Trust in AI-assisted Decision Making: Perspectives from Those Behind the System and Those for Whom the Decision is Made

Oleksandra Vereschak, Fatemeh Alizadeh, Gilles Bailly, Baptiste Caramiaux

Trust in AI-assisted Decision Making: Perspectives from Those Behind the System and Those for Whom the Decision is Made

Oleksandra Vereschak
Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR
Paris, France
vereschak@isir.upmc.fr

Fatemeh Alizadeh
University of Siegen
Siegen, Germany
fatemeh.alizadeh@uni-siegen.de

Gilles Bailly*
Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR
Paris, France
gilles.bailly@sorbonne-universite.fr

Baptiste Caramiaux*
Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR
Paris, France
baptiste.caramiaux@sorbonne-universite.fr

Abstract

Trust between humans and AI in the context of decision-making has acquired an important role in public policy, research and industry. In this context, Human-AI Trust has often been tackled from the lens of cognitive science and psychology, but lacks insights from the stakeholders involved. In this paper, we conducted semi-structured interviews with 7 AI practitioners and 7 decision subjects from various decision domains. We found that 1) interviewees identified the prerequisites for the existence of trust and distinguish trust from trustworthiness, reliance, and compliance; 2) trust in AI-integrated systems is strongly influenced by other human actors, more than the system's features; 3) the role of Human-AI trust factors is stakeholder-dependent. These results provide clues for the design of Human-AI interactions in which trust plays a major role, as well as outline new research directions in Human-AI Trust.

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

Keywords: trust, artificial intelligence, decision making, qualitative study, AI practitioners, decision subjects

1 Introduction

Decision making assisted by artificial intelligence (AI) has become more widespread in high-stakes domains, where decisions have real impacts on people's lives such as public safety [47], hiring [3] or loan approval [75]. Typically, the AI-based systems considered are based on automated processes (such as data-driven machine learning techniques) that provide assistance to human decision makers in a form of recommendations. Because Human-AI trust plays an important role in the adoption of these technologies [41] and the improvement of decision making [9], it has become a priority for their design and development, as well deployment and regulation [73]. To understand how to achieve appropriate levels of human trust in these systems, more research at the intersection of Human-Computer Interaction and social study of AI is needed.

Trust is a complex and multifaceted concept [56, 61] and several studies have focused on a better understanding of the factors that can affect Human-AI trust (e.g., [64, 74, 106, 112, 116]). In these studies, trust is predominantly investigated through the lens of users, who are the persons interacting with the AI-assisted decision making system and its recommendations in order to deliver their decision [54]. Less is known about the perspectives from which other stakeholders involved in, and impacted by the design, deployment and use of these systems, view the notion of trust in AI, while this outlook on Human-AI trust can be shaped by their role. Jakesch et al. [46] demonstrate that the ethical values embedded in AI-assisted decision making systems can hold varied significance and interpretations for different groups. For example, people working on AI, on average, considered responsible AI values less important than general public and crowdworkers that contributed to the training of AI models. Such differences might also be reflected in the understanding of and opinions on Human-AI trust of the various stakeholders. For example, Lockey et al. [60] identify that different types of users do not encounter the same issues related to Human-AI trust: trust in AI of domain experts, e.g., doctors in medical decision-making, might be particularly affected by the factors that challenge their professional knowledge, skills, identity, and reputation. In contrast, fairness-related factors might impact general users' and society's trust in AI. Therefore, examining how stakeholders other than AI users view the definitions and factors of Human-AI is essential to advance the understanding of how trust is accounted for in

*Both authors contributed equally to this research.

the development and design of AI-embedded systems assisting decision-making and whether the existing approaches match the varying needs of different stakeholders.

In this article, we investigate how two groups of stakeholders - AI practitioners and decision subjects - understand Human-AI trust. Exploring the views of these groups on Human-AI trust definitions and factors allows to understand to which extent they prioritize and value the same aspects of Human-AI trust. **AI practitioners** are involved in system design and deployment (from AI developers to project managers). Given that they make decisions that influence the shape of human-AI interaction, impacting trust in AI-based technology and its acceptance for decision-making, understanding AI practitioners' views on trust can shed light on the factors they prioritize to build trust in AI among different stakeholders. Therefore, we explore the following first two research questions: RQ1a) *According to AI practitioners, what are the critical elements of human-AI trust in decision-making?*; RQ1b) *What do AI practitioners think influences the trust of various stakeholders in AI in the context of decision-making?*

The second group is **decision subjects**, i.e., people who do not interact directly with the systems incorporating AI but are affected by the decisions made by users based on the recommendations of these systems. For example, doctors are users, and patients are decision subjects in the medical context. Although decision subjects do not generally interact with AI-based systems the same way as users, they may nevertheless want to decide whether or not they wish to be impacted by the system [36]. For example, if a patient decides that the doctor's AI-based recommendation is not fair or trustworthy, they may want to change doctors or clinics. Therefore, we are investigating the following research questions: RQ2a) *According to decision subjects, what are the critical elements of trust between humans and AI in decision-making?*; RQ2b) *What factors influence decision subjects' trust in AI?*

We thus conducted semi-structured interviews with 7 AI practitioners related to AI-assisted decision making and 7 decision subjects from various risk-sensitive contexts (finance, law, management, medicine). The questions revolved around defining trust and trustworthiness when related to AI, and what they think can affect Human-AI trust. Using thematic analysis [14, 24], we established three themes: 1) definition of trust through three prerequisite elements and differentiation from other related concepts. The interviewees define Human-AI trust similarly to the literature with vulnerability and positive expectations, and additionally propose task complexity as a trust prerequisite. Moreover, AI practitioners distinguish trust from trustworthiness and trust-related behaviors such as reliance and compliance; 2) the effect of relationships between various stakeholders on Human-AI trust. We found that the extent to which decision subjects, AI practitioners, and users trust each other has an impact on their trust towards AI and can moderate the effect of some factors on Human-AI trust; 3) stakeholder-dependency of the role and effects of some factors on Human-AI trust. We found that AI transparency, AI literacy, and interactivity of the system affect Human-AI trust differently for different stakeholders. Based on our findings, we provide a set of implications for academic researchers in HCI and AI practitioners. In particular, we recommend investigating the breaking and calibration points of trust between humans and AI beyond direct interaction with the system, and re-examining the techno-centric trust factors between humans and AI from a socio-technical point of view, as well as from the point of view of stakeholders other than users.

2 Related Work

In this section, we provide a brief overview of the methods to study Human-AI trust in assisted decision making and the different stakeholders at play with such systems.

2.1 Background on AI-embedded Systems Assisting Decision Making

While there is no universally accepted definition of AI [30], in this paper, we follow the definition provided by the European Commission: AI is a system capable of "*perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal*" [81]. Therefore, what we refer to as "AI-embedded systems assisting decision making" are the systems that analyse data to derive information used to facilitate human decision making [23, 88]. Usually, such systems provide assistance to human decision makers in a form of one or multiple recommendations, and when the system is not fully automated, it is the human who has the last word while making decisions. If the AI's recommendation differs from the decision maker's initial opinion, the decision maker finds themselves in conflict between her initial opinion and the new information received, which means that she has to choose between their opinion and the recommendation in order to make a better decision [110].

Making a better decision based on a recommendation means to be able to interpret the quality of the recommendation. However, it can sometimes be difficult to understand how a system arrived to a certain conclusion due to their "black box" nature [1, 53]. This, in turn, obfuscates understanding why a certain AI recommendation was produced, anticipating potential biases in decision making, and identifying the reasons for wrong predictions [87, 113]. When one is uncertain about how to correctly assess the quality of a recommendation [95], one can rely on their level of **trust** towards it to decide whether to stick to one's own opinion or to follow the system [100, 105]. As AI-embedded systems are becoming more widespread for assisting in making decisions have real impacts on people's lives, such as public safety [47], hiring [3] or loan approval [75], to name a few, the need for considering what contributes to human trust in the design of AI has arisen [13, 26, 34, 35, 51, 71, 92, 97, 104].

2.2 Human-AI Trust

Human-AI trust literature has two major themes of interest: defining what trust is and what factors affect it. The first line of research builds primarily on theoretical works, e.g. [45], taking a top-down approach to understanding Human-AI trust. A systematic literature review on Human-AI trust in the decision-making [103] defines Human-AI trust through three prerequisite elements, all encompassed in the trust **definition** by Lee and See [55]: *An attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability.* These three prerequisites are: vulnerability (or risk) of humans to the actions of the AI-based system, positive expectations of humans with respect to the AI-based system outcomes, and attitude as opposed to a behavior. Some scholars further define more granular facets of trust such as affective and cognitive trust [57, 70], weak and strong trust [10], warranted and unwarranted trust [45] or differentiate between trust in a particular AI tool, in people who built this tool and in AI in general [80].

The second major theme investigates Human-AI trust through a bottom-up approach, empirically studying what **factors** can affect users' trust. Glikson and Woolley [37] in a literature review of studies empirically investigating Human-AI trust factors identify that the main ones for trust in AI are: tangibility, transparency, performance (reliability), task characteristics, anthropomorphism, and socially-oriented behaviors of the system. While they did not propose any classification of the factors, almost all of them belong to a category related to the system characteristics (a category present in trust frameworks from the fields other than Human-AI interaction [2, 11, 40, 41, 84, 85]). Type of task is the only trust factor that is related to the context of interaction, rather than the interaction with the system itself. Another framework on Human-AI trust factors in the medical context [17] calls for expanding the current literature's focus on the contexts other than users' interaction with the system. Browne et al. [17] argue that considering trust factors in the contexts beyond use reflects better the entire clinical AI deployment process in the real settings and, thus, opens up the floor to new trust calibration points. As most of the work on Human-AI trust targets a single type of stakeholder - direct users of the systems, we expand the analysis of Human-AI trust definition and factors to the stakeholders other than users that are related to the Human-AI decision making systems.

2.3 Human-AI Trust and Stakeholders Other Than Users

In this article, we focused on stakeholders that are the most linked to the development or the use of AI-assisted decision making systems: **AI practitioners**, people who develop the systems; **users**, people who use these systems to make decisions; and **decision subjects**, people who are affected by those decisions (see Table 1). Additional stakeholders, however, exist, such as regulators and policy makers, whose contributions, although interesting, are out of the scope of this paper (the reader can refer to different taxonomies [8, 29, 39, 46, 90, 115] for more information).

The stakeholders that have received the most attention in the literature on Human-AI trust are the **users** of the systems [54]. Researchers have repeatedly pointed to the need to explore and assess users' trust in these systems to facilitate their adoption (see, for instance, [12, 89, 91]). It is not surprising that the literature focuses on system users, as understanding what affects their trust in the AI algorithms embedded in these systems can inform the development of interfaces and interactions that would facilitate the emergence of trust. However, different stakeholders may have different needs, expectations or roles when it comes to trust between humans and AI, and may also have an implicit impact on users' trust in systems. The research on AI with human-centered values has investigated the perspectives and needs stakeholders other than users, notably AI practitioners (e.g. [7, 28, 49, 101, 102, 111]) and decision subjects (e.g. [36, 59, 62, 63, 68, 114]). Here we first present a set of previous works that have explicitly demonstrated differences between these stakeholders when it comes to concepts such as AI ethics, explanations, or fairness. While these results are not directly about Human-AI trust, they are related to our domain and motivate our approach.

Regarding responsible AI, which aims to deploy AI-based systems in line with ethical and legal frameworks, previous work shows the importance of including the AI practitioners' perspectives to ensure that the system is designed to meet the actual needs of business and industry [42]. Typically, *AI practitioners* are pushed to quickly develop a service or a product that one can sell, which sometimes conflicts with ethical practices valued by the *users* [4, 65, 78, 107]. Regarding Explainable AI (XAI), which aims to propose means to explain AI-based predictions and help their interpretation, the usefulness of the explanation of a recommendation given by AI can vary depending on who sees it [32, 88]. A *user* might want to learn to which extent a recommendation can help them save money for instance [67], while *decision subjects* might want to know to which extent this recommendation is biased against a certain population in which they may belong [16, 108]. Regarding fairness, Smith et al. [94] take the case of microlending and show that depending on the different strategies to achieve fairness, stemming from

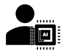
Icon	Acronym	Stakeholder	Definition
	U	Users	Individuals directly interacting with the system
	P	AI practitioners	Individuals who design, develop and deploy AI-based solutions
	DS	Decision subjects	Individuals affected by an AI-assisted decision-making system

Table 1. The different stakeholders related to the AI assisted decision making systems. This article focuses on AI practitioners and Decision subjects, two stakeholders who received less attention in the Human-AI trust literature.

Id	Role	Background	Organization	Type of AI	AI Application
P1	XAI R&D	CS and Maths	Large	CNNs	Transport, paleontology
P2	XAI R&D	Eng. and Maths	Small	OR	Task planning
P3	CEO	Maths	Small	Supervised ML	Evaluation of law cases
P4	Research mgr.	HCI	Large	OR, supervised and unsupervised ML	Project-based
P5	Research mgr.	Human Factors	Large	Not specified	Project-based
P6	CPO	Engineering	Small	ML (not specified)	Finance and business
P7	CEO	Bio. Eng. & Research	Small	Deep learning	Medical

Table 2. Characterization of AI practitioners, their companies, and AI they work with as reported by the interviewees themselves. “Small” refers to the companies with less than 20 employees, “Large” - with over 1000 employees. Explanation for abbreviations: *XAI* - explainable AI, *R&D* - research and development, *mgr.* - manager, *CEO* - chief executive officer, *CPO* - chief product officer, *CS* - computer science, *eng.* - engineering, *CNNs* - convolutional neural networks, *OR* - operations research, *ML* - machine learning.

its different definitions, Human-AI decisions favor *decision subjects*, *direct users* or *the organization behind the system.* Finally, regarding power relations in interaction, *users* and *AI practitioners* might see AI recommendations as tools, assistants or servants [50], while *decision subjects* might see the same AI recommendations as coming from someone in a more powerful position than they are. Such difference in perceived hierarchical roles between different stakeholders and AI can influence their attitude towards the system and interaction with it [21, 43, 83].

Previous work thus demonstrates the importance to study different stakeholders in the context of Human-AI interaction. In the context of Human-AI trust, Passi and Jackson [76] investigate how AI practitioners establish trust among themselves while working with data. Ammitzbøll Flügge et al. [5] and Okolo et al. [72] emphasize the importance of trust between users and decision subjects. Ferrario and Loi [32] analyze the importance of XAI for decision subjects' trust in AI. Lastly, Ramesh et al. [79] show that decision subjects overtrust AI due to seeing it as a higher authority for financial decisions. These works tend to focus on a small set of factors influencing Human-AI trust. A more global perspective of how AI practitioners and decision subjects build and perceive Human-AI trust is yet to be explored.

3 Methodology

We adopted an interview-based qualitative methodology to answer our research questions about what trust is and what that trust depends on in the context of AI-assisted decision-making from the perspective of the real-word stakeholders. The project started in 2021.

3.1 Participants

We recruited participants through a convenience sampling technique combined with snowballing among colleagues and friends, and through announcements at events and on the project's social media channels. We had two selection criteria to find interview participants: 1) they either work (as practitioners) on AI-embedded systems that support risk-sensitive decision making (e.g., in health, law, finance)[1] or they have been affected by their decisions (as decision subjects), 2) the system is used in the real world. We did not focus on any particular corporate position nor on any specific AI application in order to obtain a diversity of perspectives among interviewees. In total, we conducted 14 semi-structured interviews (7 with AI practitioners[2], 7 with AI decision subjects).

[1]Risk in risk-sensitive applications is understood as defined by the European Union (EU) regulatory framework proposal on AI: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[2]We initially contacted 14 AI practitioners, 5 of them did not reply, and 2 did not have availability for an interview

Id	Background	Decision Context
DS1	Software developer	Job application
DS2	Medical student	Access to services
DS3	Mechanical engineer	Job application
DS4	Business economics researcher	Loan application
DS5	Mechanical engineer	Job application
DS6	Accounting and project management	Job application
DS7	Computer engineer	Job application

Table 3. Characterization of decision subjects, notably their background and in what context they received a Human-AI decision.

The participation in the study was on a voluntary basis. The AI practitioners are based in Europe and Oceania, and each worked for a different company. Table 2 provides an overview of the AI practitioners' backgrounds, their roles in the company, and the application areas of AI. Three participants work on XAI (two are responsible for implementation and research, and another is the company's chief executive officer - CEO). Three other participants are senior project and product managers. The AI decision subjects are all based in Europe and had been affected by AI decision making in three different risk-sensitive areas: job application, access to services, loan application. The decision subjects we interviewed were not the people affected by the AI tools developed by the AI practitioners who participated in our study. Table 3 provides an overview of the decision subjects' backgrounds and in what context they received a Human-AI decision. Although 4 out of the 7 interviewed decision subjects have a background in computer science and engineering, we did not explicitly evaluate their prior experience with AI or their level of expertise in the field.

3.2 Interview Protocol

We conducted semi-structured interviews [66] of the recruited participants. The questions were compiled by the two first authors. They were independently reviewed by the other two authors and approved by the ethics committee of the research institution. In addition, we conducted a mock interview with an AI practitioner and a decision subject and adjusted the wording of the questions to improve their understanding. These data were not used for analysis. The questions were designed in English and translated to French and German for those participants preferring one of these languages. Interviews took place either by telephone or videoconference, whichever participants preferred. Participants could choose to allow us to record the interviews for note-taking purposes. All 14 participants agreed to do so. A total of 685 minutes were recorded, and each interview lasted an average of 50 minutes. Participants had access to our written notes before we used them in the article to ensure that their anonymity was maintained. All participants allowed us to quote them in the study.

The interview protocol consisted of four parts (Table 4) evolving around: the context with respect to their interaction with AI, Human-AI trust definitions, trust factors, and trust evaluation. In this article, we focused on the data regarding definitions and factors in the analysis. Where possible, we kept the formulation of questions identical (see Trust Definition in Table 4) for both groups of the participants. We adjusted the formulation of the questions related to the personal experiences to reflect the role of each group (example in Trust Factors, Table 4). We asked the questions around the Human-AI and trustworthiness definitions to understand what participants consider to be prerequisites of trust, that is in what contexts it is appropriate to consider Human-AI trust. We asked AI practitioners about their strategies to establish trust in their AI tool to understand what factors AI practitioners think influence trust of other stakeholders and which factors and stakeholders they prioritize. We did not explicitly refer to any group of stakeholders in our questions to let the AI practitioners spontaneously name the stakeholders relevant to the discussions around Human-AI trust. We asked decision subjects to share their experiences with receiving Human-AI decisions and, notably, what made them trust these decisions to identify the factors that influence their trust in AI. We also wanted to know whether decision subjects thought they trusted these decisions or AI in general too much or too little to gain more insights about what factors they prioritized to calibrate their trust towards more appropriate levels.

There were 8 questions in total as approximate guidance for the interviewers (Appendices A and B). When needed, we deepened the topic with follow-up questions about all the stakeholders involved in an anecdote, clarifying theoretical terminology, possible solutions to a described challenge, and whether a proposed factor always has effect on Human-AI trust.

3.3 Analysis of the Interviews

The first and second authors transcribed all interviews, removed all personal information (name of team, company, city, etc.) from the text, and assigned a code name to each interviewee, **P** for AI practitioners and **DS** for decision subjects. After transcription, the researchers deleted the audio files and allowed participants to review the interview text if they wished. The two researchers also translated the French and German texts to English and validated the translation with native speakers of the respective languages. Subsequently, the two researchers independently read all interviews at least twice, first without taking any notes and the second time highlighting the phrases or words related to people's experiences and needs with AI, to get familiarized with the data.

The further data analysis was based on the inductive thematic analysis [14, 24], that is a bottom-up approach to coding and analysis driven by the data itself. The two authors independently assigned to each highlighted phrase a code

	AI practitioners	Decision Subjects
Context	*How would you describe your role in the company? What is the main objective of your system?*	*Could you please tell me about your experience with Human-AI decision making?*
Trust Definition	*How would you define Human-AI trust in your own words? How would you define Trustworthy AI in your own words?*	
Trust Factors	*What is your strategy to establish trust of various stakeholders in your AI?*	*Have you ever trusted AI too much / too little?*
Trust Evaluation	*How would you know if someone trusts your AI?*	*Do you think AI developers consider human trust?*

Table 4. Structure and examples of questions per each group of participants. Data analysis of this paper mostly relies on answers around the definitions and factors of Human-AI trust. A full list of questions is in Appendices A and B.

that encapsulates the best its main message, focusing on the semantic content of the data. They then compared the list of highlighted phrases and their codes, discussed whether to include or not the phrases highlighted only by one of the researchers, and fine-tuned the wording of the codes for the finalized list of the selected phrases. After three iterations, the first author organized the codes in a series of sub-themes. They were further reformulated or merged with the consensus of all four authors in the process of writing the paper, and organized, under three main themes: one on the definition of trust, one on the role of interpersonal relations, and one on the divergent opinions between AI practitioners and decision subjects on the factors affecting trust (further described in the next section).

4 Findings

The thematic analysis yielded three main themes discussed in this section. We first explore the definitions of Human-AI trust from the perspectives of AI practitioners and decision subjects. Secondly, we find that both groups of respondents attribute significant importance to trust in interpersonal relationships, rather than in system characteristics. We conclude the results section by emphasizing some differences in opinions between the groups regarding the impact of AI transparency, AI literacy, and interactivity on Human-AI trust.

4.1 On the Definition of Trust

When prompted to define Human-AI trust in decision making, the interviewees identified three prerequisites for trust: positive expectations that AI will be beneficial in achieving the goals, the perceived risk associated with a decision, and the complexity of the task at hand. Importantly, the interviewees differentiated between trust, trust-related behavior, and trustworthiness.

4.1.1 Positive expectations and perceived risk are prerequisites for the emergence of trust, but the nature of risk is debated.

The interviewees state that for trust to emerge, people must have **positive expectations** that AI will help them achieve their goal and is aligned with their interest. They defined goal as *"the best answer in the shortest time"* (DS5, DS7). P6 also highlights that AI recommendations must be aligned with the goal of people interacting with or affected by the system as opposed to the technology owner's interest: *"It is important that the owner [of an AI-embedded system] does not recommend something in the company's interest"* (P6).

Moreover, the interviewees refer to the **perceived risk associated with a decision** as another prerequisite for the emergence of trust: *"When my physical integrity or money is at risk, trust becomes a consideration, especially when something important is at stake for me"* (P4)[3]. Several participants associate risk with health (DS2, DS4, DS5, DS6) or financial stability (DS4, DS5, DS6). P4 refers to risks related to economic loss or threats to life and health as universal, stating, *"... a foundation [for defining risk] would be the physical needs and individual and social integrity from the Maslow's Hierarchy."* However, some, like P5 and P2, broaden the concept of risk to include *"vulnerability"* (P5) or *"responsibility"* (P2), showing that risk extends beyond just financial or health concerns. P4 notes that what is considered risky varies from person to person, as *"not everyone has the same priorities"*. For example, DS4 found even Tinder recommendations could induce vulnerability, recounting moments when *"the algorithm says that I am ugly, something about myself that I do not want to accept"* (DS4). Therefore, DS4's experience of feeling vulnerable when their appearance was judged by AI indicates that the associated risk goes beyond monetary losses or health hazards and is closely related to one's personal vulnerabilities and priorities. This points to the situatedness of the risks involved and suggests that the mere application domain of the AI-assisted decision is not enough to indicate the level of risk associated; rather, it is the perceived risk based on individual vulnerabilities and priorities that matters.

4.1.2 Task complexity as a new prerequisite for the emergence of Human-AI trust.

Besides positive expectations and perceived risk as prerequisites for human trust in AI to emerge, some interviewees (P2, P4-P6, DS5) also mention **task complexity**. P2 and P6 describe "complex task" as a situation when a person cannot determine the quality of AI

[3]In this quote, AI practitioner P4 refers to their general reflection about what could trigger one's trust to emerge, not taking a particular perspective as an AI practitioner nor a decision subject

recommendation and, as a result, has many doubts around the final decision. DS5 agrees with P2 and P6, citing data analysis as an example of a task that is complex because: *"it is very difficult for a human to perform calculations and test the system."* A task is also perceived as more complex if the decision to make is a long-term one (P4). P5 suggests that when users face a complex task, trust emerges as a tool to mitigate the complexity: *"Sometimes you can't evaluate everything, you sort of use that quick «I just trust you, I just trust you to do the right thing»."* Interestingly, while the interviewees reported task complexity as one prerequisite for trust, it is not present in the usual definitions of trust [45, 103], which typically considers two prerequisites: "positive expectations" and "vulnerability".

4.1.3 Trust is differentiated from trust-related behaviors and trustworthiness.

Some interviewees differentiate between trust (which is defined as an attitude [45]) and trust-related behaviors. For instance P4, P5, and P6 postulate that inferring users' level of trust in AI from simply observing their behaviors could be misleading. Because users *"can have a complex and elaborate way of thinking [about AI-embedded systems and recommendations]"* (P4). P3 indicates that a user might follow AI recommendations not out of trust, but because they *"have no other solutions"* (P3). The interviewees thus clarify that it is trust-related behaviors, not trust itself, that are in action. But trust-related behaviors are useful as they can serve as *"indicators"* of trust. As P2 notes, *"as long as there aren't too many complaints, no negative comments, [...] and the user uses the solutions, we can consider that trust is not broken"* (P2).

Additionally, four interviewees (P2, P4, P5, P7) explicitly differentiate trust in AI from AI trustworthiness. Contrary to trust, which is seen as *"human reaction"* (P5), trustworthiness relates to features of the system (P2, P5), e.g., *"whether the job has been well done"* in designing and developing the system (P7). Such distinction further supports the stance that it is important to focus not only on what makes AI trustworthy, but also on what makes people trust AI [58]. Interestingly, two interviewees associated trustworthiness with AI governance, i.e., the institution or organization behind the AI. P4 states: *"For me, it [trustworthiness] is not so much a question of AI, it's more between the individual and the entity or the organization that makes the system."*

4.2 The Role of Inter-personal Relations on Trust

We discovered that trust between humans and AI is influenced by the trust among various stakeholders involved in the creation, use, and evaluation of the AI-based decision support system. Furthermore, AI certification, as a potential solution, is also contingent on trust within an interpersonal relational network.

4.2.1 The team behind AI plays an important role in (Human-AI) trust.

The interviewees indicated that an individual's trust in AI is closely linked to the level of trust they have in other stakeholders (Human-Human trust) within the socio-technical ecosystem. This ecosystem includes interactions between various stakeholders (such as AI practitioners, users, and decision subjects) and even the technical characteristics of the system. We found four cases illustrated in Figure 1. AI practitioners (P2, P4, P7) particularly emphasize Case 1: trust between the **users and the AI team**, where the AI team is the group of people behind the creation of the system. If users trust the AI team, their trust in AI *"[...] is established before the system exists. [...] Trust is very strong in the co-design phase [between users and the AI team]"* (P4). Interestingly, previous work on Human-AI trust does not generally consider trust between users and the AI team as a factor in Human-AI trust, even though it is likely to occur in real-world scenarios.

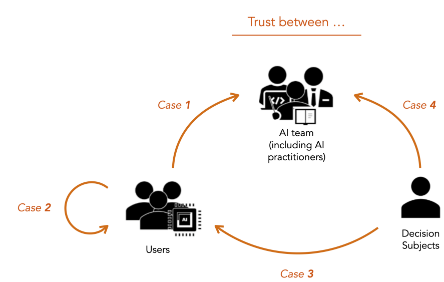


Figure 1. Schematic representation of the extent to which trust between different stakeholders groups discussed in relation to how it can affect Human-AI trust in the context of decision making.

AI practitioners (P3 and P6) also talk about the trust between the **users and other users** of the same system (Case 2). They claim that previous experiences of other users influence users' trust in AI: *"We have 10,000 users, and 90% of them say «the feedback from the AI was very interesting», now [knowing this, current users] will tend to trust the AI"* (P6). This trust in AI is further strengthened if *"a domain expert confirms what the AI recommends"* (P6).

Case 3 examines the trust that **decision subjects** place in the **users** of the system, concerning their usage and purpose. For instance, DS3 noted that trust in the system and its users are intertwined: *"There is trust in the system and trust in those who use the system [...]. They [the users] should at least tell you they are using such a system [embedding AI] so you don't lose your chance, just because you don't know how it works [...]"*. DS7 highlighted the complexity of this trust dynamic involving both humans (users) and machines from the perspective of the person being impacted by the decision: *"I don't trust mixing humans and machines. Either the decision* *should be entirely made by a machine or a human. If you have only one machine, then you know what to expect. But if you have a machine and a human, then it would be very unfair because the users' roles are not defined, and the priority is not clear."*

Finally, one **decision subject** also talked about the role of **trust in AI team** for decision subjects (Case 4). DS7 cites the example of Elon Musk and Tesla (at the time of the interviews), explaining that the trust of decision subjects in the company's high-level management influence their perceptions of and trust in the AI systems they develop: *"[he] is building trust with people through his own presence in the media [...]. People trust him and love his personality, so they trust his product even if it does not benefit them in the end."*

4.2.2 The effect of AI certification on Human-AI trust depends on who is behind it.

The interviewees (P1, P4, P6, DS1-DS3, DS5, DS7) share the view that knowing that an AI system has been certified is a factor that influences trust in that system, because *"certification has always been a way to gain confidence in technological tools, whether they are AI [or not]"* (P6). This is especially true for critical systems: "*The objective is clear - we [AI team] want certification*" (P1). P4 says that "*the certification alone should be enough [for Human-AI trust] if it is done well.*" However, some interviewees highlight the importance of who is behind the certification, rather than the sole fact of AI having been certified (DS1-DS3, DS5, DS7): *"AI certificates are very important [for Human-AI trust] if there are organizations [that issue them] that people can trust"* (DS2).

Finally, P5 and DS7 are more suspicious about certification in general because they think there is not yet enough scientific evidence that *"certification will build trust [in AI], I am not quite convinced of that yet"* (P5) or because a certification does not warrant that everything will be alright *"if there is a hack or a problem"* (DS7).

4.3 Diverging Opinions on Three Factors impacting Human-AI trust

Three factors playing a role for Human-AI trust were considered differently by AI practitioners and decision subjects, and are important to be highlighted. These factors are: AI transparency, AI literacy, and the increase of interactivity on the system. We decided to focus on them, rather the factors that both groups of stakeholders agreed on (e.g. AI performance and errors, marketing of the system, expectations about the system), as we believe that the identified diverging opinions provide interesting insights and implications for the research community.

4.3.1 AI practitioners and decision subjects do not share the same view on the role of AI transparency on trust.

AI transparency is one of the most discussed Human-AI trust factors in the interviews (P1-P7, DS3, DS4, DS7). The interviewees define two levels of transparency: a) explaining why a specific **AI recommendation** was shown and its quality, and b) explaining the **working processes** of AI development team.

The opinions about the effect of **explanations of AI recommendations** (a) on trust diverges not only between decision subjects and AI practitioners, but also among AI practitioners themselves. Some AI practitioners believe that explaining why a specific recommendation was shown can affect Human-AI trust, because it provides better understanding of how the recommendation was derived and, thus, lets estimate recommendation's quality (P2, P3, P6). At the same time, P4 strongly questions the necessity of understating for trust: *"One has to stop wondering how one can make tools that are more explainable, interpretable, or whatever, because sometimes there are tools that are not explainable in which we trust, a plane or a car, we don't know how it works inside, and yet we use them [...]"* (P4). Additionally, P1 and P7 raise concerns about the extent to which explanations can contribute to one's understanding of an AI recommendation: *"All the latest methods [of explainability] that have been developed are often so complex that humans [laypeople] do not understand them, so the methods do not help them at all"* (P1). Decision subjects further disagree with the AI practitioners supporting usefulness of AI explanations (P2, P3, P6). They state that besides AI explanations being complex (DS3, DS4), they have limited contribution to understanding of AI recommendations and, consequently, trust, because of the real world constraints: *"If people had the time to go through the explanations and review them in practice, they would have made the decision themselves in the first place"* (DS7).

Additionally, the AI practitioners (P3, P4, P7) seem to put considerable importance on transparency around the **working processes** (b) of AI development team, while decision subjects did not mention this aspect of transparency at all in connection to trust. The AI practitioners believe that the working process is the most actionable means of AI transparency for their clients, i.e. users that request development of a specific AI algorithm either for their own business or for a third party. The examples of explaining the working processes could be explaining the data, e.g., *"you have to be very, very transparent about how you prepared the data, because any AI is biased just because of the quality of the data (and also the quantity)"* (P7) and explaining the specifics of the system and AI in general, e.g. *"when we [...] try to be as transparent as possible on how it [the AI-embedded system] works, we try to explain it to them [clients], because it can be sometimes quite technical, even mathematical, and then there are no more problems, no problem of trust..."* (P4).

4.3.2 AI literacy: decision subjects perceived AI literacy as more specific and operational than AI practitioners.

There have been diverging opinions about the role of AI literacy for Human-AI trust between AI practitioners and decision subjects: while AI practitioners emphasize the need for raising general public awareness around AI, decision subjects believe in system-related literacy, i.e. more specific and operational knowledge about AI. P5 believes public education on the general understanding of AI could be beneficial for calibrating human trust in AI, *"because people will say «I do not trust AI», without really understanding what AI is"* (P5). Similarly, P7 believes that users should understand the boundaries in AI performance - what AI can do and cannot do. However, for decision subjects, it is not enough to raise public awareness about how AI works in general, because it is not specific enough, e.g. *"educational events [about AI] do not really make sense to me, because often nobody knows how the system really works"* (DS4), or not actionable enough, e.g. *"the educational sessions [about AI] do not make sense to me, how can they help?.."* (DS6). Therefore, affecting Human-AI trust through AI literacy seems to be possible by accounting for needs of a specific stakeholder group. For example, for decision subjects to understand how a Human-AI decision is made to be able to act upon it, P7 provides training tailored for their decision subjects: *"[we] create materials, [...] flyers, [...] content for patients so that they are informed, that they are not afraid of this new technology"* (P7).

4.3.3 Interactivity: exploration tool for AI practitioners, means to be included in the loop for decision subjects. AI practitioners and decision subjects agree that interactivity is another factor impacting Human-AI trust in the context of decision making (P1-P4, DS2, DS3, DS4, DS6). They also agree that interactivity is often limited. For instance, *"I give you [AI] input data - you [AI] send me back the solution, and I have no other contextual elements, elements of interaction with you"* (P2), *"I would like to have the opportunity to negotiate and influence the [the AI's] decision and say, «Hey, but look at this and that»"* (DS4) or "*these [AI] systems should be more tolerant to human error. Right now, it's so strict*" (DS6). However, their opinions differ when considering the consequences of this limited interactivity. For AI practitioners, it hampers one's ability to explore the system, *"asking for more explanations"* (P3) and establish *"a dialogue"* (P4) or *"cooperation"* (P1) between users and AI. For decision subjects, the limited interactivity leads to more serious consequences. It provides a feeling of being excluded from the loop. Decision subjects feel they lose their sense of agency. They see themselves as "*statistics*" (DS2) or simply "*filtered out*" by AI (DS3) because the system is not "flexible" (DS3) or does not allow "to negociate" (DS4).

5 Discussion

In this paper, we investigated Human-AI trust from two perspectives - what AI practitioners think is important for trust in AI of other stakeholders and what decision subjects think is important for their trust in AI. Combining these perspectives allows for understanding similarities and differences in how these different stakeholders define Human-AI trust in the context of decision making and what factors affecting Human-AI trust they prioritize. In this section, we discuss what our results mean for 1) re-envisioning what factors affect Human-AI trust in the socio-technical ecosystem; 2) defining Human-AI trust and its key prerequisite elements for its existence; and (3) in terms of stakeholders' agency over the system. Finally, we present some limitations of our study and propose future research directions that could address these limitations.

5.1 On the Important Role of Inter-personal Relations on Trust Within the Socio-technical System

Our results revealed the important role of interpersonal relationships on trust. In other words, AI practitioners and decision subjects stressed the importance of trust links with other stakeholders involved in the system: its design, development, deployment or use in real applications. Moreover, this importance seems to take precedence over the technical characteristics of the system. These results complement recent findings on the under-explored concept of social transparency for AI-assisted decision-making [31]. Through highlighting the history of other users' interactions with AI recommendations rather than the inner workings of AI, social transparency embraces the interviewees' emphasis on trust factors related to social interactions, information actionability, and expectations as a part of the system's design. In this sense, our findings about the importance of interpersonal relationships also support recent approaches arguing for trust calibration beyond direct interaction of people with AI [17].

From the different cases of trust links between stakeholders elicited in the findings, trust in the AI team (cases 1 and 4 in Figure 1) is generally absent in the literature, while respondents believe that this plays an important role in the trust between humans and AI. So far, the literature suggests that the reputation of the organization that develops AI plays a role for doctors' trust in AI recommendations [20, 93], and our study confirms this for the domains beyond medical decision making. Another difference is that users' trust in other users (case 2) is more emphasised in the academic literature than in the interviews [16, 31, 44, 72, 82]. Research shows that observing other users (especially colleagues) trusting the recommendations of the system can increase one's own trust in AI [31, 44]. However, from the interviews, AI practitioners often serve as intermediaries between users and convey feedback as product reviews. Finally, the relationship of trust between decision subjects an other users (case 4) is barely present in the interviews. The academic literature shows that if decision subjects (e.g., a patient) trust the direct user (e.g., a clinician) and the direct user trusts the AI recommendations, then they would also trust the AI recommendations [72] and vice versa [16, 27].

Our results suggest that these bonds of trust are either transversal (e.g. users to users) or upstream (e.g. users to AI team). We believe that trust, in this case, relates to the people who have either more expertise on the domain and technology or means of actions over the technology (such as the AI team). Therefore, these trust links might take an even more important role for decision subjects than other stakeholders. In fact, we saw this in the perception of the role of AI transparency on trust. Contrary to AI practitioners, decision subjects do not see how transparency can affect their trust in AI since explanations might be difficult to understand and the additional information about AI or a specific system is usually not actionable. Specifically, neither the interviewed AI practitioners, nor the literature provide ample reflections for the role of transparency for trust in AI of decision subjects. Transparency is, hence, viewed as a factor affecting primarily users' trust in AI, targeting their needs for quality evaluation of an AI recommendation and for refining their mental model about AI, which does not necessary encompass actionability and contestability - the needs of decision subjects [68, 114].

Research implications.

1. **Investigating the points of Human-AI trust breakdowns and calibrations beyond direct interaction with the system**. To this end, research needs to involve more fieldwork with the various stakeholders to understand how trust in AI is shaped and influenced within the complex web of relationships among AI practitioners, decision subjects, users, and other stakeholders, and identify key patterns and dynamics of trust flow among these stakeholders.
2. **Re-examining the techno-centric trust factors between humans and AI with a social lens**. Following the example of Ehsan et al. [31], who proposed the term of social transparency, moving away from providing more information about how AI works to more information about how other users make decisions with the system, we envision other Human-AI trust factors can be relooked in the same manner. For instance, in addition to reporting AI accuracy, one can inform users about how AI recommendations affected the performance of other users.

5.2 On the Prerequisites for the Existence of Trust

In order to understand what AI practitioners and decision subjects expect from a system they trust, we analysed how these stakeholders understand trust, i.e. what essential elements, or prerequisites, they associate with this notion. Both groups elicited the need for positive expectations and a situation of vulnerability. These two prerequisites are how theoretical work in the literature defines trust. This was unexpected, because trust is a complex and abstract theoretical concept that leads to frequent theoretical confusions [45, 58, 103]. It remains that we found a more nuanced outlook on the key elements of trust in comparison with the academic literature. The interviewees' discussions highlight that vulnerability and positive expectations cannot be boiled down to monetary losses and high levels of accuracy as they are often presented in the empirical studies [103]. Vulnerability denotes a state in which someone feels the possibility of being emotionally attacked, and therefore finds themselves in a position of weakness. In our results, we had the example of a judgement based on physical appearance. So these prerequisites for the existence of trust depend on the individual or the community with which the individual identifies. Recent examples of the behaviour of algorithms that discriminate against a certain population, such as black women [19], place them in a vulnerable position more than other individuals. Additionally, decision subjects report to feel vulnerable, because they have no control over how the data they share about themselves for Human-AI decision making get interpreted by the users in charge of these decisions [27]. Sometimes, in order to appear cooperative, they provide more data about themselves than needed, which puts them at risk of "algorithmic stigmatization" [6, 83] - wrongfully assigned a certain label "at risk", e.g., risk of recidivism, child maltreatment, suicidal tendencies, based on the an algorithmic assemblage.

Our results also highlighted a new prerequisite for the existence of trust, namely the complexity of the task. Behind this prerequisite is the idea that if the task is simple, it can be easily solved by the person using the system or receiving a decision from it. Thus, if one knows the right answer, evaluating the quality of AI recommendation is straightforward, the conflict of between one's own opinion and the AI recommendation does not emerge, and consequently, neither does the state of trust. However, there is an ambiguity about the definition of complexity. Typically, we could envisage two scenarios. Firstly, complexity can arise from the impossibility for a human to process a large amount of information (for example, a large amount of data in a database) in order to produce a decision. Secondly, complexity can arise from a lack of expertise, either related to the decision domain, or related to the underlying AI techniques. If task complexity is a prerequisite for the emergence of trust, along with vulnerability and positive expectations, this implies that future research should study it and include it in the way experimental tasks are designed to focus on trust, rather than confidence [103].

Research implications.

1. **Understanding the role of vulnerability in AI-based decision-making systems**. As our findings indicate that feeling vulnerability to AI-based decision-making systems can go beyond monetary gains and losses, especially in the case of decision subjects, qualitative studies, such as interviews and case studies,

could be utilized to gain deeper insights into individual and community experiences of vulnerability towards AI in order to inform further research on trust.

2. **Incorporating complexity into experimental studies of trust**. Given that task complexity could be a key element in trust formation, it should be accounted for in designing experiments that study trust in AI to distinguish between trust and confidence in the system's recommendations. Future research should investigate what aspects of task should be considered to vary the task complexity as well as to which extent it contributes to the formation of trust as a function of different levels of task complexity.

5.3 On the Notion of Agency over the System

AI practitioners and decision subjects both stress the importance of AI interactivity for trust, but their views on the purpose of interactivity differ. For AI practitioners, interactivity is a means to explore AI recommendations. From this point of view, they agree with what previous work has shown about the fact that interactivity contributes to explore to which extent nuances are accounted for AI recommendations [82]. Other works have shown, in addition, that interactivity also contributes to the refinement of the mental model about AI [22] and gives a sense of striving to improve decision making [72]. Decision subjects, on the other hand, see interactivity as a way of getting involved in the decision-making loop. In other words, they see interactivity first as a way of being represented in the decision-making process, before being able to formalise what this representation could bring in terms of understanding the system's mechanisms and creating a mental model of its behaviour. We therefore see, in these different opinions between the stakeholders, a difference in power relationships. AI practitioners have the means to act on the system, and are therefore in a position to imagine what these means can bring them.

This interpretation suggests that interactivity is related to the notion of agency. In fact, decision subjects discuss the sense of agency and its relationship to Human-AI trust more than AI practitioners, which is expected considering the mentioned frustrations about their lack of actionability and power over the systems. This means that decision subjects value more the factors of trust linked to their inclusion in the decision-making loop in comparison with AI practitioners. It is an empowerment over Human-AI decisions so as not to feel solely "*part of the statistics*", as put by DS2. These findings align with the prior work [46] showing that different groups of stakeholders prioritize ethical values differently. Our findings extend this line of research by demonstrating this for trust and underlines the importance of undertaking a multi-stakeholder approach [115] for Human-AI trust.

That being said, although the academic literature on Human-AI trust examining the interactivity of AI recommendations have led to certain results, as those mentioned above, this research remains scarce [15, 22, 38, 72, 82]. Moreover, these studies are primarily about users rather than decision subjects. Similarly, while previous work has investigated the relationship between agency and trust in AI, it focuses exclusively on the agency of direct users (e.g., [18, 33, 48, 86, 96, 98, 109]). Additionally, in all these articles, participants are fully aware to which extent they have control over AI recommendations, and their level of agency remains unchanged throughout the experiment. Hence, the issue of varying levels of control over AI is not largely studied in the Human-AI trust literature in the context of decision making. Moreover, in the interviews, agency is mostly referred to as ability to contest a Human-AI decision, while in the literature, it is mainly represented as control over seeing an AI recommendation: full - AI recommendations are optional and appear on demand [18, 52, 52, 86, 96, 99], limited - mandatory AI recommendations that appear immediately [18, 33, 52, 77, 86, 96, 98, 99] or only after users' initial decision [18, 33], and none - AI recommendations executed autonomously [52, 69, 77, 98, 99]. Therefore, it remains unclear to which extent the solution of "introducing four levels [of AI recommendations] instead of the binary [...]" proposed by P7 to increase the sense of agency for decision subjects would work.

Research implications.

1. **Investigating the role of different mechanisms of interactivity for trust in AI of various stakeholders.** Our findings indicate that interactivity plays a different role for decision subjects than for AI users, and thus might affect their trust in AI not through the same mechanisms. HCI researchers could conduct in-depth studies to examine how different interactive features (e.g., feedback loops, adjustable parameters) empower decision subjects or change trust links between them and practitioners or users.
2. **Investigating agency as human capability instead of a feature of the system.** Our results have shown the importance of human agency, particularly for decision subjects. While agency tends to be seen as a feature of the system (e.g., providing means to act on system behavior), it is also related to people's perception of actions on the system and their representation by the system. In the same way as trust, this concept needs to be better understood from a human-centric point of view in the context of interactions with AI-based decision-making systems.

5.4 Future Work Directions

In this article, we interviewed representatives from a varied panel of decision domains (e.g. medicine, finance, recruitment). Although our main objective was to study the factors of trust between humans and artificial intelligence for risk-sensitive applications, each domain may nuance the effects on trust due to the diversity of decision-making flows, types of stakeholders involved, etc., which is one of the limitations in the interpretation of the study's results. Considering that type of task and level of risk also have impact on Human-AI trust, it could be interesting to conduct a cross-domain comparison to see to which extent they put importance on the same Human-AI trust factors. Understanding the differences and similarities between various task domains can inform researchers and policy makers on higher level classification of domains [54]. Additionally, we did not account for individual differences such as gender, age, and explicitly assess prior experience with AI, and other demographic information in our analysis while these factors can further influence how certain ethical values are prioritized [46].

Secondly, we considered two types of stakeholders that are not users - AI practitioners and decision subjects. While there is no widely established categorization, some researchers propose a set of 11 stakeholders' groups [8] that are connected to the AI ecosystem, spanning from policy makers that work on high level strategies to hiring managers that recruit AI developers. An interesting research direction would be to extend the presented research to these stakeholders and inspect differences and commonalities in findings.

Lastly, we took an organization-focused approach to studying Human-AI trust when talking to AI practitioners. In other words, the AI-embedded systems that they are responsible for are developed, trained, designed, deployed, and monitored by the same company. However, nowadays AI technologies are often a product of "algorithmic supply chains" [25], that is multiple independent actors are responsible for commissioning different phases of production and deployment. As these actors have distributed responsibility over the outcomes of Human-AI decisions with imperfect control over how their work is used further down in an algorithmic supply chain, this can raise additional concerns over whether an AI recommendation produced by "many hands" can be trusted. Further investigating implications for Human-AI trust resulting from such a production set-up can shed light on new nuances about interpersonal dynamics between different stakeholders and identify new potential Human-AI trust breakdown points and factors that affect it.

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. This research was supported by the ARCOL project (ANR-19-CE33-0001) – Interactive Reinforcement Co-Learning, from the French National Research Agency. We wish to acknowledge and thank everyone involved in this research, and express our sincere gratitude to the anonymous reviewers for their constructive comments.

References

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[2] B. D. Adams, L. Bruyn, S. Houde, and P. Angelopoulos. 2003. *Trust in Automated Systems literature review*. Technical Report. Defence Research and Development Canada, Toronto, Ontario, Canada. 124 pages. https://cradpdf.drdc-rddc.gc.ca/PDFS/unc17/p520342.pdf

[3] Ifeoma Ajunwa, Sorelle Friedler, Carlos E Scheidegger, and Suresh Venkatasubramanian. 2016. Hiring by algorithm: predicting and preventing disparate impact. *Available at SSRN* 2746078 (2016), 29.

[4] Sanna J. Ali, Angèle Christin, Andrew Smart, and Riitta Katila. 2023. Walking the Walk of AI Ethics: Organizational Challenges and the Individualization of Risk among Ethics Entrepreneurs. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 217–226. https://doi.org/10.1145/3593013.3593990

[5] Asbjørn Ammitzbøll Flügge, Thomas Hildebrandt, and Naja Holten Møller. 2021. Street-Level Algorithms and AI in Bureaucratic Decision-Making: A Caseworker Perspective. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 40 (April 2021), 23 pages. https://doi.org/10.1145/3449114

[6] Nazanin Andalibi, Cassidy Pyle, Kristen Barta, Lu Xian, Abigail Z. Jacobs, and Mark S. Ackerman. 2023. Conceptualizing Algorithmic Stigmatization. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 373, 18 pages. https://doi.org/10.1145/3544548.3580970

[7] Narges Ashtari, Ryan Mullins, Crystal Qian, James Wexler, Ian Tenney, and Mahima Pushkarna. 2023. From Discovery to Adoption: Understanding the ML Practitioners' Interpretability Journey. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 2304–2325. https://doi.org/10.1145/3563657.3596046

[8] Jacqui Ayling and Adriane Chapman. 2022. Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics* 2 (08 2022). https://doi.org/10.1007/s43681-021-00084-x

[9] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence* (Honolulu, Hawaii, USA) *(AAAI'19/IAAI'19/EAAI'19)*. AAAI Press, Washington D.C, USA, Article 300, 9 pages. https://doi.org/10.1609/aaai.v33i01.33012429

[10] Riccardo Baratella, Glenda Amaral, Tiago Prince Sales, Renata Guizzardi, and Giancarlo Guizzardi. forthcoming. The Many Facets of Trust. In *Formal Ontology in Information Systems*. IOS Press, Amsterdam, Netherlands.

[11] Jason M. Bindewald, Christina F. Rusnock, and Michael E. Miller. 2018. Measuring Human Trust Behavior in Human-Machine Teams. In *Advances in Human Factors in Simulation and Modeling*, Daniel N. Cassenti (Ed.). Springer International Publishing, Cham, 47–58.

[12] Ann M Bisantz and Younho Seong. 2001. Assessment of operator trust in and utilization of automated decision-aids under different framing conditions. *International Journal of Industrial Ergonomics* 28, 2 (2001), 85–97.

[13] 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

[14] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* American Psychological Association, Washington, 57–71.

[15] 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

[16] 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

[17] Jacob Browne, Saskia Bakker, Bin Yu, Peter Lloyd, and S. Ben Allouch. 2022. Trust in Clinical AI: Expanding the Unit of Analysis. In *HHAI2022: Augmenting Human Intellect*, Vol. 354. IOS Press, Amsterdam, Netherlands, 96–113.

[18] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (April 2021), 21 pages. https://doi.org/10.1145/3449287

[19] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[20] Eleanor R. Burgess, Ivana Jankovic, Melissa Austin, Nancy Cai, Adela Kapuścińska, Suzanne Currie, J. Marc Overhage, Erika S Poole, and Jofish Kaye. 2023. Healthcare AI Treatment Decision Support: Design Principles to Enhance Clinician Adoption and Trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 15, 19 pages. https://doi.org/10.1145/3544548.3581251

[21] Federico Cabitza, Andrea Campagner, Riccardo Angius, Chiara Natali, and Carlo Reverberi. 2023. AI Shall Have No Dominion: On How to Measure Technology Dominance in AI-Supported Human Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 354, 20 pages. https://doi.org/10.1145/3544548.3581095

[22] 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

[23] Claude Castelluccia and Daniel Le Métaye. 2019. *Understanding algorithmic decision-making: Opportunities and challenges.* EU Publications, Strasbourg, France. 1–79 pages. https://doi.org/10.2861/536131

[24] Victoria Clarke and Virginia Braun. 2013. Teaching thematic analysis: Overcoming challenges and developing strategies for effective learning. *The Psychologist* 26 (02 2013), 120–123.

[25] Jennifer Cobbe, Michael Veale, and Jatinder Singh. 2023. Understanding Accountability in Algorithmic Supply Chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1186–1197. https://doi.org/10.1145/3593013.3594073

[26] 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

[27] Vedant Das Swain, Lan Gao, William A Wood, Srikruthi C Matli, Gregory D. Abowd, and Munmun De Choudhury. 2023. Algorithmic Power or Punishment: Information Worker Perspectives on Passive Sensing Enabled AI Phenotyping of Performance and Wellbeing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 246, 17 pages. https://doi.org/10.1145/3544548.3581376

[28] Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. 2023. Investigating Practices and Opportunities for Cross-Functional Collaboration around AI Fairness in Industry Practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 705–716. https://doi.org/10.1145/3593013.3594037

[29] Advait Deshpande and Helen Sharp. 2022. Responsible AI Systems: Who Are the Stakeholders?. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 227–236. https://doi.org/10.1145/3514094.3534187

[30] Yanqing Duan, John S. Edwards, and Yogesh K Dwivedi. 2019. Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management* 48 (2019), 63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

[31] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 82, 19 pages. https://doi.org/10.1145/3411764.3445188

[32] Andrea Ferrario and Michele Loi. 2022. How Explainability Contributes to Trust in AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1457–1466. https://doi.org/10.1145/3531146.3533202

[33] Riccardo Fogliato, Alexandra Chouldechova, and Zachary Lipton. 2021. The Impact of Algorithmic Risk Assessments on Human Predictions and Its Analysis via Crowdsourcing Studies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 428 (oct 2021), 24 pages. https://doi.org/10.1145/3479572

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

[35] 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

[36] Meric Altug Gemalmaz and Ming Yin. 2022. Understanding Decision Subjects' Fairness Perceptions and Retention in Repeated Interactions with AI-Based Decision Systems. In *Proceedings of the 2022*

AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 295–306. https://doi.org/10.1145/3514094.3534201

[37] 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

[38] Akshit Gupta, Debadeep Basu, Ramya Ghantasala, Sihang Qiu, and Ujwal Gadiraju. 2022. To Trust or Not To Trust: How a Conversational Interface Affects Trust in a Decision Support System. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 3531–3540. https://doi.org/10.1145/3485447.3512248

[39] H. Güngör. 2020. Creating Value with Artificial Intelligence: A Multi-stakeholder Perspective. *Journal of Creating Value* 6, 1 (2020), 72–85. https://doi.org/10.1177/2394964320921071 arXiv:https://doi.org/10.1177/2394964320921071

[40] 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

[41] 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

[42] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. *Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?* Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3290605.3300830

[43] Yoyo Tsung-Yu Hou, Wen-Ying Lee, and Malte Jung. 2023. "Should I Follow the Human, or Follow the Robot?" — Robots in Power Can Have More Influence Than Humans on Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 114, 13 pages. https://doi.org/10.1145/3544548.3581066

[44] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 659, 14 pages. https://doi.org/10.1145/3411764.3445385

[45] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 624–635. https://doi.org/10.1145/3442188.3445923

[46] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 310–323. https://doi.org/10.1145/3531146.3533097

[47] Anil Kalhan. 2013. Immigration policing and federalism through the lens of technology, surveillance, and privacy. *Ohio St. LJ* 74 (2013), 1105.

[48] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "Because AI is 100% Right and Safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 158, 18 pages. https://doi.org/10.1145/3491102.3517533

[49] Jee Young Kim, William Boag, Freya Gulamali, Alifia Hasan, Henry David Jeffry Hogg, Mark Lifson, Deirdre Mulligan, Manesh Patel, Inioluwa Deborah Raji, Ajai Sehgal, Keo Shaw, Danny Tobey, Alexandra Valladares, David Vidal, Suresh Balu, and Mark Sendak. 2023. Organizational Governance of Emerging Technologies: AI Adoption in Healthcare. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1396–1417. https://doi.org/10.1145/3593013.3594089

[50] Taenyun Kim, Maria D. Molina, Minjin (MJ) Rheu, Emily S. Zhan, and Wei Peng. 2023. One AI Does Not Fit All: A Cluster Analysis of the Laypeople's Perception of AI Roles. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 29, 20 pages. https://doi.org/10.1145/3544548.3581340

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

[52] 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

[53] Francesca Lagioia and Giuseppe Contissa. 2020. The strange case of Dr Watson : liability implications of AI evidence-based decision support systems in health care. *Eur. J. Leg. Stud.* 12, 2 (2020), 241–289.

[54] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1369–1385. https://doi.org/10.1145/3593013.3594087

[55] 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

[56] 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

[57] 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

[58] Q.Vera Liao and S. Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1257–1268. https://doi.org/10.1145/3531146.3533182

[59] Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. 2022. The Conflict Between Explainable and Accountable Decision-Making Algorithms. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2103–2113. https://doi.org/10.1145/3531146.3534628

[60] Steven Lockey, Nicole Gillespie, Daniel Holm, and Ida Asadi Someh. 2021. A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions. In *HICSS-54*. Hawaii International Conference on System Sciences, Hawaii, USA. https:

//doi.org/10.24251/HICSS.2021.664

[61] 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

[62] Henrietta Lyons, Tim Miller, and Eduardo Velloso. 2023. Algorithmic Decisions, Desire for Control, and the Preference for Human Review over Algorithmic Review. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 764–774. https://doi.org/10.1145/3593013.3594041

[63] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. 2022. What's the Appeal? Perceptions of Review Processes for Algorithmic Decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 580, 15 pages. https://doi.org/10.1145/3491102.3517606

[64] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 759, 19 pages. https://doi.org/10.1145/3544548.3581058

[65] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 52 (apr 2022), 26 pages. https://doi.org/10.1145/3512899

[66] 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

[67] Marco Marabelli and Sue Newell. 2019. Algorithmic Decision-making in the US Healthcare Industry: Good for Whom? *Academy of Management Proceedings* 2019 (08 2019), 15581. https://doi.org/10.5465/AMBPP.2019.15581

[68] Amelie Marian. 2023. Algorithmic Transparency and Accountability through Crowdsourcing: A Study of the NYC School Admission Lottery. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 434–443. https://doi.org/10.1145/3593013.3594009

[69] 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

[70] 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

[71] 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/

[72] Chinasa T. Okolo, Srujana Kamath, Nicola Dell, and Aditya Vashistha. 2021. "It Cannot Do All of My Work": Community Health Worker Perceptions of AI-Enabled Mobile Health Applications in Rural India. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 701, 20 pages. https://doi.org/10.1145/3411764.3445420

[73] High-Level Expert Group on AI (AI HLEG). 2019. *Building Trust in Human-Centric Artificial Intelligenc.* Technical Report. European Commission, Brussels, Belgium. 11 pages. https://ec.europa.eu/jrc/communities/en/community/digitranscope/document/building-trust-human-centric-artificial-intelligence

[74] Jeroen Ooge, Shotallo Kato, and Katrien Verbert. 2022. Explaining Recommendations in E-Learning: Effects on Adolescents' Trust. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 93–105. https://doi.org/10.1145/3490099.3511140

[75] Rachel O'Dwyer. 2018. Algorithms are making the same mistakes as humans assessing credit scores. *Retrieved April* 17 (2018), 2019.

[76] 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

[77] 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

[78] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where Responsible AI Meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 7 (April 2021), 23 pages. https://doi.org/10.1145/3449081

[79] Divya Ramesh, Vaishnav Kameswaran, Ding Wang, and Nithya Sambasivan. 2022. How Platform-User Power Relations Shape Algorithmic Accountability: A Case Study of Instant Loan Platforms and Financially Stressed Users in India. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1917–1928. https://doi.org/10.1145/3531146.3533237

[80] Mark Ryan. 2020. In AI we trust: Ethics, artificial intelligence, and reliability. *Sci. Eng. Ethics* 26, 5 (Oct. 2020), 2749–2767.

[81] S Samoili, Cobo M Lopez, E Gomez Gutierrez, G De Prato, F Martinez-Plumed, and B Delipetrev. 2020. *AI WATCH. Defining Artificial Intelligence.* Technical Report KJ-NA-30117-EN-N (online). European Commission, Luxembourg (Luxembourg). https://doi.org/10.2760/382730(online)

[82] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2021. A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-Welfare. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 348 (oct 2021), 41 pages. https://doi.org/10.1145/3476089

[83] Devansh Saxena, Erina Seh-Young Moon, Aryan Chaurasia, Yixin Guan, and Shion Guha. 2023. Rethinking "Risk" in Algorithmic Systems Through A Computational Narrative Analysis of Casenotes in Child-Welfare. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 873, 19 pages. https://doi.org/10.1145/3544548.3581308

[84] Kristin E. Schaefer, Deborah R. Billings, James L. Szalma, Jeffrey K. Adams, Tracy Sanders, Jessie Y. C. Chen, and Peter A. Hancock. 2014. *A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Human-Robot Interaction.* Technical Report. U.S. Army Research Laboratory.

[85] Kristin E. Schaefer, Jessie Y. C. Chen, James L. Szalma, and P. A. Hancock. 2016. A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors* 58, 3 (2016), 377–400.

https://doi.org/10.1177/0018720816634228 PMID: 27005902.

[86] 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

[87] Matthew U Scherer. 2015. Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *SSRN Electron. J.* 29 (2015), 48.

[88] Timothée Schmude, Laura Koesten, Torsten Möller, and Sebastian Tschiatschek. 2023. On the Impact of Explanations on Understanding of Algorithmic Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 959–970. https://doi.org/10.1145/3593013.3594054

[89] Jakob Schöffer, Yvette Machowski, and Niklas Kühl. 2021. A Study on Fairness and Trust Perceptions in Automated Decision Making. In *Joint Proceedings of the ACM IUI 2021 Workshops, April 13–17, 2021, College Station, USA (CEUR Workshop Proceedings)*, Vol. 2903. RWTH Aachen, Aachen, Germany, 170005.

[90] Ian A Scott, Stacy M Carter, and Enrico Coiera. 2021. Exploring stakeholder attitudes towards AI in clinical practice. *BMJ Health & Care Informatics* 28, 1 (2021), 7. https://doi.org/10.1136/bmjhci-2021-100450 arXiv:https://informatics.bmj.com/content/28/1/e100450.full.pdf

[91] Younho Seong, Ann M Bisantz, and Ann M Bisantz. 2002. Judgment and Trust in Conjunction with Automated Decision Aids: A Theoretical Model and Empirical Investigation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 46, 3 (2002), 423–427. https://doi.org/10.1177/154193120204600344 arXiv:https://doi.org/10.1177/154193120204600344

[92] 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

[93] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. 2023. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 754, 18 pages. https://doi.org/10.1145/3544548.3581075

[94] Jessie J. Smith, Anas Buhayh, Anushka Kathait, Pradeep Ragothaman, Nicholas Mattei, Robin Burke, and Amy Voida. 2023. The Many Faces of Fairness: Exploring the Institutional Logics of Multistakeholder Microlending Recommendation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1652–1663. https://doi.org/10.1145/3593013.3594106

[95] Janet A. Sniezek and Lyn M. Van Swol. 2001. Trust, Confidence, and Expertise in a Judge-Advisor System. *Organizational Behavior and Human Decision Processes* 84, 2 (2001), 288–307. https://doi.org/10.1006/obhd.2000.2926

[96] 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

[97] 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

[98] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 160, 17 pages. https://doi.org/10.1145/3491102.3517732

[99] 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

[100] Lyn M Van Swol and Janet A Sniezek. 2005. Factors affecting the acceptance of expert advice. *Br J Soc Psychol* 44, Pt 3 (Sept. 2005), 443–461.

[101] Rama Adithya Varanasi and Nitesh Goyal. 2023. "It is Currently Hodgepodge": Examining AI/ML Practitioners' Challenges during Co-Production of Responsible AI Values. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 251, 17 pages. https://doi.org/10.1145/3544548.3580903

[102] 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

[103] 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 (Oct. 2021), 39 pages. https://doi.org/10.1145/3476068

[104] 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.

[105] Xiuxin Wang and Xiufang Du. 2018. Why Does Advice Discounting Occur? The Combined Roles of Confidence and Trust. *Front Psychol* 9 (Nov. 2018), 2381.

[106] Xinru Wang and Ming Yin. 2023. Watch Out for Updates: Understanding the Effects of Model Explanation Updates in AI-Assisted Decision Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 758, 19 pages. https://doi.org/10.1145/3544548.3581366

[107] David Gray Widder, Derrick Zhen, Laura Dabbish, and James Herbsleb. 2023. It's about Power: What Ethical Concerns Do Software Engineers Have, and What Do They (Feel They Can) Do about Them?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 467–479. https://doi.org/10.1145/3593013.3594012

[108] 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

[109] 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

[110] Ilan Yaniv and Eli Kleinberger. 2000. Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation. *Organizational Behavior and Human Decision Processes* 83, 2 (2000), 260 – 281. https://doi.org/10.1006/obhd.2000.2909

[111] Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas. 2023. Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People + AI Guidebook. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 356, 13 pages. https://doi.org/10.1145/3544548.3580900

[112] 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

[113] Kun-Hsing Yu and Isaac S Kohane. 2019. Framing the challenges of artificial intelligence in medicine. *BMJ Quality & Safety* 28, 3 (2019), 238–241. https://doi.org/10.1136/bmjqs-2018-008551 arXiv:https://qualitysafety.bmj.com/content/28/3/238.full.pdf

[114] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: The Effects of Explanations, Human Oversight, and Contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 134, 21 pages. https://doi.org/10.1145/3544548.3581161

[115] Mireia Yurrita, Dave Murray-Rust, Agathe Balayn, and Alessandro Bozzon. 2022. Towards a Multi-Stakeholder Value-Based Assessment Framework for Algorithmic Systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 535–563. https://doi.org/10.1145/3531146.3533118

[116] 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