnaires are [11, 39, 68, 137, 142]. If an existing questionnaire needs modifications to better fit in the context, researchers should also ensure that these changes are consistent with the trust definition as per G11.

8.1.4 Single-item Questionnaires. Some questionnaires can have a single question (n = 24) which asks participants either to rate the trust in the system, e.g. "How much did you trust our machine learning algorithm's predictions on the first twenty speed dating participants?" [241] or to rank the systems, e.g. "Rank the agents in order of trust" [21]. Single-item questionnaires have the advantage to be quick for participants to answer [186], but they are generally less appropriate to study complex constructs like trust [182]. Specifically, when it comes to measuring appropriate trust, there is an issue in determining which score on the Likert scale (e.g., 3 or 4) was exactly appropriate trust. The bigger question is whether "rating" trust is insightful and meaningful enough and if participants can objectively assign a score to their trust levels. There is a need to:

RO8 Better understand **whether single-item questionnaires capture trust as well as other measures**.

8.1.5 Psychometric Statistics for Replication. It is a good practice once the participants' responses are collected, to use *psychometric statistics* to verify whether a reused or modified questionnaire still measures trust accurately in a new, independent study. However, we found that only 14 papers out of 34 (41.2%) reported psychometric statistics in data analysis. This echoes McEvily and Tortoriello's review on Human-Human trust [140], showing that most studies in the field did not report enough information on psychometric statistics in the questionnaires' analysis. Moreover, when it is done, the analysis often uses Cronbach's alpha [172] requiring several conditions to hold true, which are rather strict (e.g. unidimensionality of the construct). Additionally, reported alone, Cronbach's alpha gives little insight about the questionnaire [202]. The ω *coefficient* [41, 49, 217] might be more appropriate as it has more relaxed requirements, and other statistics such as confirmatory factor analysis (CFA) [101, 147, 174] need to be reported, too. See [173, 194] for more information about different types of psychometric statistics and how to implement them. We thus strongly encourage the community to:

G12 Adopt the practice of **reporting psychometric statistics** to ensure that a reused or modified trust questionnaire yielded data of good quality.

8.2 Trust-related Behavioral Measures

In our corpus, 25 papers (37.9% of the 66 quantitative papers) record logs about participants' activity and use them to derive what is often referred to as "behavioral measures" of trust. As trust cannot be always inferred from a behavior (section 4.2), it might be misleading to refer to these measures as "behavioral trust measures". To avoid confusion, we preferably:

G13 Use the term ***trust-related behavioral measures*** instead of *behavioral trust measures* [141].

We identify four types of trust-related behavioral measures.

8.2.1 Trust-related Behavioral Measures Based on Following Recommendations. Figure 4 illustrates different processes to make a decision and the associated quantitative measures. Two measures are independent of the process, **Decision Time**, i.e. how fast a recommendation is accepted (n = 2, 8% of the 25 papers with trust-related behavioral measures, [56, 247]) and Compliance. **Compliance** is the number of times participants follow the systems' recommendations (n = 18, 72%), both correct and incorrect ones. It is then possible to calculate:

- *appropriate compliance*: correct recommendations accepted (n = 2) and incorrect recommendations non-accepted (n = 2) ;
- *overcompliance*: incorrect recommendations accepted (n = 3);
- *undercompliance*: correct recommendations rejected (n = 1).

When the recommendation is not initially provided (case 2, Figure 4), it is also possible to estimate **Reliance** - the number of times participants asked for a recommendation (n = 4, 16%) and to derive [210]:

- *appropriate reliance*: *requested* recommendation when it was *beneficial* and *did no request* recommendation when it was too *costly*;
- *overreliance*: requested recommendation when it was too costly (n = 1);
- *underreliance*: did not request recommendation when it was beneficial (n = 1).

Additionally, when participants are free to ask several (typically up to two) recommendations [21, 75, 239], the first one being automatically given (case 3, Figure 4). We can thus derive the following measures, where how quickly a recommendation is accepted is considered to be indicative of high levels of trust:

- *Agreement*, when the initial recommendation is immediately accepted;
- *Moderate agreement* when asking for a second recommendation and accepting it;
- *Moderate disagreement* when asking for a second recommendation and rejecting it;
- *Disagreement* when the initial recommendation is immediately rejected;
- *Levels of questioning*, how many times an additional recommendation was asked.

Finally, when participants indicate an initial decision (before receiving the recommendation (case 4, Figure 4), we can estimate the **Switch ratio**, the number of times a participant who initially disagreed with the system decided to follow its recommendation in the end (n = 3, 12%). It is assumed the higher the switch ratio is, the higher the levels of trust are.

Only 4 papers in the corpus break down trust-related behavioral measures into more granular ones. These measures can provide more nuanced insights about the way participants integrate AI-based system's recommendations in their decision making relative to the system's performance. For example, low participants' compliance rate can be interpreted differently depending on whether most of the recommendations were wrong or not. Researchers are then encouraged to:

G14 Use **trust-related behavioral measures relative to the system's performance** to be able to assess their appropriate, over-, and under-levels.

8.2.2 Other Trust-related Behavioral Measures. [216] links the amount of money shared with the system as a trust-related behavioral measure, inspired by game theory situations such as *Prisoner's Dilemma* [44, 45, 125]. However, such games are criticized for confounding trust with altruism [37] and betrayal aversion [17, 55], and for the lack of stability [25, 99, 208], and hence, should not be preferred for measuring trust-related behaviors. Finally, measures were related to scenario-specific events such as how long the brakes were hold for and with what intensity [61] in an automated vehicle.

8.2.3 Physiological Measures. In quest of collecting objective trust data, which is not under participants' control and, hence, is less subjective than responses to trust questionnaires, researchers start turning to physiological measures [155]. There is some evidence that higher levels of stress are associated with lower levels of trust. For instance, **Heart Rate Variability** (HRV) (n = 2, 3% of the 66 papers with quantitative studies) measures the variability of time interval between heartbeats [200]. As elevated levels of stress can be indicated by low HRV, this could also be an indicator of lower levels of trust. Another example is **Galvanic Skin Response** (GSR) (n = 2, 3%) which measures the intensity of an experienced emotion with the electrical conductance of the skin, which varies with sweat [188]. High levels of stress can be generally indicated by high GSR, and hence, could be potentially linked to lower levels of trust [148]. However, in our corpus, no relationship has been found between these measures (HRV and GSR) and trust [61, 77, 77, 232].

Another was is **Electroencephalography** (EEG) (n = 2, 3%), which records activity of the brain. This approach is more promising as there is some evidence that the predominant brain areas correlated with trust are the frontal and occipital ones [226], but the papers in our corpus either did not deeply explore the EEG data [77] or used it primarily for a preliminary model construction [2]. Finally, **hand trajectories** [59], easily captured with a computer mouse has recently been shown to reflect the evolution of decision making as well as hesitations [59, 132]. While the relationship between hand trajectory and trust is yet to be determined, this measure can provide additional information in comparison with integral measures discussed above. Research community could benefit from:

RO9 Exploring more **fundamental correlates between physiological sensing (e.g. EEG, mouse trajectories) and trust** in Human-AI interaction.

9 QUALITATIVE METHODS

Qualitative methods produce less structured data than the quantitative ones. They might thus aid in discovering new aspects of trust and build new theories [119]. We identified 10 qualitative methods to *collect* data in non-retrospective or retrospective ways among 34 papers with qualitative studies. We also identified 3 methods to analyze the collected data.

9.1 Non-retrospective Methods

Only a small number of studies use qualitative methods *while* a participant is interacting with the system. **Think-aloud protocol** (n = 3, [24, 48, 60]) 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 happens with methods used post experiment. The papers use this method to investigate participants' decision-making with a system and what role trust played in the process. **Observations** in the field (n = 1; [238]) is used to understand doctors' daily routine and decision-making process. However, this method might not be appropriate as trust does not always translate in a behavior (see section 4.2). It remains useful for preparing potential interview questions about trust post experiment.

9.2 Retrospective Methods

Retrospective methods are used after the experiment. We distinguish interview-based methods, which received a lot of attention, and non interview-based methods.

9.2.1 Interview-based methods. **Semi-structured** interviews are the most common type of interviews ($n = 24$, 82.6% of 29 studies with interviews) as they both provide control over the topic while leaving room for unexpected insights [131]. In our corpus, they primary focus on understanding participants' general experience with the system, decision-making process or general perceptions and attitudes towards a system. Only 3 semi-structured interviews primarily focus on Human-AI trust [136, 207, 238], rather than considering it as one of the multiple factors of users' experience to evaluate.

Non-structured interviews ($n = 1$, [77]) and **in-depth** interviews ($n = 1$, [97]) have been used in our corpus to study participants' general experience with AI. They both allow for gathering more personal, sensitive or confidential information, which is especially appropriate for discussing a topic of trust. In particular, in-depth interviews tend to be longer, useful to build a relationship with an interviewee and to ask more detailed questions.

We found one instance of **focus groups** (or group interviews) to study participants' general attitude to AI [233]. Focus groups are less time-consuming and less expensive than the above types of interviews but there is a risk that the responses of one person bias the rest of the participants. Moreover, some participants might get too shy to express their real opinions, especially for such personal topics as trust [163]. With this method, the paper in our corpus explore trust of people with a specific background - members of marginalized communities [233].

The following interview-based methods are especially appropriate to study the *dynamics of trust*, i.e. how trust evolves over time. **Critical incident technique** [57] 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 [7]. When applied to trust, it is especially useful to study real life cases in which trust was established, destroyed or repaired [152]. Researchers directly ask participants what aspects of others' behavior was important for trust weakening or strengthening. This information can in turn be applied, for example, towards improving patients' experience during a medical visit [229, 236] or enhancing intercultural business negotiations [152]. Although this method is an established method, we did not find it in our corpus.

Repertory grid[12] is an interview-based method relying on card sorting. During the interview, participants establish links between different elements (words, objects) which is useful to make some concepts emerging. The main advantage of this method is to minimize the researchers' influence on a study by reducing the interviewer's input and maximising the interviewee's output [9]. Researchers are thus less prompted to introduce their preconceived assumptions about whether and how an element of the studied environment affects trust. They can then study participants' understanding of trust, its development, breakage and repair processes with a reduced interviewer bias, which enhances validity of the data. This method has also been used in Human-Human trust [119], but not in our corpus, probably because it is quite time consuming. It is more suitable for

[12] An example of repertory grid based on studying trust in organizational settings. The interviewer presents a random pair of words, *elements*, related to work settings: face-to-face contact, lengthy detailed contracts, frequent emails etc. The participant indicates if these elements are similar/dissimilar with regards to trust, and explains why: "face-to-face contact and lengthy detailed contracts are similar, because they represent 'engagement' (keyword)" or "dissimilar, because the former is associated with 'transparency' (keyword) and the latter with 'bureaucracy' (keyword)". Later on, the participant indicates whether there is a link between each of the combinations of elements and keywords, which later will be translated into a cognitive map with points proximity determined by words similarity [12] (see more in Chapters 13 and 14 of [119].)

studying small groups where individual differences play an important role.

To conclude, several interview-based methods are available to study trust. Some of them, **Critical Incident Technique and Repertory Grid, should be more largely considered in Human-AI trust (see later G16)** as they have been demonstrated useful, especially to study the dynamics of trust, in other domains (e.g. Human-Human trust).

Our analysis also reveals the lack of information to evaluate or replicate interviews as well as to compare the findings between papers. For instance, only few papers provide question examples ($n = 5$, 17.2% of 29 papers with interviews) or describe the general topics of the interviews ($n = 12$, 41.4%). Among them, only [27, 97] mention they conducted a pilot study to identify the prominent questions and to refine their wording to study Human-AI trust. It is thus difficult to assess to what extent the questions were really understandable for the participants.

G15 **Reporting on qualitative studies should experience more of empirical rigor in Human-AI community** to support evaluation and replication of interviews in the context of AI-assisted decision making.

9.2.2 Non Interview-based Methods. Non interview-based methods are generally less used. However, they might be useful as they are simple and fast to collect data. For instance, some studies just let participants leave any **comment** they wish after the experiment ($n = 2$) or opt for a **open-ended question** ($n = 3$) (i.e., what-how-why questions) to study participants' general attitude to AI [107], to understand participants' decision-making [219], and to directly investigate participants' trust [71].

Another method is **UX curve** ($n = 1$, [60]), used for understanding the reasons behind long-term system use or abandonment (more about it here [110]). Participants draw a line which represents their experience with a system, saying out loud what events changed it and if they affected it positively and negatively and by how much. This method serves to get accurate and chronological insights about what, in what direction and by how much affected participants' trust and experience during an interaction with a system.

Finally, **open-card sorting** ($n = 1$, [48]) identifies what are the most important factors for participants' trust. Participants rank various pre-selected prompts and few ones introduced by them in order of importance for their trust in a system (for more details about the method [206]). Overall, it is not yet established how efficient these methods, marginally used, are to assess trust in our context. More work is needed to more systematically compare these qualitative measures.

9.3 Analyzing Qualitative Data

21 papers (out of 34) explicitly state the method used to analyze the data. These methods are: Grounded Theory, Thematic Analysis, and Discourse Analysis.

Grounded Theory aims to generate hypotheses based on the themes and categories found in the qualitative data. Consequently, the findings are the presentation of a new theory that includes the core themes [58]. Usually, these themes emerge from the data after it is annotated with *open* and *axial* coding. Open coding is aimed at summarizing small portions of text with one or two words - codes, and axis coding organizes these codes into groups. Researchers then study how these groups interact with each other to establish a theory or framework [58]. 6 papers in our corpus analyze their data in this manner [4, 34, 123, 165, 221, 243].

Unlike Grounded Theory, **Thematic Analysis** focuses on identifying the themes most relevant to the research objectives of a paper, without necessarily exploring the relationship between them[13].

Therefore, the main findings are presented as a description of the most important themes. 16 papers in our corpus state using Thematic Analysis as new theory development was not their objective.

Rather than analyzing what participants say, **Discourse Analysis** focuses on how they say it [175], i.e., the type of vocabulary, grammar, non-verbal communication used. The advantage of this approach in comparison with the above mentioned ones is that it could supply researchers with insights from the cues which participants are unlikely to voluntarily control. 1 paper in the corpus analyzed the way participants spoke to robots before making a decision, particularly the amount of words used, while studying their trust [234].

In complement to the previous methods identified from the reviewed corpus, we also introduce a method used in Human-Human trust literature. **Hermeneutics**[14] is suitable to analyze not only interviews transcript but also existing stories published in popular media outlets (e.g. [95, 189]). With the rising media coverage and popularity of workshops and webinars related to AI, researchers should **consider Hermeneutics to interpret the current narratives (see G16 below)** as an alternative data source on users' trust.

This method is most widely employed by historians and theologians to interpret human actions and their outcomes [133]. It offers a toolbox for finding patterns and common threads in texts to justify their interpretation and theories drawn from them. Gerard Breeman, researcher in trust and politics, finds hermeneutics useful for investigating the reasons why people trust and why exactly those reasons were given in that specific context [19]. Before analyzing the text, a research determines key factors that can influence trust in a certain scenario based on theory. This framework becomes a guiding thread for a researcher while analysing the text to find passages that either confirm or go against the theory. In the final step, a researcher update the framework they established incorporating new insights from the text [19]. The main limitations of hermeneutics is that this method relies on preselected concepts, which might lead to a biased interpretation and sub-optimal understanding of the case. Plus, it focuses on analyzing a very specific event, which could hinder results' generalization.

To sum up, very few qualitative studies ($n = 4$, 11.8% of 34 qualitative papers) consider Human-AI trust as their central focus. The rest of the qualitative studies in our corpus view trust as one of the multiple factors of users' experience. This finding is similar to the one by [61], which urges to use qualitative methods for deepening our understanding of Human-AI trust as little is known about its nature and how different it is from Human-Human and Human-machine trust. Some qualitative methods for studying trust stemming from other domains could be found useful in the Human-AI Interaction research, too, being that for studying different aspects of trust (i.e. dynamics of trust) or for having a tool to analyze a different type of data (i.e. media reports). We encourage the community to:

G16 **Adopt under-used qualitative methods for studying trust** in Human-AI interaction such as Critical Incident Technique, Repertory Grid and Hermeneutics.

10 DISCUSSION

Trust has recently emerged as key concept in Human-AI Interaction. While many studies investigated the factors influencing trust, our approach focuses on **how to evaluate trust** in the context of AI-assisted decision making. This survey offers a lens on existing methodologies and highlights

[13] Though developing a new theory is not a goal of Thematic Analysis, the emerged themes and their relationships can be further studied with the Grounded Theory Approach for a potential theory or framework development [58].

[14] You can find a more detailed description of the method in Chapter 15 of [119]. For more trust studies, using hermeneutic analysis, refer to Breeman [20] and von Sinner [223].

the difficulties of properly studying this multi-faceted and dynamic construct. This survey also provides an opportunity to improve validity and replicability of future experiments by proposing practical guidelines. Finally, it identifies challenges and research opportunities. We now discuss these different contributions.

10.1 Main Findings and guidelines

We summarize the main findings from our analysis of 83 papers investigating trust and AI-assisted decision making.

Our first finding (**F1**) is that trust definitions are often incomplete or even not provided . Established definitions exist in related fields [52, 138, 192] as well as HCI [116], but few studies explicitly mention any. However, trust is a multi-dimensional construct and Human-AI interaction is a recent field of research, it is, thus, important to clearly define trust to avoid conflicting terminology and misunderstanding in the community. In particular, we found (**F2**) that the three key elements of trust are not always incorporated in the reviewed studies. The sense of *vulnerability* is often missing or questionable due to the lack of realistic outcomes in the experiments. The system is not always introduced in a way that participants have *positive expectations*. Finally, several methods capture participants' behaviors while trust is an *attitude*. Consequently, several papers investigated constructs related to trust such as distrust, confidence or reliance, rather than trust. It is important that Human-AI community adopts the theoretical evidence establishing the difference between these related constructs [52, 145, 203].

We derived several guidelines from these two findings. In particular, we recommend to provide a clear definition of trust (**G1**), to introduce task outcomes **(G7)**, to control participants' expectations in the beginning of an experiment (**G9**) through instructions or system's performance or to clarify that common quantitative measures based on users' logs are generally trust-related behavior measures, i.e. do not necessary capture the attitude (**G13**).

We also found that (**F3**) there is a large variability among the designs and the measures used to assess trust. For instance, there is no "standard" task, nor procedure, nor questionnaire. While it could be explained by a variety of scenarios in the real life, it also appears that the relevance or validity of existing methods are still under debate. For instance, it is not clear to what extent behavioral, especially physiological, measures can be used as a proxy to capture trust [5, 153] or whether single-item questionnaires are as robust as multi-question questionnaires [182].

We derived several guidelines from this finding. We suggest to consider the different interaction flows (**G6**) illustrated in Figure 4 before choosing the final one to ensure it fits the research question, the envisioned scenarios as well as the target measures. We also recommend using established questionnaires that comprise all the key elements defining trust (**G11**). Lastly, more information should be reported regarding the modifications and analysis performed (e.g. in questionnaires, **G12**) and methods used (e.g. interview questions, **G15**) to foster replicability and increase scientific rigor

Beyond that, we identified (**F4**) a profound conflict between the importance of investigating the dynamics of trust and the (temporal) constraints of laboratory experiments. Indeed, trust can be developed, damaged or repaired, but the underlying mechanisms of this in Human-AI interaction are still not well understood. It thus requires interaction phases long enough to make these different phenomena happening, but also fine-grained measures to precisely capture them.

Regarding this finding, we recommend to favor interactions over a long period of time (**G11**) and to include, for instance, questionnaires at different stages of the experiment. However, laboratory experiments are often limited to one or two hours, and methods such as questionnaires are not always appropriate to reflect on users' attitude at this level of granularity. This raises several research opportunities to go beyond this trade-off.

10.2 Challenges and Research Opportunities

We identified two main classes of research opportunities. The first one is further investigation regarding **methodologies** to study AI-assisted decision making. We already acknowledged one major challenge of studying trust experimentally: the conflict between the importance of the dynamics of trust and the constraints of laboratory experiments. Future work could investigate novel practical methods which do not break the interaction flow and do not make the experiment longer (**R07**, **R08**, **R09**). In particular, several novel measures have been recently introduced, e.g. EEG, mouse trajectory. Future research could investigate to what extent these components have an impact on Human-AI trust and whether there is a relationship between them and trust (**R09**). More generally, it is important to foster connections between the Human-AI and Human-Human trust communities and to investigate how to transpose methods from other fields to the Human-AI interaction one . We propose several quantitative and qualitative methods used for studying Human-Human trust to apply for Human-AI trust, but it is not an exhaustive list. This paper hopes to promote further exploration of other fields studying trust to enrich the pool of trust methods in Human-AI Interaction community. We would also like to note that within the Human-AI Interaction community, we covered the part represented by ACM Digital Library, and hence, further exploration of methods used in this community in AAAI Digital library, IEEE Xplore and HCI journals will be beneficial.

The second class of research opportunities is the investigation of **factors** on Human-AI trust. A main challenge is to incorporate the key elements of trust (vulnerability, positive expectations, and attitude) in the experimental protocol in Human-AI interaction setting. For instance, further work should investigate the impact of task outcomes (**R06**) and scenarios (**R04**, **R05**) on trust. Another challenge is to better understand the role of individual differences (e.g. prior experience, self-confidence (**R01**)), groups (**R02**) or stakeholders (**R03**) on trust in the context of AI-assisted decision making.

Lastly, our guidelines and research opportunities are based predominantly on studies conducted in the laboratory settings with systems' mock-ups or prototypes, and further evaluation is needed on how efficiently they can be used with implemented systems in real-life settings.

10.3 AI in AI-based decision-making systems

In this paper, we have deliberately considered AI-based decision-making systems in a relatively wide sense. We did not make strong constraints on the AI technology involved (for instance, if it relies on machine learning or knowledge-based models). As a matter of fact, AI has become an umbrella term that encompasses different types of technology achieving a wide range of highly complex tasks (speech recognition, character generation, content-based recommendation, etc.). This has two implications in our work.

First, this approach allowed us to extract generic guidelines that could be used by designers, developers and HCI practitioners independently of the type of AI involved in the system, as long as the goal of the algorithm is to provide recommendations to users in a decision-making process. That said, we are aware that the study of trust is, to some extent, context-dependent, and certain results may change according to the type of system considered. For instance, we mentioned several studies that indicated that 60% - 70% accuracy could be considered to be a threshold for investigating users' trust in AI-embedded decision-support systems [241, 244, 248]. But this threshold may vary according to the application domain and the task at hand. Nonetheless, we believe that the proposed guidelines capture the fundamental elements that ensure trust to be assessed in this context.

Second, the fact that AI is considered as a generic technology able to achieve complex and high-level cognitive tasks, leads researchers in Human-AI interaction to borrow concepts and

methods from other fields (especially cognitive science, psychology and behavioural economics) in order to study the phenomena at play. In this work, we have extensively used the literature on Human-Human trust as proxy to help us draw the lines of an experimental methodology to assess trust in Human-AI interaction, i.e. we provided tools to assess if trust has formed and developed in AI-assisted decision-making. While interaction with AI-based systems has undoubtedly its own peculiarities compared to interaction with humans, we believe that, by relying on fundamental components of trust (identified from behavioural psychology studies), we broach a more universal approach to trust assessment in Human-AI interaction.

11 CONCLUSION

In conclusion, this work can benefit different types of audience. Primarily, this work can benefit to designers who look for operational guidelines to study the impact of AI-embedded systems on trust. Second, this work can also benefit to researchers through the identification of under-explored factors (e.g. participants' profile) and research opportunities regarding the methods. Third, educators can include our findings in their lectures on Human-AI interaction, too. Finally, public policy actors may see work as a framework to assess trustworthy interaction. Maybe more importantly, we expect to foster connections between the Human-AI and Human-Human trust communities. Trust is a multi-disciplinary construct requiring endeavours across fields.

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. It was also supported by the ELEMENT project (ANR-18-CE33-0002) and the ARCOL project (ANR-19-CE33-0001) from the French National Research Agency.

REFERENCES

[1] Sander Ackermans, Debargha Dey, Peter Ruijten, Raymond H. Cuijpers, and Bastian Pfleging. 2020. The Effects of Explicit Intention Communication, Conspicuous Sensors, and Pedestrian Attitude in Interactions with Automated Vehicles. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376197

[2] 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

[3] 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

[4] Alper Alan, Enrico Costanza, Joel Fischer, Sarvapali D. Ramchurn, Tom Rodden, and Nicholas R. Jennings. 2014. A Field Study of Human-Agent Interaction for Electricity Tariff Switching. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS '14)*. International Foundation for Autonomous Agents and Multiagent Systems, New York, NY, USA, 965–972.

[5] Carlos Alós-Ferrer and Federica Farolfi. 2019. Trust Games and Beyond. *Frontiers in Neuroscience* 13, 887 (Sept. 2019), 1–14. https://doi.org/10.3389/fnins.2019.00887

[6] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, and et al. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 3, 13 pages. https://doi.org/10.1145/3290605.3300233

[7] 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

[8] Sean Andrist, Erin Spannan, and Bilge Mutlu. 2013. Rhetorical Robots: Making Robots More Effective Speakers Using Linguistic Cues of Expertise. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*

(HRI '13). IEEE Press, New York, NY, USA, 341–348.

[9] Melanie J. Ashleigh and Edgar Meyer. 2011. Deepening the understanding of trust: combining repertory grid and narrative to explore the uniqueness of trust. In *Handbook of Research Methods on Trust*, Fergus Lyon, Guido Möllering, and Mark Saunders (Eds.). Edward Elgar, Cheltenham, UK; Northampton, MA, USA, Chapter 14, 138–148.

[10] Maryam Ashoori and Justin D. Weisz. 2019. In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes.

[11] Benoit A. Aubert and Barbara L. Kelsey. 2003. Further Understanding of Trust and Performance in Virtual Teams. *Small Group Research* 34, 5 (2003), 575–618. https://doi.org/10.1177/1046496403256011 arXiv:https://doi.org/10.1177/1046496403256011

[12] Reinhard Bachmann. 2011. Utilising repertory grids in macro- level comparative studies. In *Handbook of Research Methods on Trust*, Fergus Lyon, Guido Möllering, and Mark Saunders (Eds.). Edward Elgar, Cheltenham, UK; Northampton, MA, USA, Chapter 13, 130–137.

[13] 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

[14] Roy F. Baumeister. 1984. Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology* 46, 3 (1984), 610–620. https://doi.org/10.1037/0022-3514.46.3.610

[15] 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

[16] 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.

[17] Iris Bohnet, Fiona Greig, Benedikt Herrmann, and Richard Zeckhauser. 2008. Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *American Economic Review* 98, 1 (2008), 294–310. http://dx.doi.org/10.1257/aer.98.1.294

[18] S. Boon and J. Holmes. 1991. The dynamics of interpersonal trust: resolving uncertainty in the ace of risk. In *Cooperation and Prosocial Behaviour*, R. Hinde and J. Gorebel (Eds.). Cambridge University Press, Cambridge, 190–211.

[19] Gerard Breeman. 2011. Hermeneutic methods 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 15, 149–160.

[20] Gerard Engelbert Breeman. 2006. *Cultivating trust : how do public policies become trusted.* Ph.D. Dissertation. Dept. of Public Administration, Faculty of Social and Behavioural Sciences, Leiden University.

[21] Tom Bridgwater, Manuel Giuliani, Anouk van Maris, Greg Baker, Alan Winfield, and Tony Pipe. 2020. Examining Profiles for Robotic Risk Assessment: Does a Robot's Approach to Risk Affect User Trust?. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*. Association for Computing Machinery, New York, NY, USA, 23–31. https://doi.org/10.1145/3319502.3374804

[22] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-Making in Child Welfare Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300271

[23] Marc Brysbaert. 2019. How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition* 2, 1, 28. https://doi.org/10.5334/joc.72

[24] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[25] Terence Burnham, Kevin McCabe, and Vernon Smith. 2000. Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavior & Organization* 43, 1 (2000), 57–73. https://EconPapers.repec.org/RePEc:eee:jeborg:v:43:y:2000:i:1:p:57-73

[26] 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

[27] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput.*

Interact. 3, CSCW (2019), 24. https://doi.org/10.1145/3359206

[28] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[29] COLIN F. CAMERER and ROBIN M. HOGARTH. 1999. The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. *Journal of Risk and Uncertainty* 19, 1/3 (1999), 7–42. http://www.jstor.org/stable/41760945

[30] 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

[31] Raja Chatila, Virginia Dignum, Michael Fisher, Fosca Giannotti, Katharina Morik, Stuart Russell, and Karen Yeung. 2021. *Trustworthy AI*. Springer International Publishing, Cham, 13–39. https://doi.org/10.1007/978-3-030-69128-8_2

[32] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren. 2020. The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. *Computer Graphics Forum* 39, 3 (2020), 713–756. https://doi.org/10.1111/cgf.14034 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14034

[33] Shih-Yi Chien, Michael Lewis, Katia Sycara, Jyi-Shane Liu, and Asiye Kumru. 2018. The Effect of Culture on Trust in Automation: Reliability and Workload. *ACM Trans. Interact. Intell. Syst.* 8, 4, Article 29 (Nov. 2018), 31 pages. https://doi.org/10.1145/3230736

[34] Michael Chromik, Florian Lachner, and Andreas Butz. 2020. *ML for UX? - An Inventory and Predictions on the Use of Machine Learning Techniques for UX Research*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3419249.3420163

[35] 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

[36] 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

[37] James Cox. 2004. How to identify trust and reciprocity. *Games and Economic Behavior* 46, 2 (2004), 260–281. https://EconPapers.repec.org/RePEc:eee:gamebe:v:46:y:2004:i:2:p:260-281

[38] Henriette Cramer, Vanessa Evers, Nicander Kemper, and Bob Wielinga. 2008. Effects of Autonomy, Traffic Conditions and Driver Personality Traits on Attitudes and Trust towards In-Vehicle Agents. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 03 (WI-IAT '08)*. IEEE Computer Society, New York, NY, USA, 477–482. https://doi.org/10.1109/WIIAT.2008.326

[39] Steven C. Currall and Timothy A. Judge. 1995. Measuring trust between organizational boundary role persons. *Organizational Behavior and Human Decision Processes* 64, 2 (1995), 151–170. https://doi.org/10.1006/obhd.1995.1097

[40] Shuchisnigdha Deb, Lesley Strawderman, Daniel W. Carruth, Janice DuBien, Brian Smith, and Teena M. Garrison. 2017. Development and validation of a questionnaire to assess pedestrian receptivity toward fully autonomous vehicles. *Transportation Research Part C: Emerging Technologies* 84 (2017), 178 – 195. https://doi.org/10.1016/j.trc.2017.08.029

[41] Lifang Deng and Wai Chan. 2017. Testing the Difference Between Reliability Coefficients Alpha and Omega. *Educational and psychological measurement* 77, 2 (Apr 2017), 185–203. https://doi.org/10.1177/0013164416658325

[42] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of Robot Failures and Feedback on Real-Time Trust. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI '13)*. IEEE Press, New York, NY, USA, 251–258.

[43] Morton Deutsch. 1958. Trust and suspicion. *Journal of Conflict Resolution* 2, 4 (1958), 265–279. https://doi.org/10.1177/002200275800200401 arXiv:https://doi.org/10.1177/002200275800200401

[44] Morton Deutsch. 1960. The Effect of Motivational Orientation upon Trust and Suspicion. *Human Relations* 13, 2 (1960), 123–139. https://doi.org/10.1177/001872676001300202 arXiv:https://doi.org/10.1177/001872676001300202

[45] M Deutsch. 1960. Trust, trustworthiness, and the F scale. *Journal of abnormal and social psychology* 61 (July 1960), 138—140. https://doi.org/10.1037/h0046501

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

[47] Kurt Dirks and Donald Ferrin. 2002. Trust in Leadership: Meta-Analytic Findings and Implications for Research and Practice. *The Journal of applied psychology* 87 (09 2002), 611–28. https://doi.org/10.1037//0021-9010.87.4.611

[48] Jaimie Drozdal, Justin Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. 2020. Trust in AutoML: Exploring Information Needs for Establishing Trust in Automated Machine Learning Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. Association for Computing Machinery, New York, NY, USA, 297–307. https://doi.org/10.1145/3377325.3377501

[49] Thomas J. Dunn, Thom Baguley, and Vivienne Brunsden. 2014. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology* 105, 3 (2014), 399–412.

https://doi.org/10.1111/bjop.12046 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjop.12046

[50] Malin Eiband, Sarah Theres Völkel, Daniel Buschek, Sophia Cook, and Heinrich Hussmann. 2019. When People and Algorithms Meet: User-Reported Problems in Intelligent Everyday Applications. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 96–106. https://doi.org/10.1145/3301275.3302262

[51] Fredrick Ekman, Mikael Johansson, and Jana Sochor. 2016. Creating Appropriate Trust for Autonomous Vehicle Systems: A Framework for HMI Design. *IEEE Transactions on Human-Machine Systems* 48, 1 (01 2016), 95–101.

[52] Anthony M. Evans and Joachim I. Krueger. 2009. The Psychology (and Economics) of Trust. *Social and Personality Psychology Compass* 3, 6 (2009), 1003–1017. https://doi.org/10.1111/j.1751-9004.2009.00232.x

[53] Xiaocong Fan, Sooyoung Oh, Michael McNeese, John Yen, Haydee Cuevas, Laura Strater, and Mica R. Endsley. 2008. The Influence of Agent Reliability on Trust in Human-Agent Collaboration. In *Proceedings of the 15th European Conference on Cognitive Ergonomics: The Ergonomics of Cool Interaction (ECCE '08)*. Association for Computing Machinery, New York, NY, USA, 8. https://doi.org/10.1145/1473018.1473028

[54] D. S. Fareri, L. J. Chang, and M. R. Delgado. 2012. Effects of direct social experience on trust decisions and neural reward circuitry. *Front Neurosci* 6 (2012), 148.

[55] Ernst Fehr. 2009. ON THE ECONOMICS AND BIOLOGY OF TRUST. *Journal of the European Economic Association* 7, 2-3 (2009), 235–266. https://doi.org/10.1162/JEEA.2009.7.2-3.235 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1162/JEEA.2009.7.2-3.235

[56] Shi Feng and Jordan Boyd-Graber. 2019. What Can AI Do for Me? Evaluating Machine Learning Interpretations in Cooperative Play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 229–239. https://doi.org/10.1145/3301275.3302265

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

[58] Jerry Floersch, Jeffrey L. Longhofer, Derrick Kranke, and Lisa Townsend. 2010. Integrating Thematic, Grounded Theory and Narrative Analysis: A Case Study of Adolescent Psychotropic Treatment. *Qualitative Social Work* 9, 3 (2010), 407–425. https://doi.org/10.1177/1473325010362330 arXiv:https://doi.org/10.1177/1473325010362330

[59] Jonathan B. Freeman. 2018. Doing Psychological Science by Hand. *Current Directions in Psychological Science* 27, 5 (2018), 315–323. https://doi.org/10.1177/0963721417746793 arXiv:https://doi.org/10.1177/0963721417746793

[60] Anna-Katharina Frison, Laura Aigner, Philipp Wintersberger, and Andreas Riener. 2018. Who is Generation A? Investigating the Experience of Automated Driving for Different Age Groups. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '18)*. Association for Computing Machinery, New York, NY, USA, 94–104. https://doi.org/10.1145/3239060.3239087

[61] Anna-Katharina Frison, Philipp Wintersberger, Andreas Riener, Clemens Schartmüller, Linda Ng Boyle, Erika Miller, and Klemens Weigl. 2019. In UX We Trust: Investigation of Aesthetics and Usability of Driver-Vehicle Interfaces and Their Impact on the Perception of Automated Driving. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300374

[62] Ernestine Fu, Mishel Johns, David A. B. Hyde, Srinath Sibi, Martin Fischer, and David Sirkin. 2020. Is Too Much System Caution Counterproductive? Effects of Varying Sensitivity and Automation Levels in Vehicle Collision Avoidance Systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376300

[63] 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

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

[65] 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

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

[67] Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable Active Learning (XAL): Toward AI Explanations as Interfaces for Machine Teachers. *Proc. ACM Hum.-Comput. Interact.* 4, 3, Article 235 (Jan. 2021), 28 pages. https://doi.org/10.1145/3432934

[68] Nicole Gillespie. 2003. *Measuring trust in working relationships: The behavioral trust inventory.* Melbourne Business School, Melbourne, Australia.

[69] Nicole Gillespie. 2011. Measuring trust in organizational contexts: An overview of survey-based measures. In *Handbook of Research Methods on Trust*, Fergus Lyon, Guido Möllering, and Mark Saunders (Eds.). Edward Elgar, Cheltenham, UK; Northampton, MA, USA, Chapter 17, 175–188.

[70] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (2018), 80–89.

[71] Alyssa Glass, Deborah L. McGuinness, and Michael Wolverton. 2008. Toward Establishing Trust in Adaptive Agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI '08)*. Association for Computing Machinery, New York, NY, USA, 227–236. https://doi.org/10.1145/1378773.1378804

[72] Ella Glikson and Anita Woolley. 2020. Human trust in artificial intelligence: Review of empirical research (in press). *The Academy of Management Annals* 14, 2 (August 2020), 62. https://doi.org/10.5465/annals.2018.0057

[73] Uri Gneezy and Aldo Rustichini. 2000. Pay Enough or Don't Pay at All. *The Quarterly Journal of Economics* 115, 3 (2000), 791–810. http://www.jstor.org/stable/2586896

[74] Dietz Graham and Den Hartog Deanne N. 2006. Measuring trust inside organisations. *Personnel Review* 35, 5 (01 Jan 2006), 557–588. https://doi.org/10.1108/00483480610682299

[75] Dara Gruber, Ashley Aune, and Wilma Koutstaal. 2018. Can Semi-Anthropomorphism Influence Trust and Compliance? Exploring Image Use in App Interfaces. In *Proceedings of the Technology, Mind, and Society (TechMindSociety '18)*. Association for Computing Machinery, New York, NY, USA, 6. https://doi.org/10.1145/3183654.3183700

[76] Jonathan Grudin. 2009. AI and HCI: Two Fields Divided by a Common Focus. *AI Magazine* 30, 4 (September 2009), 48–57. https://doi.org/10.1609/aimag.v30i4.2271

[77] Kunal Gupta, Ryo Hajika, Yun Suen Pai, Andreas Duenser, Martin Lochner, and Mark Billinghurst. 2019. In AI We Trust: Investigating the Relationship between Biosignals, Trust and Cognitive Load in VR. In *25th ACM Symposium on Virtual Reality Software and Technology (VRST '19)*. Association for Computing Machinery, New York, NY, USA, 10. https://doi.org/10.1145/3359996.3364276

[78] Andreas Gutscher. 2007. A Trust Model for an Open, Decentralized Reputation System. In *Trust Management*. Springer US, New Brunswick, Canada, 285–300. https://doi.org/10.1007/978-0-387-73655-6_19

[79] Özgür Gürerk, Andrea Bönsch, Lucas Braun, Christian Grund, Christine Harbring, Thomas Kittsteiner, and Andreas Staffeldt. 2014. Experimental Economics in Virtual Reality.

[80] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. de Visser, and Raja Parasuraman. 2011. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors* 53, 5 (2011), 517–527. https://doi.org/10.1177/0018720811417254

[81] Jason L. Harman, John O'Donovan, Tarek Abdelzaher, and Cleotilde Gonzalez. 2014. Dynamics of Human Trust in Recommender Systems. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*. Association for Computing Machinery, New York, NY, USA, 305–308. https://doi.org/10.1145/2645710.2645761

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

[83] 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

[84] Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. 2013. Presenting System Uncertainty in Automotive UIs for Supporting Trust Calibration in Autonomous Driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '13)*. Association for Computing Machinery, New York, NY, USA, 210–217. https://doi.org/10.1145/2516540.2516554

[85] Ralph Hertwig and Andreas Ortmann. 2003. *Economists' and Psychologists' Experimental Practices: How They Differ, Why They Differ, And How they Could Converge*. Vol. 1. Oxford University Press, Oxford, United Kingdom, Chapter 13, 253–272. https://books.google.fr/books?id=fOI31h_G6UkC&pg=PA260&lpg=PA260&dq=financial+incentives+and+trust+experiment&source=bl&ots=-CRrjQeHv_&sig=ACfU3U2ID0VJinKgmlUgpFsomoQMDO2GnQ&hl=en&sa=X&ved=2ahUKEwiA4O3C27XqAhVNOBoKHWGJBakQ6AEwDnoECAsQAQ#v=onepage&q=financial%20incentives%20and%20trust%20experiment&f=false

[86] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (2015), 407–434. https://doi.org/10.1177/0018720814547570

[87] Kai Holländer, Philipp Wintersberger, and Andreas Butz. 2019. Overtrust in External Cues of Automated Vehicles: An Experimental Investigation. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '19)*. Association for Computing Machinery, New York, NY, USA, 211–221. https://doi.org/10.1145/3342197.3344528

[88] John Holmes and John Rempel. 1985. Trust in Close Relationships. *Journal of Personality and Social Psychology* 49 (07 1985). https://doi.org/10.1037//0022-3514.49.1.95

[89] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proc. ACM Hum.-Comput. Interact.* 4, 1, Article 068 (May 2020), 26 pages. https://doi.org/10.1145/3392878

[90] 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

[91] Hsiao-Ying Huang and Masooda Bashir. 2017. Personal Influences on Dynamic Trust Formation in Human-Agent Interaction. In *Proceedings of the 5th International Conference on Human Agent Interaction (HAI '17)*. Association for Computing Machinery, New York, NY, USA, 233–243. https://doi.org/10.1145/3125739.3125749

[92] 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

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

[94] 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

[95] Joi Ito. 2018. What the Boston School Bus Schedule Can Teach Us About AI. https://www.wired.com/story/joi-ito-ai-and-bus-routes/

[96] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

[97] Zhuochen Jin, Shuyuan Cui, Shunan Guo, David Gotz, Jimeng Sun, and Nan Cao. 2020. CarePre: An Intelligent Clinical Decision Assistance System. *ACM Trans. Comput. Healthcare* 1, 1 (2020), 20. https://doi.org/10.1145/3344258

[98] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (01 Sep 2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[99] Noel D. Johnson and Alexandra A. Mislin. 2011. Trust games: A meta-analysis. *Journal of Economic Psychology* 32, 5 (June 2011), 865–889. https://doi.org/10.1016/j.joep.2011.05.00

[100] Angie M. Johnston, Candice M. Mills, and Asheley R. Landrum. 2015. How do children weigh competence and benevolence when deciding whom to trust? *Cognition* 144 (2015), 76 – 90. https://doi.org/10.1016/j.cognition.2015.07.015

[101] K. G. Jöreskog. 1967. A GENERAL APPROACH TO CONFIRMATORY MAXIMUM LIKELIHOOD FACTOR ANALYSIS. *ETS Research Bulletin Series* 1967, 2 (1967), 183–202. https://doi.org/10.1002/j.2333-8504.1967.tb00991.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1967.tb00991.x

[102] Daniel Kahneman. 2000. *Evaluation by Moments: Past and Future.* Cambridge University Press & Russell Sage Foundation, New York, USA, Chapter 38, 693–708. https://doi.org/10.1017/CBO9780511803475.039

[103] 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

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

[105] 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

[106] Melissa A. Koenig and Vikram K. Jaswal. 2011. Characterizing Children's Expectations About Expertise and Incompetence: Halo or Pitchfork Effects? *Child Development* 82, 5 (2011), 1634–1647. http://www.jstor.org/stable/41289869

[107] Agnieszka Kolasinska, Ivano Lauriola, and Giacomo Quadrio. 2019. Do People Believe in Artificial Intelligence? A Cross-Topic Multicultural Study. In *Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good (GoodTechs '19)*. Association for Computing Machinery, New York, NY, USA, 31–36. https://doi.org/10.1145/3342428.3342667

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

[109] Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. 2020. Effects of Proactive Dialogue Strategies on Human-Computer Trust. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 107–116. https://doi.org/10.1145/3340631.3394840

[110] Sari Kujala, Virpi Roto, Kaisa Väänänen, Evangelos Karapanos, and Arto Sinnelä. 2011. UX Curve: A method for evaluating long-term user experience. *Interact. Comput.* 23 (2011), 473–483. https://doi.org/10.1016/j.intcom.2011.06.005

[111] Philipp Kulms and Stefan Kopp. 2019. More Human-Likeness, More Trust? The Effect of Anthropomorphism on Self-Reported and Behavioral Trust in Continued and Interdependent Human-Agent Cooperation. In *Proceedings of Mensch Und Computer 2019 (MuC'19)*. Association for Computing Machinery, New York, NY, USA, 31–42. https://doi.org/10.1145/3340764.3340793

[112] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[113] Asheley R. Landrum, Candice M. Mills, and Angie M. Johnston. 2013. When do children trust the expert? Benevolence information influences children's trust more than expertise. *Developmental Science* 16, 4 (2013), 622–638. https://doi.org/10.1111/desc.12059 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/desc.12059

[114] 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

[115] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270. https://doi.org/10.1080/00140139208967392

[116] 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

[117] 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

[118] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Co-Design and Evaluation of an Intelligent Decision Support System for Stroke Rehabilitation Assessment. *Proc. ACM Hum.-Comput. Interact.* 4, 2, Article 156 (Oct. 2020), 27 pages. https://doi.org/10.1145/3415227

[119] 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

[120] 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

[121] 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

[122] Ian Li, Jodi Forlizzi, Anind Dey, and Sara Kiesler. 2007. My Agent as Myself or Another: Effects on Credibility and Listening to Advice. In *Proceedings of the 2007 Conference on Designing Pleasurable Products and Interfaces (DPPI '07)*. Association for Computing Machinery, New York, NY, USA, 194–208. https://doi.org/10.1145/1314161.1314179

[123] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376590

[124] 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

[125] James L. Loomis. 1959. Communication, the Development of Trust, and Cooperative Behavior. *Human Relations* 12, 4 (1959), 305–315. https://doi.org/10.1177/001872675901200402 arXiv:https://doi.org/10.1177/001872675901200402

[126] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

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

[128] Nikolas Luhmann. 2000. Familiarity, Confidence, Trust: Problems and Alternatives. In *Trust: Making and Breaking Cooperative Relations*, Diego Gambetta (Ed.). Basil Blackwell, Oxford, United Kingdom, 94–107.

[129] 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

[130] Maria A. Madsen and Shirley Gregor. 2000. Measuring Human-Computer Trust. In *Proceedings of the 11 th Australasian Conference on Information Systems*. Australasian Conference on Information Systems (ACIS), Brisbane, Australia, 6–8.

[131] Danielle Magaldi and Matthew Berler. 2020. *Semi-structured Interviews*. Springer International Publishing, Cham, 4825–4830. https://doi.org/10.1007/978-3-319-24612-3_857

[132] Mora Maldonado, Ewan Dunbar, and Emmanuel Chemla. 2019. Mouse tracking as a window into decision making. *Behavior Research Methods* 51, 3 (01 Jun 2019), 1085–1101. https://doi.org/10.3758/s13428-018-01194-x

[133] C. Mantzavinos. 2020. Hermeneutics. In *The Stanford Encyclopedia of Philosophy* (spring 2020 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University, Stanford, CA, USA.

[134] Ronald Marshall. 2003. Building trust early: The influence of first and second order expectations on trust in international channels of distribution. *International Business Review* 12 (08 2003), 421–443. https://doi.org/10.1016/S0969-5931(03)00037-4

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

[136] Steffen Maurer, Rainer Erbach, Issam Kraiem, Susanne Kuhnert, Petra Grimm, and Enrico Rukzio. 2018. Designing a Guardian Angel: Giving an Automated Vehicle the Possibility to Override Its Driver. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '18)*. Association for Computing Machinery, New York, NY, USA, 341–350. https://doi.org/10.1145/3239060.3239078

[137] 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

[138] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review* 20, 3 (1995), 709–734. http://www.jstor.org/stable/258792

[139] Daniel J. McAllister. 1995. Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations. *The Academy of Management Journal* 38, 1 (1995), 24–59. http://www.jstor.org/stable/256727

[140] Bill McEvily and Marco Tortoriello. 2011. Measuring trust in organisational research: Review and recommendations. *Journal of Trust Research* 1, 1 (2011), 23–63. https://doi.org/10.1080/21515581.2011.552424

[141] 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

[142] 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

[143] 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

[144] Stephanie M. Merritt. 2011. Affective Processes in Human–Automation Interactions. *Human Factors* 53, 4 (2011), 356–370. https://doi.org/10.1177/0018720811411912 arXiv:https://doi.org/10.1177/0018720811411912

[145] Joachim Meyer and John D. Lee. 2013. Trust, Reliance, and Compliance. In *The Oxford Handbook of Cognitive Engineering*, John D. Lee and Alex Kirlik (Eds.). Oxford University Press, Oxford, UK, 1–29. https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199757183.001.0001/oxfordhb-9780199757183-e-6

[146] Microsoft. 2018. *Responsible bots: 10 guidelines for developers of conversational AI.* Technical Report. Microsoft, USA. 5 pages. https://www.microsoft.com/en-us/research/publication/responsible-bots/

[147] Michael Moore. 2012. *Confirmatory factor analysis.* The Guilford Press, NY, USA, 361–379.

[148] Drew M. Morris, Jason M. Erno, and June J. Pilcher. 2017. Electrodermal Response and Automation Trust during Simulated Self-Driving Car Use. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61, 1 (2017), 1759–1762. https://doi.org/10.1177/1541931213601921 arXiv:https://doi.org/10.1177/1541931213601921

[149] B.M. Muir. 1989. *Operators' Trust in and Use of Automatic Controllers in a Supervisory Process Control Task.* University of Toronto, Toronto, Canada. https://books.google.fr/books?id=T94NSwAACAAJ

[150] Lea S. Müller, Sarah M. Meeßen, Meinald T. Thielsch, Christoph Nohe, Dennis M. Riehle, and Guido Hertel. 2020. *Do Not Disturb! Trust in Decision Support Systems Improves Work Outcomes under Certain Conditions.* Association for Computing Machinery, New York, NY, USA, 229–237. https://doi.org/10.1145/3404983.3405515

[151] Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. 2019. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 119 (Nov. 2019), 36 pages. https://doi.org/10.1145/3359221

[152] 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.

[153] Michael Naef and Jürgen Schupp. 2009. *Measuring Trust: Experiments and Surveys in Contrast and Combination.* IZA Discussion Papers 4087. Institute of Labor Economics (IZA). https://ideas.repec.org/p/iza/izadps/dp4087.html

[154] Manisha Natarajan and Matthew Gombolay. 2020. Effects of Anthropomorphism and Accountability on Trust in Human Robot Interaction. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*. Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/3319502.3374839

[155] Domen Novak. 2014. *Engineering Issues in Physiological Computing.* 17–38. https://doi.org/10.1007/978-1-4471-6392-3_2

[156] Kenya Freeman Oduor and Christopher S. Campbell. 2007. Deciding When to Trust Automation in a Policy-Based City Management Game: Policity. In *Proceedings of the 2007 Symposium on Computer Human Interaction for the Management of Information Technology (CHIMIT '07)*. Association for Computing Machinery, New York, NY, USA, 2–es. https://doi.org/10.1145/1234772.1234775

[157] Institute of Business Ethics. 2018. *Business Ethics and Artificial Intelligence.* Technical Report. Internet Society, London, UK. https://www.ibe.org.uk/resource/ibe-briefing-58-business-ethics-and-artificial-intelligence-pdf.html

[158] Royal College of Physicians. 2018. *Artificial intelligence (AI) in health.* Technical Report. Royal College of Physicians, London, UK. 1 pages. https://www.rcplondon.ac.uk/projects/outputs/artificial-intelligence-ai-health

[159] 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/

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

[161] Roobina Ohanian. 1990. Construction and Validation of a Scale to Measure Celebrity Endorsers' Perceived Expertise, Trustworthiness, and Attractiveness. *Journal of Advertising* 19, 3 (oct 1990), 39–52. https://doi.org/10.1080/00913367.1990.10673191

[162] Special Interest Group on Artificial Intelligence. 2019. *Dutch Artificial Intelligence Manifesto.* Technical Report. ICT Research Platform Nederland, The Netherlands. 15 pages. http://ii.tudelft.nl/bnvki/wp-content/uploads/2018/09/Dutch-AI-Manifesto.pdf

[163] Tobias O.Nyumba, Kerrie Wilson, Christina J. Derrick, and Nibedita Mukherjee. 2018. The use of focus group discussion methodology: Insights from two decades of application in conservation. *Methods in Ecology and Evolution* 9, 1 (2018), 20–32. https://doi.org/10.1111/2041-210X.12860 arXiv:https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12860

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

[165] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. In *Proceedings of the 2019 Conference on Computer Supported Cooperative Work (CSCW '19)*, Vol. 3. Association for Computing Machinery, New York, NY, USA, 15. https://doi.org/10.1145/3359204

[166] Dhaval Parmar, Stefán Ólafsson, Dina Utami, Prasanth Murali, and Timothy Bickmore. 2020. *Navigating the Combinatorics of Virtual Agent Design Space to Maximize Persuasion.* International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1010–1018.

[167] 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

[168] P. Ivan Pavlov. 2010. Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex. *Annals of neurosciences* 17, 3 (Jul 2010), 136–141. https://doi.org/10.5214/ans.0972-7531.1017309 25205891[pmid].

[169] Carl J. Pearson, Allaire K. Welk, William A. Boettcher, Roger C. Mayer, Sean Streck, Joseph M. Simons-Rudolph, and Christopher B. Mayhorn. 2016. Differences in Trust between Human and Automated Decision Aids. In *Proceedings of the Symposium and Bootcamp on the Science of Security (HotSos '16).* Association for Computing Machinery, New York, NY, USA, 95–98. https://doi.org/10.1145/2898375.2898385

[170] Brandon S. Perelman, Arthur W. Evans III, and Kristin E. Schaefer. 2020. Where Do You Think You're Going? Characterizing Spatial Mental Models from Planned Routes. *J. Hum.-Robot Interact.* 9, 4, Article 23 (May 2020), 55 pages. https://doi.org/10.1145/3385008

[171] 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

[172] J. Paul Peter. 1979. Reliability: A Review of Psychometric Basics and Recent Marketing Practices. *Journal of Marketing Research* 16, 1 (1979), 6–17. http://www.jstor.org/stable/3150868

[173] Gjalt-Jorn Peters. 2014. The alpha and the omega of scale reliability and validity: Why and how to abandon Conbach's alpha and the route towards more comprehensive assessment od scale quality. *Euro Health Psychologist* 16 (01 2014), 56–69.

[174] Jonathan A. Plucker. 2003. Exploratory and Confirmatory Factor Analysis in Gifted Education: Examples with Self-Concept Data. *Journal for the Education of the Gifted* 27, 1 (2003), 20–35. https://doi.org/10.1177/016235320302700103 arXiv:https://doi.org/10.1177/016235320302700103

[175] J. Potter and D. Edwards. 1996. *Discourse Analysis.* Macmillan Education UK, London, 419–425. https://doi.org/10.1007/978-1-349-24483-6_63

[176] Pearl Pu and Li Chen. 2006. Trust Building with Explanation Interfaces. In *Proceedings of the 11th International Conference on Intelligent User Interfaces (IUI '06).* Association for Computing Machinery, New York, NY, USA, 93–100. https://doi.org/10.1145/1111449.1111475

[177] David V. Pynadath, Ning Wang, Ericka Rovira, and Michael J. Barnes. 2018. Clustering Behavior to Recognize Subjective Beliefs in Human-Agent Teams. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '18).* International Foundation for Autonomous Agents and Multiagent Systems, New York, NY, USA, 1495–1503.

[178] Bako Rajaonah, Franoise Anceaux, Nicolas Tricot, and Marie-Pierre Pacaux-Lemoine. 2006. Trust, Cognitive Control, and Control: The Case of Drivers Using an Auto-Adaptive Cruise Control. In *Proceedings of the 13th Eurpoean Conference on Cognitive Ergonomics: Trust and Control in Complex Socio-Technical Systems (ECCE '06).* Association for Computing Machinery, New York, NY, USA, 17–24. https://doi.org/10.1145/1274892.1274896

[179] Bako Rajaonah, Françoise Anceaux, and Fabrice Vienne. 2006. Study of driver trust during cooperation with adaptive cruise control. *Le travail humain* 69, 2 (2006), 99–127. https://doi.org/10.3917/th.692.0099

[180] Samantha Reig, Selena Norman, Cecilia G. Morales, Samadrita Das, Aaron Steinfeld, and Jodi Forlizzi. 2018. A Field Study of Pedestrians and Autonomous Vehicles. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '18)*. Association for Computing Machinery, New York, NY, USA, 198–209. https://doi.org/10.1145/3239060.3239064

[181] Robin M. Richter, Maria Jose Valladares, and Steven C. Sutherland. 2019. Effects of the Source of Advice and Decision Task on Decisions to Request Expert Advice. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 469–475. https://doi.org/10.1145/3301275.3302279

[182] Loo Robert. 2002. A caveat on using single-item versus multiple-item scales. *Journal of Managerial Psychology* 17, 1 (01 Jan 2002), 68–75. https://doi.org/10.1108/02683940210415933

[183] 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

[184] Jr Robert B. Lount, Chen-Bo Zhong, Niro Sivanathan, and J. Keith Murnighan. 2008. Getting Off on the Wrong Foot: The Timing of a Breach and the Restoration of Trust. *Personality and Social Psychology Bulletin* 34, 12 (2008), 1601–1612. https://doi.org/10.1177/0146167208324512 arXiv:https://doi.org/10.1177/0146167208324512 PMID: 19050335.

[185] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. 2016. Overtrust of Robots in Emergency Evacuation Scenarios. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI '16)*. IEEE Press, New York, NY, USA, 101–108.

[186] Mark A. Robinson. 2018. Using multi-item psychometric scales for research and practice in human resource management. *Human Resource Management* 57, 3 (2018), 739–750. https://doi.org/10.1002/hrm.21852 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/hrm.21852

[187] John T. Roscoe. 1969. *Fundamental research statistics for the behavioral sciences.* New York Holt, Rinehart and Winston. http://openlibrary.org/books/OL5685768M

[188] Elizabeth Rosenzweig. 2015. Usability Testing. In *Successful User Experience: Strategies and Roadmaps*, Elizabeth Rosenzweig (Ed.). Morgan Kaufmann, Boston, MA, USA, Chapter 7, 131 – 154. https://doi.org/10.1016/B978-0-12-800985-7.00007-7

[189] Casey Ross and Ike Swetlitz. 2017. IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. https://www.statnews.com/2017/09/05/watson-ibm-cancer/

[190] Jennifer M. Ross. 2008. *Moderators of trust and reliance across multiple decision aids.* Ph.D. Dissertation. Department of Psychology in the College of Sciences at the University of Central Florida.

[191] 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

[192] Denise Rousseau, Sim Sitkin, Ronald Burt, and Colin Camerer. 1998. Not So Different After All: A Cross-discipline View of Trust. *Academy of Management Review* 23 (July 1998). https://doi.org/10.5465/AMR.1998.926617

[193] Nicole Salomons, Michael van der Linden, Sarah Strohkorb Sebo, and Brian Scassellati. 2018. Humans Conform to Robots: Disambiguating Trust, Truth, and Conformity. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. Association for Computing Machinery, New York, NY, USA, 187–195. https://doi.org/10.1145/3171221.3171282

[194] Willem E. Saris and Irmtraud N. Gallhofer. 2007. *Criteria for the Quality of Survey Measures.* John Wiley & Sons, Ltd, Hoboken, New Jersey, USA, 173–217. https://doi.org/10.1002/9780470165195.ch9 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470165195.ch9

[195] Kristin E. Schaefer. 2013. *The Perception And Measurement Of Human-robot Trust.* Ph.D. Dissertation. Department of Psychology in the College of Sciences at the University of Central Florida.

[196] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I Can Do Better than Your AI: Expertise and Explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 240–251. https://doi.org/10.1145/3301275.3302308

[197] Hanna Schneider, Julia Wayrauther, Mariam Hassib, and Andreas Butz. 2019. Communicating Uncertainty in Fertility Prognosis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300391

[198] Haeseung Seo, Aiping Xiong, and Dongwon Lee. 2019. Trust It or Not: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation. In *Proceedings of the 10th ACM Conference on Web Science (WebSci '19)*. Association for Computing Machinery, New York, NY, USA, 265–274. https://doi.org/10.1145/3292522.3326012

[199] 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

[200] Fred Shaffer and J. P. Ginsberg. 2017. An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in public health* 5 (28 Sep 2017), 258–258. https://doi.org/10.3389/fpubh.2017.00258 29034226[pmid].

[201] Ameneh Shamekhi, Q. Vera Liao, Dakuo Wang, Rachel K. E. Bellamy, and Thomas Erickson. 2018. Face Value? Exploring the Effects of Embodiment for a Group Facilitation Agent. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173965

[202] Klaas Sijtsma. 2008. On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika* 74, 1 (11 Dec 2008), 107. https://doi.org/10.1007/s11336-008-9101-0

[203] 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

[204] Internet Society. 2017. *Artificial intelligence and machine learning: policy paper.* Technical Report. Internet Society, Reston, Virginia, United States. https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/

[205] 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/

[206] Donna Spencer and Todd Warfel. 2004. Card sorting: a definitive guide. https://boxesandarrows.com/card-sorting-a-definitive-guide/

[207] Nicole Sultanum, Michael Brudno, Daniel Wigdor, and Fanny Chevalier. 2018. More Text Please! Understanding and Supporting the Use of Visualization for Clinical Text Overview. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173996

[208] Haoye Sun, Willem J. M. I. Verbeke, Rumen Pozharliev, Richard P. Bagozzi, Fabio Babiloni, and Lei Wang. 2019. Framing a trust game as a power game greatly affects interbrain synchronicity between trustor and trustee. *Social Neuroscience* 14, 6 (Dec. 2019), 635–648. https://doi.org/10.1080/17470919.2019.1566171

[209] Harini Suresh, Natalie Lao, and Ilaria Liccardi. 2020. Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making. In *12th ACM Conference on Web Science* (Southampton, United Kingdom) *(WebSci '20)*. Association for Computing Machinery, New York, NY, USA, 315–324. https://doi.org/10.1145/3394231.3397922

[210] Steven C. Sutherland, Casper Harteveld, and Michael E. Young. 2015. The Role of Environmental Predictability and Costs in Relying on Automation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2535–2544. https://doi.org/10.1145/2702123.2702609

[211] Richard S. Sutton and Andrew G. Barto. 1998. *Introduction to Reinforcement Learning* (1st ed.). MIT Press, Cambridge, MA, USA.

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

[213] 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

[214] Hiroyuki Tokushige, Takuji Narumi, Sayaka Ono, Yoshitaka Fuwamoto, Tomohiro Tanikawa, and Michitaka Hirose. 2017. Trust Lengthens Decision Time on Unexpected Recommendations in Human-Agent Interaction. In *Proceedings of the 5th International Conference on Human Agent Interaction (HAI '17)*. Association for Computing Machinery, New York, NY, USA, 245–252. https://doi.org/10.1145/3125739.3125751

[215] Ilaria Torre, Emma Carrigan, Rachel McDonnell, Katarina Domijan, Killian McCabe, and Naomi Harte. 2019. The Effect of Multimodal Emotional Expression and Agent Appearance on Trust in Human-Agent Interaction. In *Motion, Interaction and Games (MIG '19)*. Association for Computing Machinery, New York, NY, USA, 6. https://doi.org/10.1145/3359566.3360065

[216] Ilaria Torre, Jeremy Goslin, Laurence White, and Debora Zanatto. 2018. Trust in Artificial Voices: A "Congruency Effect" of First Impressions and Behavioural Experience. In *Proceedings of the Technology, Mind, and Society (TechMindSociety '18)*. Association for Computing Machinery, New York, NY, USA, 6. https://doi.org/10.1145/3183654.3183691

[217] Italo Trizano-Hermosilla and Jesús M. Alvarado. 2016. Best Alternatives to Cronbach's Alpha Reliability in Realistic Conditions: Congeneric and Asymmetrical Measurements. *Frontiers in Psychology* 7 (2016), 769. https://doi.org/10.3389/fpsyg.2016.00769

[218] UNI Global Union. 2017. *10 Principles for Ethical AI.* Technical Report. UNI Global Union, Nyon, Switzerland. 10 pages. http://www.thefutureworldofwork.org/opinions/10-principles-for-ethical-ai/

[219] Hanneke Hooft van Huysduynen, Jacques Terken, and Berry Eggen. 2018. Why Disable the Autopilot?. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '18)*. Association for Computing Machinery, New York, NY, USA, 247–257. https://doi.org/10.1145/3239060.3239063

[220] Peter-Paul van Maanen, Francien Wisse, Jurriaan van Diggelen, and Robbert-Jan Beun. 2011. Effects of Reliance Support on Team Performance by Advising and Adaptive Autonomy. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 02 (WI-IAT '11)*. IEEE Computer Society, New York, NY, USA, 280–287. https://doi.org/10.1109/WI-IAT.2011.117

[221] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174014

[222] 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.

[223] Rudolf von Sinner. 2005. Trust and Convivência. *The Ecumenical Review* 57, 3 (2005), 322–341. https://doi.org/10.1111/j.1758-6623.2005.tb00554.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1758-6623.2005.tb00554.x

[224] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[225] Lin Wang, Pei-Luen Patrick Rau, Vanessa Evers, Benjamin Krisper Robinson, and Pamela Hinds. 2010. When in Rome: The Role of Culture & Context in Adherence to Robot Recommendations. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI '10)*. IEEE Press, New York, NY, USA, 359–366.

[226] M. Wang, A. Hussein, R. F. Rojas, K. Shafi, and H. A. Abbass. 2018. EEG-Based Neural Correlates of Trust in Human-Autonomy Interaction. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, Bangalore, India, 350–357.

[227] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. The Impact of POMDP-Generated Explanations on Trust and Performance in Human-Robot Teams. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, New York, NY, USA, 997–1005.

[228] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. Trust Calibration within a Human-Robot Team: Comparing Automatically Generated Explanations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI '16)*. IEEE Press, New York, NY, USA, 109–116.

[229] 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.

[230] Lawrence R. Wheeless and Janis Grotz. 1977. The measurement of trust and its relationship to self-disclosure. *Human Communication Research* 3, 3 (1977), 250–257. https://doi.org/10.1111/j.1468-2958.1977.tb00523.x

[231] T. Whelan. 2008. Social Presence in Multi-User Virtual Environments : A Review and Measurement Framework for Organizational Research.

[232] Philipp Wintersberger, Tamara von Sawitzky, Anna-Katharina Frison, and Andreas Riener. 2017. Traffic Augmentation as a Means to Increase Trust in Automated Driving Systems. In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter (CHItaly '17)*. Association for Computing Machinery, New York, NY, USA, 7. https://doi.org/10.1145/3125571.3125600

[233] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174230

[234] Jun Xiao, John Stasko, and Richard Catrambone. 2007. The Role of Choice and Customization on Users' Interaction with Embodied Conversational Agents: Effects on Perception and Performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1293–1302. https://doi.org/10.1145/1240624.1240820

[235] Yaqi Xie, Indu P Bodala, Desmond C. Ong, David Hsu, and Harold Soh. 2019. Robot Capability and Intention in Trust-Based Decisions across Tasks. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI '19)*. IEEE Press, New York, NY, USA, 39–47.

[236] 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

[237] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. 2020. How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. Association for Computing Machinery, New York, NY, USA, 189–201. https://doi.org/10.1145/3377325.3377480

[238] 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

[239] X. Jessie Yang, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah. 2017. Evaluating Effects of User Experience and System Transparency on Trust in Automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)*. Association for Computing Machinery, New York, NY, USA, 408–416. https://doi.org/10.1145/2909824.3020230

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

[241] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300509

[242] Louise C. Young and Gerald S. Albaum. 2002. *Developing a measure of trust in retail relationships : a direct selling application*. School of Marketing, University of Technology of Sydney, Sydney Broadway, N.S.W, Australia.

[243] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. *Keeping Designers in the Loop: Communicating Inherent Algorithmic Trade-Offs Across Multiple Objectives*. Association for Computing Machinery, New York, NY, USA, 1245–1257. https://doi.org/10.1145/3357236.3395528

[244] Kun Yu, Shlomo Berkovsky, Dan Conway, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2016. Trust and Reliance Based on System Accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (UMAP '16)*. Association for Computing Machinery, New York, NY, USA, 223–227. https://doi.org/10.1145/2930238.2930290

[245] 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

[246] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I Trust My Machine Teammate? An Investigation from Perception to Decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 460–468. https://doi.org/10.1145/3301275.3302277

[247] Beste F. Yuksel, Penny Collisson, and Mary Czerwinski. 2017. Brains or Beauty: How to Engender Trust in User-Agent Interactions. *ACM Trans. Internet Technol.* 17, 1 (2017), 20. https://doi.org/10.1145/2998572

[248] 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

Received January 2021 ; revised April 2021 ; accepted May 2021