



**HAL**  
open science

# Understanding Human-AI Trust in the Context of Decision Making through the Lenses of Academia and Industry: Definitions, Factors, and Evaluation

Oleksandra Vereschak

► **To cite this version:**

Oleksandra Vereschak. Understanding Human-AI Trust in the Context of Decision Making through the Lenses of Academia and Industry: Definitions, Factors, and Evaluation. Artificial Intelligence [cs.AI]. Sorbonne Université, 2022. English. NNT : 2022SORUS552 . tel-04125248

**HAL Id: tel-04125248**

**<https://theses.hal.science/tel-04125248>**

Submitted on 12 Jun 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



# Thèse

présentée à

Sorbonne Université

Ecole Doctorale N° 391 (SMAER)  
Sciences mécaniques, acoustique, électronique & robotique

Institut des Systèmes Intelligents et de Robotique (ISIR)

par

Oleksandra Vereschak

pour obtenir le diplôme de  
Doctorat de Sorbonne Université

---

Understanding Human-AI Trust in the Context of Decision Making  
through the Lenses of Academia and Industry:  
Definitions, Factors, and Evaluation

---

*Thèse soutenue : 12/12/2022*

Pr. N. Krämer	Directrice de recherche au Département de psychologie sociale : médias et communication - l'université de Duisbourg et Essen	Rapporteure
Pr. M. Lewkowicz	Professeure à l'université de Technologie de Troyes	Rapporteure
Dr. C. Pelachaud	Directrice de recherche au CNRS - Sorbonne Université, ISIR	Examinatrice, présidente du jury
Pr. N. van Berkel	Professeur à l'université d'Aalborg	Examineur
Dr. G. Bailly	Directeur de recherche au CNRS - Sorbonne Université, ISIR	Directeur de thèse
Dr. B. Caramiaux	Chargé de recherche au CNRS - Sorbonne Université, ISIR	Co-encadrant de thèse
Dr. T. Baudel	Directeur de recherche & Master Inventor à l'IBM France R&D Lab	Invité



# *Abstract*

With the rise of AI-embedded systems assisting decisions in the context of medicine, justice, recruiting, Human-AI trust has become an utmost design priority. Numerous governments and large enterprises as well as researchers propose various strategies on how to foster trust in AI-embedded systems. However, trust is a complex, multifaceted concept, and Human-AI trust, being a recent research avenue, faces several challenges. On the theoretical level, the difference between trust and other related theoretical concepts (e.g. reliance, compliance, and trustworthiness) needs to be understood as well as the factors affecting Human-AI trust. On the methodological level, trust is difficult to assess, and appropriate protocols have to be understood.

In this thesis, I tackle these challenges empirically through two lenses - academia and industry. I first conduct a systematic literature review of empirical studies on Human-AI trust in the context of decision making to get an overview of how trust is defined and evaluated in academia. However, as most studies are focusing on users' trust investigated in the controlled lab setting with AI mock-ups, I go further to investigate to which extent these findings hold true for other stakeholders with AI-embedded systems deployed in the market. To do so, I conduct a series of semi-structured interviews on the topic of Human-AI trust definitions and evaluation with people who develop and design AI-embedded systems assisting decision making and with people who are affected by these decisions.

I argue that theoretical understanding of Human-AI trust directly affects experimental protocol and measures choices. Drawing from the social sciences literature, I propose guidelines on improving experimental protocols for studying Human-AI trust in the context of decision making. I also demonstrate that discussing theoretical concepts, such as Human-AI trust, with laypeople of different backgrounds not only can validate the academic theories, but also potentially contribute to theoretical advancement. Lastly, I provide an overview of factors that can affect Human-AI trust in the context of decision making and, based on the comparison between the findings of academia and industry, I highlight research opportunities and design implications for academic researchers and AI practitioners.

This thesis provides theoretical and empirical evidence on Human-AI trust in the context of decision making and opens the ways to support trust in Human-AI interaction.

**Keywords:** *Human-AI trust, decision making, systematic review, semi-structured interviews, industry, experimental protocol*



# Résumé

Avec l'essor des systèmes d'aide à la décision intégrant l'intelligence artificielle (AI) dans le domaine médical, judiciaire ou du recrutement, la confiance entre l'humain et l'IA est devenue une priorité dans la conception de ces systèmes. De nombreux gouvernements et grandes entreprises ainsi que des chercheurs proposent diverses stratégies pour favoriser la confiance dans les systèmes intégrant l'IA. Cependant, la confiance est un concept complexe et multidimensionnel, et la confiance entre l'humain et l'IA, qui est un sujet de recherche récent, fait face à plusieurs défis. Sur le plan théorique, la différence entre la confiance et d'autres concepts théoriques proches (par exemple, la conformité) doit être comprise, de même que les facteurs affectant la confiance entre l'humain et l'IA. Sur le plan méthodologique, la confiance est difficile à évaluer, et il est nécessaire de définir des protocoles appropriés.

Dans cette thèse, je traite de ces défis de manière empirique à travers deux perspectives : académique et industrielle. J'effectue d'abord une revue systématique de la littérature des études empiriques sur la confiance entre l'humain et l'IA dans le contexte de la prise de décision afin d'obtenir une vue globale de la façon dont la confiance est définie et évaluée dans le monde académique. Cependant, comme la plupart de ces études sont en laboratoire et avec des maquettes d'IA, je poursuis cette analyse pour savoir dans quelle mesure ces résultats sont valables sur le terrain avec des vrais systèmes intégrant l'IA et différentes parties prenantes. Pour cela, je mène une série d'entretiens semi-structurés autour de la définition et de l'évaluation de la confiance entre l'humain et l'IA. Les participants sont soit des personnes qui développent ou conçoivent des systèmes intégrant l'IA pour l'aide à la prise de décision ou bien des personnes qui sont affectées par ces décisions.

Je soutiens que la compréhension théorique de la confiance entre l'humain et l'IA influence directement le choix des protocoles expérimentaux et des mesures utilisés pour les études empiriques. En m'inspirant de la littérature en sciences sociales, je propose des recommandations pour améliorer ces protocoles expérimentaux. Je démontre également que la discussion de concepts théoriques, tels que la confiance entre l'humain et l'IA, avec des personnes ordinaires de différents profils peut non seulement valider les théories académiques, mais aussi contribuer à l'avancement de la théorie. Enfin, je donne un aperçu des facteurs qui peuvent affecter la confiance entre l'humain et l'IA dans le contexte de la prise de décision. En comparant les résultats provenant du monde académique avec ceux provenant de l'industrie, je souligne les opportunités pour la recherche pour les chercheurs académiques et des implications pour la conception

pour les professionnels de l'IA.

Cette thèse fournit des preuves théoriques et empiriques sur la confiance entre l'humain et l'IA dans le contexte de la prise de décision et ouvre des voies pour promouvoir la confiance dans l'interaction entre l'humain et l'IA.

**Mots-clés** : *Confiance humain-IA, prise de décision, revue systématique, entretiens semi-structurés, industrie, protocole expérimental*

## Acknowledgments

Дякую. Merci. Thank you. These words of gratitude go to everyone whom I have crossed paths with and who made me smile at least once throughout this 3-year research journey.

Gilles and Baptiste, my supervisors, I have been **so** lucky to work with you. Thank you for being open-minded researchers, supportive and present mentors, and most importantly, “simply” understanding and cool human beings. I am forever grateful for this experience, it would not have been possible without you.

My HCI Sorbonne colleagues, those who are still at ISIR and not anymore, thank you for creating the chill, wholesome environment to work and to have fun. Special mentions go to the Ho5 gang for being awesome partners-in-research-and-procrastination; to Katerina for contagious excitement about research (best of luck with your PhD, cannot wait to see more of your work!); to Nacho and Reyhaneh for making me feel welcome and included when I used to be a shy intern trying to navigate a new research domain. The most special mention is reserved, of course, for my PhD sis Clara The One and Only for the *special type* of energy she brings into my life. You’ve been there since the legendary clean-up till 24 hours before the manuscript submission. Go finish your PhD as well, and let’s go to Hawaii???

I thank the entire ISIR, SCAI, and CAPSULE communities for kindness, support, and availability. I am glad to have had a unique opportunity to be a part of these communities and I will always look back fondly on my time here.

My work has also been greatly affected by the people I had a chance to meet at CSCW and CHI conferences (online and offline). A special shout-out goes to Jacob Browne for his unconditional support, curiosity, and wholesomeness. To anyone reading this, do check out his research. I am looking forward to learning more about Human-AI trust from you!

My research would not be possible without all the participants who agreed to be interviewed and to take part in the experiments, the reviewers and examiners who challenged it and provided precious feedback, and Sci-Hub which removed the barriers in the way of science.

My friends, in France and elsewhere, thank you for the sweet memories consisting of board games, picnics, museum visits, various food experiences, dancing, theatre, diving, occasional hikes, random visits, simply sharing funny memes or checking in when it is most needed. Garric fam, Theatre fam, Rome fam, and UA fam, can’t wait to celebrate with you all!

My family, thank you for your constant love and support, even at times when you needed it more than me, so strong that thousands of kilometers between us feel non-existent. I am sure that we will celebrate this achievement as well as other victories together very soon.





# Contents

<b>List of Figures</b>	<b>xiii</b>
<b>List of Tables</b>	<b>xiv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Context . . . . .	1
1.1.1 Trust . . . . .	3
1.1.2 Taxonomy of Decision Support Systems . . . . .	5
1.2 Problem Statement . . . . .	8
1.3 Research Approach . . . . .	9
1.4 Research Methods . . . . .	10
1.5 Contributions of the Research . . . . .	11
1.6 Overview of the Thesis . . . . .	13
1.7 Publications and Collaborations . . . . .	14
<b>2 How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies</b>	<b>17</b>
2.1 Objectives and Approach . . . . .	18
2.2 Research in Human-AI Interaction . . . . .	19
2.2.1 Empirical Research on Trust in AI . . . . .	19
2.2.2 Human-AI Guidelines . . . . .	20
2.2.3 Multidisciplinary Constructs in AI . . . . .	20
2.3 Systematic Review Methodology . . . . .	21
2.3.1 Keywords Identification . . . . .	21
2.3.2 Selection Criteria . . . . .	22
2.3.3 Papers Selection . . . . .	23
2.3.4 Corpus Overview . . . . .	24
2.3.5 Corpus Analysis . . . . .	24
2.3.6 Review Structure and Summary . . . . .	25
2.3.7 A Practical Example . . . . .	25
2.4 Trust Definitions . . . . .	27
2.4.1 Definitions in Human-AI Trust . . . . .	27

2.4.2	Elements of Trust . . . . .	28
2.4.3	Constructs Related to Trust . . . . .	31
2.5	Participants . . . . .	33
2.5.1	Experience and Expertise . . . . .	33
2.5.2	Groups and Stakeholders . . . . .	34
2.6	Task . . . . .	35
2.6.1	Interaction flow in the Decision Making Process .	36
2.6.2	Feedback . . . . .	37
2.6.3	Task Outcomes . . . . .	38
2.7	Procedure and Design . . . . .	40
2.7.1	Introducing the System Performance . . . . .	40
2.7.2	Experimental Design . . . . .	41
2.7.3	Assessing Pre-, Post-, or Dynamic Trust . . . . .	41
2.8	Summary of Findings for Experimental Protocol . . . . .	42
2.9	Quantitative Measures . . . . .	44
2.9.1	Questionnaires . . . . .	44
2.9.2	Trust-related Behavioral Measures . . . . .	48
2.10	Qualitative Methods . . . . .	51
2.10.1	Non-retrospective Methods . . . . .	51
2.10.2	Retrospective Methods . . . . .	52
2.10.3	Analyzing Qualitative Data . . . . .	54
2.11	Summary of Findings for Trust Measures . . . . .	56
2.12	Discussion . . . . .	57
2.12.1	Main Findings and guidelines . . . . .	58
2.12.2	Challenges and Research Opportunities . . . . .	59
2.12.3	AI in AI-based decision-making systems . . . . .	60
2.13	Conclusion . . . . .	61
<b>3</b>	<b>Human-AI Trust in the Context of Decision Making through the Lens of AI Practitioners and Decision Subjects</b>	<b>63</b>
3.1	Objectives and Approach . . . . .	64
3.2	Related Work . . . . .	65
3.2.1	Human-AI Trust . . . . .	65
3.2.2	Stakeholders of AI-embedded Systems for Deci- sion Making . . . . .	66
3.3	Methodology . . . . .	68
3.3.1	Participants . . . . .	69
3.3.2	Interview Protocol . . . . .	70
3.3.3	Analysis of the Interviews and Comparison with the Academic Literature . . . . .	71
3.3.4	Result presentation . . . . .	72
3.4	Trust and Trustworthiness in Human-AI Interaction . . .	72
3.4.1	Key Elements of Human-AI Trust . . . . .	72
3.4.2	Key elements of AI Trustworthiness . . . . .	75

3.5	Trust Factors Related to the Socio-Technological Context	76
3.5.1	Human-Human Trust . . . . .	77
3.5.2	Time Dynamics . . . . .	78
3.5.3	Type of Task . . . . .	79
3.5.4	Marketing . . . . .	80
3.5.5	Summary . . . . .	81
3.6	Trust Factors Related to the Systems' Development and Design . . . . .	82
3.6.1	Performance . . . . .	83
3.6.2	Transparency . . . . .	84
3.6.3	Interactivity . . . . .	86
3.6.4	AI Certification . . . . .	87
3.6.5	Summary . . . . .	88
3.7	Trust Factor Related to People's Preferences and Experiences . . . . .	89
3.7.1	Agency . . . . .	90
3.7.2	Expectations about AI Recommendations . . . . .	91
3.7.3	AI Literacy . . . . .	92
3.7.4	Domain Expertise . . . . .	93
3.7.5	Summary . . . . .	94
3.8	General Discussion . . . . .	94
3.8.1	Discussing Our Research Questions . . . . .	95
3.8.2	Future Work Directions . . . . .	96
<b>4</b>	<b>Discussion, Future Perspectives, and Conclusions</b>	<b>99</b>
4.1	Progress on Research Problems . . . . .	99
4.1.1	RQ1 What differentiates Human-AI trust from other related constructs, such as reliance, compliance, trustworthiness, etc.? . . . .	99
4.1.2	RQ2 How to evaluate trust in the context of decision making? . . . . .	100
4.1.3	RQ3 What factors affect Human-AI trust in the context of decision making? . . . . .	101
4.1.4	RQ4 Do the academic postulations about trust definition, factors, and evaluation of Human-AI trust reflect the real world considerations? . . . . .	102
4.1.5	Pieces of Trust Puzzle Brought Together . . . . .	103
4.2	Scientific Contributions . . . . .	104
4.3	Further Perspectives . . . . .	105
4.3.1	Short-term . . . . .	106
4.3.2	Mid- and Long-term . . . . .	109
4.4	Conclusion . . . . .	110
<b>A</b>	<b>Trust Definitions</b>	<b>113</b>

<b>B</b>	<b>Selected Human-Human Trust Questionnaires</b>	<b>115</b>
B.1	Behavioral Trust Inventory [Gillespie, 2003] . . . . .	116
B.2	Trust Questionnaire [Currall and Judge, 1995] . . . . .	117
B.3	Trust for Management Questionnaire [Mayer, 1999] . . . . .	118
<b>C</b>	<b>Trust Questionnaires Used in Human-AI Literature</b>	<b>119</b>
C.1	Human Trust in Automation Scale [Jian et al., 2000] . . . . .	119
C.2	Human-Robot Trust Questionnaire [Schaefer, 2013] . . . . .	120
C.3	Trust in Management Questionnaire [Mayer, 1999] . . . . .	121
C.4	Trust in Automation [Muir, 1989] . . . . .	121
C.5	Trust in Teammate [Ross, 2008] . . . . .	121
C.6	Human-Computer Trust Scale (HCT) [Madsen and Gregor, 2000] . . . . .	122
C.7	Trust in Automation [Chien et al., 2018] . . . . .	123
C.8	Semantic Pairs for Credibility [Ohanian, 1990] . . . . .	124
C.9	Trust in Automation Questionnaire [Merritt, 2011] . . . . .	124
C.10	Pedestrian Receptivity Questionnaire [Deb et al., 2017] . . . . .	125
C.11	Trust in E-Commerce [McKnight et al., 2002] . . . . .	126
<b>D</b>	<b>Interview Questions for AI Practitioners</b>	<b>129</b>
D.0.1	Background . . . . .	129
D.0.2	Understanding of Human-AI Trust . . . . .	129
D.0.3	Evaluation of Human-AI Trust . . . . .	130
<b>E</b>	<b>Interview Questions for Decision Subjects</b>	<b>131</b>
E.0.1	Experiences with Human-AI Decision Making . . . . .	131
E.0.2	Understanding of Human-AI Trust . . . . .	132
E.0.3	Opinions about Industry's Efforts towards Trust-worthy AI . . . . .	132
<b>F</b>	<b>Systematic selection of empirical studies of Human-AI trust in the decision making context</b>	<b>133</b>
<b>G</b>	<b>Summary of all the discussed Human-AI trust factors in the context of decision making</b>	<b>135</b>

## *List of Figures*

1.1	Examples of AI-embedded systems assisting decision making in the thesis' scope.	6
1.2	Schematic representation of the decision-making scenarios in the thesis' scope.	7
1.3	Problematics explored in this thesis.	10
2.1	Research questions addressed in Chapter 2.	18
2.2	Papers search and selection process.	22
2.3	Distribution of the reviewed papers across the publishing venues and years.	24
2.4	Key elements of trust and related theoretical constructs.	32
2.5	4 types of decision-making flows and measures associated to them.	36
2.6	Summary of empirical methods used in the reviewed papers.	45
3.1	Research questions addressed in Chapter 3.	64
4.1	Summary of the principal findings.	100
4.2	Reminder: problematics explored in this thesis .	103



## *List of Tables*

- 2.1 16 guidelines and 9 research opportunities regarding Human-AI trust evaluation. 26
- 2.2 List of trust definitions in Human-AI papers. 29
- 3.1 Three types of stakeholders most tightly linked to Human-AI decision making. 67
- 3.2 Overview of the decision subjects' profiles. 69
- 3.3 Overview of the AI practitioners' profiles. 70
- 3.4 The main questions of the semi-structured interviews, both for AI practitioners and decision subjects. 71
- 3.5 Comparison of trust and trustworthiness definitions as given by the interviewees and the literature. 73
- 3.6 Human-AI trust factors related to socio-technological context. 76
- 3.7 Types of Human-Human trust discussed in relation to Human-AI trust by the interviewees and the literature. 77
- 3.8 Human-AI trust factors related to systems' design and development. 82
- 3.9 Human-AI trust factors related to people's preferences and experiences. 89





*“Never trust anything that can think for itself if you can’t see where it keeps its brain.”*

— J.K. Rowling, *Harry Potter and the Chamber of Secrets*



# 1

## *Introduction*

### 1.1 Context

Everyday humans are involved in **decision making**, whether they notice it or not: when to leave to arrive on time, what to buy, what job to accept. However, we rarely make decisions alone and turn decision making into a social process by soliciting advice from other people, especially when it comes to difficult decisions [Sniezek and Van Swol, 2001; Van Swol and Sniezek, 2005]. When receiving advice from other people, one can enter in a conflict between their initial opinion and the newly received information when they differ: which one is more pertinent to make a better decision [Yaniv and Kleinberger, 2000]? As there is often uncertainty around the quality of the given advice [Sniezek and Van Swol, 2001], researchers showed that one's **trust** in the advice is one of the important contributors that catalyzes advice taking. The main reason is that it serves as a mental shortcut to resolve the conflict between the opinions (personal and the one of the advisor) [Van Swol and Sniezek, 2005; Wang and Du, 2018].

Technology can also help people in making decisions. There exists a class of systems designed with this purpose – decision support system (DSS) [Bertl et al., 2022; Madhavan and Wiegmann, 2007]. These systems can aggregate information from numerous sources, assist in organizing and analyzing it [Bertl et al., 2022], often surpassing human capabilities and speed, which makes them especially advantageous in high-stake scenarios [Madhavan and Wiegmann, 2007]. Since

these systems can provide recommendations, one might enter in the same conflict between their own opinion and newly received information, but this time it is offered by a system, rather than a human. However, as humans have tendency to view technological systems as social actors, they can form “relationships” with systems and experience emotions and attitudes just like they would with other humans [Cannon-Bowers and Salas, 1998; Reeves and Nass, 1996]. Therefore, human trust in a DSS and its recommendations is as equally important for decision making as with human advice [Madhavan and Wiegmann, 2007].

Traditional DSS, however, has several limitations. First of all, they demonstrate limited performance for ill-structured problems [Er, 1988; Vohra and Das, 2011], that is problems that evoke a highly variable set of solutions with no criterion to determine which one is correct or false, for example, developing a new marketing strategy [Cats-Baril and Huber, 1987; Voss and Post, 1988]. This is because data capture and collection has been proven challenging for traditional DSS [Er, 1988; Vohra and Das, 2011], especially in the decision domains where new evidence is produced at a rapid pace like medicine [Lagioia and Contissa, 2020]. Therefore, such systems are also not capable of adapting to unknown, new situations [Phillips-Wren, 2013].

To overcome these limitations, a newer generation of DSS now embed Artificial Intelligence (AI) to benefit from powerful computing tools to better aggregate, integrate, manage and analyze big, complex data [Bertl et al., 2022; Gupta et al., 2022; He and Li, 2017; Phillips-Wren, 2013]. In this thesis, I refer to such systems as **AI-embedded systems assisting decision making** (I will detail their different types in Section 1.1.2). While there is no universally accepted definition of AI [Duan et al., 2019], in this thesis, I 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*” [Samoili et al., 2020]. DSS can embed a diverse set of different AI techniques at once [Lagioia and Contissa, 2020], which lay on a spectrum between data-driven AI and knowledge-based AI (as classified by [Mattioli et al., 2022]): artificial neural networks, genetic algorithms, decision trees, fuzzy logic techniques, to name a few [Aljaaf et al., 2015; Phillips-Wren, 2013].

In contrast to traditional DSS, AI-embedded systems assisting decision making are more adapted to solve unstructured, ambiguous problems,

respond appropriately and timely to a new situation, aggregate and learn from new data and past interactions [Phillips-Wren, 2013]. Because of these capabilities, they have become more widespread in high-stake domains, where the problems can be unique and non-trivial [Lagioia and Contissa, 2020] and decisions have real impacts on people's lives, such as public safety [Kalhan, 2013], hiring [Ajunwa et al., 2016] or loan approval [O'Dwyer, 2018]. AI-embedded systems assisting decision making, however, bring new challenges to the decision makers. One of them is that they are a "black box," that is it is difficult to understand how a system arrived to a certain conclusion [Adadi and Berrada, 2018; Lagioia and Contissa, 2020]. It is because, in contrast to some traditional DSS, developers do not program their interpretation of the structure of the decision to make. Instead, they develop an architecture that "connects the dots" and makes its own model from a large quantity of data (in deep learning, for example) [London, 2019]. This, in turn, obfuscates understanding why a certain AI recommendation was produced, anticipation of potential biases in decision making, and identification of the reasons for wrong predictions [Scherer, 2015; Yu and Kohane, 2019]. As this prevents users from adequately evaluating the quality AI recommendations and solving the conflict between their own opinion and what AI suggests, the topic of human trust in AI-embedded systems assisting decision making is ever so important. International institutions (European Commission [2020], G20 [2019]) and governments (USA [Defense Innovation Board, 2019; White House Office, 2020], Estonia [AI Taskforce, 2019], or France [Villani et al., 2018]) have highlighted the need for considering trust in the design of AI, because it plays an important role in the adoption of these technologies [Hoff and Bashir, 2015] and the improvement of decision making [Bansal et al., 2019]. In the private sectors, companies such as AXA Research Fund [2019], Accenture Federal Services [2019], or KPMG [2019] are also taking this path of research in order to foster trust by going beyond system's accuracy and tackling the issue of black-box and non-deterministic nature of AI through promoting privacy, security, algorithm accountability and transparency. Thus, designing and ensuring peoples' trust in AI has raised interest in the Human-Computer Interaction (HCI), and it is especially crucial for the context of decision making.

### 1.1.1 Trust

Trust affects many spheres of our lives, and the proof of this is the numerous fields that studied it: Philosophy [Baier, 1986; Lagerspetz, 2010], Psychology [Colquitt et al., 2011; Simpson, 2007], Sociology [Gambetta and Gambetta, 2000; Misztal, 1996], Economics [Akerlof,

1970; Braynov, 2002], Management [Fulmer and Gelfand, 2012; Zheng et al., 2008], Human-Computer Interaction [Hoff and Bashir, 2015; Lee and See, 2004; Muir, 1994], to name a few. Consequently, trust can be directed towards different types of entities: physical individuals, online individuals, an organization, and a piece of technology. In this thesis when talking about trust in AI, I adopt the definition of trust in automation: *“An attitude that an agent will achieve an individual’s goal in a situation characterized by uncertainty and vulnerability”* [Lee and See, 2004]. Currently, there is no definition of trust derived specifically for AI. Chapter 2 provides more details about different definitions of trust from various research domains used in Human-AI interaction and why I decide to favor this one. Note that some definitions propose to further distinguish between affect-based trust and cognition-based trust [McAllister, 1995] or unquestioned and calculated trust [Markova and Gillespie, 2008], but in this thesis, I look at trust in AI in more general terms.

Literature on trust has two predominant axes: understanding the conditions necessary for trust to exist and exploring what factors affect levels of trust once it is established. For one’s trust in an entity to exist, the most common condition is for this entity to be trustworthy [Schoorman et al., 2007]. This condition is broken down in three main elements: ability (possession of relevant and sufficient competence to provide a good recommendation), benevolence (being well-meaning to the one who trusts), and integrity (adherence to the values and principles acceptable by the one who trusts) [Schoorman et al., 2007]. In this thesis, I also explore the conditions necessary for trust in AI to exist. However, instead of focusing on the qualities of an entrusted entity, I aim to highlight the elements that differentiate trust from other theoretical concepts such as confidence, distrust, reliance, compliance, etc. Therefore, I turn my attention to the elements from the environment that can trigger trust on a cognitive level, which I review in Chapter 2.

For trust levels to change, factors related to the person itself, the system, and the context come in play [Adams et al., 2003; Bindewald et al., 2018; Hancock et al., 2011; Hoff and Bashir, 2015; Schaefer et al., 2014, 2016]. Generally, the most studied factors are related to the systems’ performance [Hancock et al., 2011], for example, its accuracy and errors, followed by the ones related to the context, e.g. task difficulty, and then by the user-associated ones, e.g. domain expertise [Hancock et al., 2011]. Since trust in AI is a relatively young field of research, there is only two literature reviews that summarize and categorize factors that influence trust in AI [Browne et al., 2022; Glikson and Woolley, 2020]. As Glikson and Woolley [2020] consider trust in AI in a general context

and Browne et al. [2022] look at the factors pertinent to the medical decision making, I complete their reviews with investigating factors that affect trust in AI in a larger decision making context in Chapter 3.

### 1.1.2 Taxonomy of Decision Support Systems

To better describe the AI-embedded systems assisting decision making I consider in this thesis, I will use the taxonomy of traditional DSS for I view such systems as a subset of DSS. As DSS can assist decision making in different ways, for various users, in numerous domains, there are many approaches to categorize them [Aqel et al., 2019; Power, 2002]: based on User Relationship (active, passive, cooperative) [Jelassi et al., 1987], Scope (personal, group, organizational) [Hackathorn and Keen, 1981], Specificity (custom-made vs vendor-ready-made) [Turban and Aronson, 1997], Type and Frequency of decision making (ad-hoc and institutional) [Donovan and Madnick, 1977], to name a few.

For my scope, I use the categorization of Power [2002], inspired by Alter [1976], that relies on the mode of assistance. Power [2002] identifies 5 main categories of DSS depending on how exactly they support decision-making process:

1. *Data-driven DSS*: tools that aid in analysis of large amount of structured data through accessing more detailed information, broader summary or change the viewed data dimensions [O'Brien and Marakas, 2007].
2. *Knowledge-driven DSS*: tools that provide recommendations via the knowledge stored as rules, relationships or probabilities [O'Brien and Marakas, 2007].
3. *Model-driven DSS*: tools that provide access to a model and allows for its manipulation.
4. *Document-driven DSS*: tools that gather, retrieve, classify, and manage unstructured documents.
5. *Communications-driven DSS*: tools that facilitate communication and collaboration between humans.

In this thesis, I focus on data- and knowledge-driven AI-embedded systems. I also note that I differentiate DSS from a recommendation system [Liang, 2008]. A recommendation system provides recommendations through analyzing previous users' behaviors and can



infer their interests and preferences, e.g. a movie recommendation on Netflix. While recommendation systems can assist in decision making, e.g. choosing a movie to watch or an item to purchase, I focus on the decision-making processes that are not based on users' preferences. It is because I am interested in scenarios that have considerable implications on someone's life (health, financial security) as I believe the issue of trust is particularly pertinent here. I also highlight that I do not consider systems that belong to "decision *making* technologies" [Stohr and Viswanathan, 1999], which automate decision making replacing users rather than assisting them.

To be even more concrete, here are the most representative examples of the AI-embedded systems I consider in this thesis:



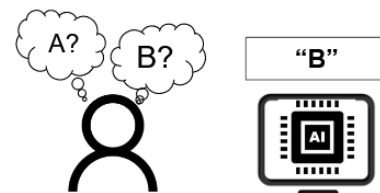
**Figure 1.1:** Schematic representations of existing AI-embedded systems assisting decision making for simplification purposes that in the scope of the thesis. a) IBM Watson for Oncology, AI-embedded system for cancer treatment; b) Hirevue, AI-embedded system for talent acquisition; c) Stockbot, AI-embedded systems for financial investments.

- AI-embedded system for cancer treatment, one of the examples is IBM Watson for Oncology (Figure 1.1-a). The system uses natural language processing and a variety of search techniques to analyze both structured and unstructured medical data- and knowledge-bases to produce a list of suggested cancer treatment for a specific patient. The list has ranking based on the confidence scores [Horizon Cancer Center, 2015; Lagioia and Contissa, 2020; STATNews, 2017] (classification task). The system is developed by IBM Watson, purchased by hospitals who are clients, and the principal users are medical doctors (domain experts) whose decisions based on AI recommendations will affect patients. Even though the system showed higher accuracy rates than clinicians, after 4 years of development and deployment and investment of \$62 million, the doctors in MD Anderson Cancer Center in Houston, USA, stopped using its recommendations completely [Lohr, 2021; Ross and Swetlitz, 2017]. Lack of doctors' trust in AI recommendations was among the reasons for its abandonment, because the doctors could not understand how it derived these recommendations and were not ready to take upon the responsibility in case AI makes a mistake [Lagioia and Contissa, 2020].

- AI-embedded system for talent acquisition, one of the examples is Hirevue<sup>1</sup> (Figure 1.1-b). The candidates upload their video replies, and the system's AI analyzes their tone of voice, used words, and facial expression to evaluate and rank them comparing their performance to the one of the actual employees of a company (classification task). Hirevue develops the system, sells their solution to their clients (companies), where the main users are HR team whose hiring decisions based on AI recommendations affect the perspective candidates. Such systems usually raise issues of trust due to the lack of transparency of the selection process, potential biases (see Amazon case [Anonymous, 2016]), and claims that AI is not adept at tasks that require social and empathetic skills [Figueroa-Armijos et al., 2022; Hunkenschroer and Kriebitz, 2022].
- AI-embedded systems for financial investments, one of the examples is Stockbot [Mohanty et al., 2022]. Using a prediction model based on an artificial neural network (long short-term memory, LSTR) trained on the past prices of a stock, the system predicts future stock prices and provides a recommendation on whether to buy or to sell a certain stock (regression and classification tasks) [Ghorbel, 2022]. While this specific system is not deployed in the market, usually it is a company that develops such systems and sells their solution directly to users, who might not necessarily be domain experts. The decisions they make based on AI recommendations typically affect their own losses and revenues. This direct consequence could be one of the sources of trust issues [Burke and Hung, 2021] alongside with the fact that on a more global scale algorithmic trading can contribute to financial instability [Arena et al., 2018].

In a nutshell, I do not focus on any specific decision domain, a particular task (classification or regression), type of user (domain expert or not) nor a set of stakeholders (e.g. only users are affected by AI recommendations or someone else who does not interact with AI directly). The scenarios I can consider can be summarized as the following represented in Figure 1.2: a user has to make a decision that can have substantial consequences for them or someone else, AI-embedded system provides information useful for the decision, and the user has the last say in the decision.

<sup>1</sup> <https://www.hirevue.com/>



**Figure 1.2:** Schematic representation of the decision-making scenarios I consider in the thesis: a user has to make a decision that can have substantial consequences for them or someone else, AI-embedded system provides information useful for the decision, and the user has the last say in the decision.

## 1.2 Problem Statement

When I was planning to run an experiment to investigate how one factor affects human trust in AI in the context of decision making, I ran into several methodological challenges. First of all, I did not know what is the best method to measure trust: through 1-item questionnaires, multi-item questionnaires or behavioral measures. Additionally, each method included a plethora of measurement tools, which further complicated the choice. After a brief literature review, I realized that the Human-AI Trust community does not follow a standard experimental protocol on which I could rely to run my study. I have also noticed a theoretical confusion in terminology between trust, reliance, confidence, and trustworthiness in the research articles. High level policy reports and guidelines promoting the idea of building AI-embedded systems that people can trust also often remained vague about what Human-AI trust means.

I argue that to efficiently develop and design AI-embedded systems people can trust in the context of decision making, we need to have a strong foundational understanding of this concept [Gille et al., 2020]. Therefore, in this thesis, the question I address is the following: *What is Human-AI trust in the context of decision making?* This question can further be broken down into four research questions:

**RQ 1** *What differentiates Human-AI trust from other related constructs, such as reliance, compliance, trustworthiness, etc.?*

To disentangle trust from other related concepts, one must identify the key elements that contribute to trust existence, also sometimes called conditions or prerequisites for trust to form. My aim in this thesis to highlight the ones that differentiate trust from reliance, distrust, and confidence. Understanding what key elements differentiate trust from these concepts has a direct impact on the choice of experimental protocol and measures of trust.

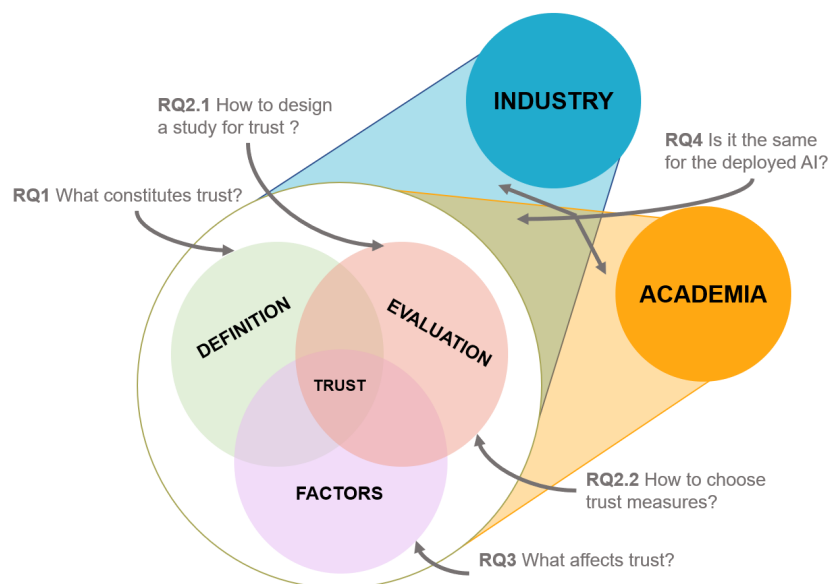
Note that I differentiate key elements of trust from trust factors. The former are what trigger trust mechanism on a cognitive level, and the latter are what change the levels of trust. For example, high accuracy of AI recommendations can increase trust in AI, but trust in AI can exist even if high accuracy is absent. Hence, it is a trust factor, but not a key element.

- RQ 2** *How to evaluate Human-AI trust in the context of decision making?*  
 To evaluate a concept, one has to design an appropriate experimental protocol as well as to choose fitting measures. Consequently, this implies two sub-challenges:
- RQ 2.1** What do we have to include in the experimental protocol for Human-AI trust to exist in the context of decision making?
  - RQ 2.2** How to choose the appropriate measures for Human-AI trust in the context of decision making?
- RQ 3** *What factors affect Human-AI trust in the context of decision making?*  
 Trust factors is what affects trust level, either positively or negatively. Having a comprehensive overview of these factors can have design and governance implications.
- RQ 4** *Do the academic postulations about trust definition, factors, and evaluation of Human-AI trust reflect the real world considerations?*  
 Most of our knowledge about Human-AI trust comes from academic findings, which focus on the perspectives of users and investigate their interactions with AI mock-ups. Thus, we know little about whether Human-AI trust has the same meaning in the real-world context, that is for the stakeholders other than users that exist in the Human-AI decision making ecosystem and for the systems deployed in the market. I refer to the latter as “industry” perspective.

### 1.3 Research Approach

In this thesis, my goal is to understand how both academia and industry view and evaluate Human-AI trust in the context of decision making. I approach is thus two-fold: theoretical and empirical.

Firstly, I build on social and cognitive sciences to understand how to define and evaluate trust. I transpose this knowledge and compare with the current overview of the literature on Human-AI trust in the context of decision making. It is, thus, possible to, first of all, disentangle Human-AI trust from other related theoretical concepts (RQ 1). It also lets me to identify shortcomings in the empirical methodologies



**Figure 1.3:** Problematics explored in this thesis.

used in Human-AI Interaction community (RQ 2). Lastly, I conduct the first review of factors that affect Human-AI trust exclusively in the decision-making context.

Secondly, to compare the academic findings about trust definitions, factors and evaluations with the current views from the industry, I empirically explore people's reflections through interviews. Specifically, I am interested in the reflections of AI practitioners, people who develop, design, and deploy these systems, and decision subjects, people who do not interact with AI, but are affected by AI recommendations. This provides me with a perspective on Human-AI trust that does not focus just on users and comprises the influence of the socio-technical context the systems are deployed in. Through a comparison between the findings from academia derived through the theoretical approach, I identify research opportunities - theoretical, empirical, and design - related to human trust and decision making unexplored in Human-AI Interaction community. I also highlight the aspects of Human-AI trust well explored in academia that the industry can benefit from.

## 1.4 Research Methods

My work includes these research methods:

- **Systematic Review.** Called "a study of studies" [Institute for Qual-

ity and Efficiency in Health Care, 2016], a systematic review provides a very detailed, often quantified summary of the articles selected on a specific topic [Okoli, 2015]. It allows to have a comprehensive overview of the state of a research field, including trends in the topics studied, methodologies used, and results. I conducted a systematic review to understand how trust is defined and evaluated in Human-AI Interaction community presented in Chapters 2 and 3.

- **Theoretical Literature Review.** I reviewed social and cognitive sciences literature on methodology of studying human-human trust. It allowed me to identify shortcomings in the emerging methodological trends in Human-AI Interaction community and to propose guidelines to overcome them presented in Chapter 2. The literature review also shed the light on the research opportunities to be explored yet.
- **Semi-structured Interviews.** Semi-structured interviews is a type of interviews where an interviewer investigates a set of pre-defined topics, but at the same time is free to change their order and to introduce follow-up questions to deepen the discussion [Adams, 2015; Whiting, 2008]. I used this method to explore whether AI practitioners consider Human-AI trust in their working practices and to study decision subjects experiences and needs with Human-AI decision making presented in Chapter 3. Obtaining this information with semi-structured interviews provided insights on how AI practitioners and decision subjects define Human-AI trust with their own terms and what they think about trust evaluation.
- **Thematic Analysis.** Thematic analysis is a method to analyze qualitative data and identify meaningful patterns in it [Clarke and Braun, 2013]. It is an iterative process where researchers assign codes to the (transcribed) text and group them into larger themes. The themes are usually the important aspects of the studied phenomenon. I used thematic analysis to interpret the semi-structured interviews of AI practitioners and decision subjects in Chapter 3.

## 1.5 Contributions of the Research

This thesis contains 4 types of contributions: *methodological*, *survey*, *theoretical*, and *empirical* as categorized by Wobbrock and Kientz [2016]. Methodological contributions inform researchers how to conduct their investigations. Survey contributions summarize academic work on a

research topic and describe existing trends and gaps. Theoretical contributions yield new or improved definitions, frameworks or models. Empirical contributions produce a new piece of knowledge based on collected data [Wobbrock and Kientz, 2016].

- Methodological
  - *Guidelines for Empirical Protocols and Trust Measures.* I provide 8 guidelines to standardize the design of empirical protocol for studies on Human-AI trust in the context of decision making, emphasising the importance of integrating elements from trust definition. I also provide 6 guidelines to facilitate the choice and use of trust measures, emphasising that the nature of trust. This contribution is presented in Chapter 2 and addresses RQ2.
- Survey
  - *Landscape of Current Trends in Protocols and Measures for Human-AI Trust Evaluation.* As an outcome of the systematic literature review, I summarize and categorize the existing protocol choices per each standard section of an empirical protocol in the current studies on Human-AI trust in the context of decision making. I do the same for the existing qualitative and quantitative trust measures. This way the community has a compact overview of all the possible ways to design their studies and a straightforward access to the repertoire of trust measures used in the research community. This contribution is presented in Chapter 2 and addresses RQ2.
  - *Landscape of Factors that Affect Human-AI Trust in the Context of Decision Making.* As an outcome of the systematic literature review, I provide a structured overview of all the trust factors considered in the studies on Human-AI trust in the context of decision making, organizing them in three groups. This contribution is presented in Chapter 3 and addresses RQ3.
- Theoretical
  - *Difference between Trust and Related Concepts.* I highlight the key elements that differentiate trust from confidence, distrust, reliance, compliance, and trustworthiness. This contribution is presented in Chapter 2 and addresses RQ1.
  - *Research and Design Implications around Human-AI Trust Factors.* I

identify the research and design opportunities for academic researchers and AI practitioners. For academic researchers, they are mostly related to the investigation of socio-contextual trust factors, e.g. trust between AI users and AI team, relative performance, social transparency, and surprising AI recommendations. For AI practitioners, they are related to the practitioners' design, development, and deployment practices, e.g. the way they communicate about their AI, types of AI explanations, and individual differences of users. This contribution is presented in Chapter 3 and addresses RQ3.

- Empirical
  - *Human-AI Trust Definition and Factors as Seen in the Industry*. As an outcome of the semi-structured interviews, I describe how AI practitioners and decision subjects define Human-AI trust with their own words. I also discover what they think can affect Human-AI trust in the context of decision making and which of these factors they consider the most important. This contribution is presented in Chapter 3 and addresses RQ4.

## 1.6 Overview of the Thesis

The main body of this thesis consists of two articles that tackle the above-identified research questions. The table below describes the link of each chapter with the research questions:

<b>Chapter 2</b>	<i>How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies</i>	
	In this chapter, I present a systematic literature review of definitions of Human-AI trust as well as empirical methodologies and measures used to evaluate it in the context of decision making. The first part of the chapter focuses on trust definitions, and specifically identifying the elements that differentiate trust from other trust-related theoretical concepts.	Academia: trust definition ( <b>RQ1</b> )
	The second part of the chapter focuses on summarizing and categorizing empirical methods and measures to evaluate Human-AI trust present in the reviewed papers. I also provide guidelines and research opportunities to standardize empirical protocols and to choose appropriate measures.	Academia: trust evaluation ( <b>RQ2</b> )



- Chapter 3** *Human-AI Trust in the Context of Decision Making through the Lens of AI Practitioners and Decision Subjects*  
 In this chapter, I present reflections of people from the industry, AI practitioners and decision subjects, on Human-AI trust definition and factors obtained with semi-structured interviews. I also compare these findings to the ones from the systematic literature review. In the first part of the chapter, I compare how the interviewees define trust and whether they differentiate it from other related theoretical concepts.  
 In the second part of the chapter, I compare what factors play an important role for Human-AI trust according to the interviewees versus the reviewed papers. I also provide research and design opportunities for academic researchers and AI practitioners.
- Academia and industry: trust definition (RQ3&4)
- Academia and industry: trust factors (RQ3&4)
- Chapter 4** *Discussion, Future Perspectives, and Conclusions*  
 This chapter summarizes the findings and the contributions of my thesis. I also propose the short-, mid- and long-term perspectives stemming from my work.
- Industry: trust evaluation (RQ4)

## 1.7 Publications and Collaborations

The content of this thesis is built on three articles:

- 🏆 **Honorable Mention Award.** Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. “How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies.” In *Proc. ACM Hum.-Comput. Interact.* 5 (CSCW2021), 39 pages. (hal-03280969v2)
- Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. “What AI Practitioners Say about Human-AI Trust: Its Role, Importance, and Factors That Affect It,” working paper. In *International Conference on Hybrid Human-Artificial Intelligence (HHAI2022)*, 8 pages.
- Oleksandra Vereschak, Fatemeh Alizadeh, Gilles Bailly, and Baptiste Caramiaux. “Human-AI Trust in the Context of Decision Making through the Lens of AI Practitioners and Decision Subjects.” *Under review*, submitted in September 2022, 22 pages.

This work also resulted in the participation and co-organization of the following workshops:

- Oleksandra Vereschak, Gilles Bailly, Baptiste Caramiaux. On the Way to Improving Experimental Protocols to Evaluate Users’ Trust

in AI-Assisted Clinical Decision Making. CHI'21 Workshop: Realizing AI in Healthcare: Challenges Appearing in the Wild, 2021. ⟨hal-03418706⟩

- Oleksandra Vereschak, Gilles Bailly, Baptiste Caramiaux. On the Way to Improving Experimental Protocols to Evaluate Users' Trust in AI-Assisted Decision Making. CHI'21 Workshop: Towards Explainable and Trustworthy Autonomous Physical Systems, 2021. ⟨hal-03418712⟩
- Fatemeh Alizadeh, Oleksandra Vereschak, Dominik Pins, Gunnar Stevens, Gilles Bailly, Baptise Caramiaux. Building Appropriate Trust in Human-AI Interactions. 20th European Conference on Computer-Supported Cooperative Work (ECSCW 2022), Jun 2022, Coimbra, Portugal. ⟨hal-03724018⟩



# 2

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

Empirical studies can advance understanding of what Human-AI trust is, what factors affect it and to which extent. As mentioned in Chapter 1, human trust is a complex concept to study due to its multifaceted and multidisciplinary nature and the theoretical confusion with other related constructs such as confidence, distrust, reliance, compliance, and trustworthiness. Therefore, constructing an experimental protocol and choosing appropriate trust measures can be a challenge. To address this, the literature does not yet provide guidelines that support the empirical study of Human-AI trust in the context of decision making.

In this chapter, we<sup>1</sup> focus on how to evaluate trust between human users and AI-embedded systems in decision making. Specifically, we investigate what differentiates trust from other theoretical constructs, what needs to be included and controlled for in the design of an experimental protocol and what trust measures to choose. We, thus, present a comprehensive survey of the Human-AI trust definitions and the experimental methodologies set to investigate it in the context of de-

<sup>1</sup>Main portion of this chapter was published in ACM Conference On Computer-Supported Cooperative Work And Social Computing [Vereschak et al., 2021]. Thus, any use of “we” in this chapter refers to the authors of this work: Oleksandra Vereschak, Gilles Bailly, Baptiste Caramiaux.

cision making. Through this systematic literature review, we aim to summarize and categorize the most common elements of trust definitions alongside with current empirical approaches in the research on Human-AI trust in the context of decision making. We are also set to identify good practices and potential caveats among these approaches. Finally, we draw guidelines and research opportunities for the empirical study of Human-AI trust in the context of decision making.

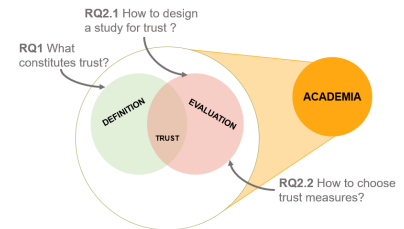
## 2.1 Objectives and Approach

This chapter addresses two research questions of this thesis: **DEFINITION-RQ1** *What differentiates Human-AI trust from other related constructs, such as reliance, compliance, trustworthiness, etc.?* and **EVALUATION-RQ2** *How to evaluate Human-AI trust in the context of decision making?* through the lens of **ACADEMIA** (Figure 2.1). This chapter shares 4 objectives. The first one stems from RQ1, targeting to disentangle trust from other theoretical constructs. We analyze the existing trust definitions presented in the empirical studies on Human-AI trust in decision making to find common elements between them. We present our findings in Section 2.4.

The second objective focuses on the design of the experimental protocol and is related to **RQ 2.1** of this thesis: *What do we have to include in the experimental protocol for Human-AI trust to exist in the context of decision making?* Having identified the key elements of trust in Section 2.4, we investigate to which extend they are incorporated and controlled for in the current methodological practices of studying Human-AI trust in the context of decision making through a systematic literature review. We summarize and categorize our findings following a typical structure of an experimental protocol design: Participants (Section 2.5), Task (Section 2.6), and Procedure and Design (Section 2.7).

The third objective focuses on the choice of measures and is related to **RQ 2.2** of this thesis: *How to choose the appropriate measures for Human-AI trust in the context of decision making?* Using the systematic literature review, we describe and categorize different trust measures used in the corpus, diving them in two groups - Quantitative (Section 2.9) and Qualitative (Section 2.10).

Finally, the forth objective is to provide a standardized guidance on the design of experimental protocols, choice of trust measures, and possible research directions; it encompasses RQ 2.1 and RQ 2.2. Draw-



**Figure 2.1:** Research questions investigated in this chapter.

ing from literature on social and cognitive sciences exploring Human-Human trust, we propose a set of guidelines to help researchers in the design of experimental protocols and in choice of the trust measures. The main purpose is to prevent the identified caveats in the study of trust in the specific context of AI-assisted decision making. In complement to guidelines, we identify research opportunities regarding the elaboration of practical methods to studying trust and its dynamics in laboratory experiments or the investigation of relevant factors (e.g. individual differences, task outcomes) on Human-AI trust.

This chapter is structured as follows. We first present the methodology behind the systematic review of empirical studies investigating Human-AI trust in the decision making context. We then present and discuss results alongside with guidelines and research opportunities for the experimental protocol and measures respectively. We start each results' subsection with a categorized summary of the methods used in the corpus, which we discuss in the light of social and cognitive sciences literature on Human-Human trust, highlighting strengths and limitations. Based on this, we provide guidelines and research opportunities stemming from our discussion, summarized in Table 2.1, Section 2.3.

## 2.2 Research in Human-AI Interaction

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

### 2.2.1 *Empirical Research on Trust in AI*

Many empirical studies (e.g., [Ashoori and Weisz, 2019; Hancock et al., 2011; Hoff and Bashir, 2015; Lee and See, 2004; Yin et al., 2019]) investigate the impact of factors related to user, system and task on trust while interacting with an AI. For instance, Yin et al. [2019] explores the impact of stated and actual system accuracy on users' trust, and Feng and Boyd-Graber [2019] studies how experts and novices of the given task react to Machine Learning recommendations. Recently, Glikson and Woolley [2020] reviewed, synthesised and discussed these empirical findings. While their focus was on factors that affect users' trust,

they also remarked the need to address the great variance of measures used to study trust in AI. The authors urged to refer to other disciplines in an effort to improve the current research methodology on trust in AI with human subjects. Our work does so by drawing from social and cognitive sciences, and henceforth, opens a cross-disciplinary dialogue about the practices suitable for studying trust. Our main focus is, thus, investigating *how to study* trust rather than *factors that affect* trust. Readers can refer to Glikson and Woolley [2020] for a general overview of the latter, to Chatzimparmpas et al. [2020] for a review of advances in visualization techniques related to trust, and to Chatila et al. [2021] for a discussion of the role of explainable AI in the context of trust.

### 2.2.2 *Human-AI Guidelines*

An increasing number of Human-AI guidelines provide both high and low level suggestions on how to build systems that users can trust. They focus on different aspects such as transparency [Microsoft, 2018; Royal College of Physicians, 2018; Special Interest Group on Artificial Intelligence, 2019; UNI Global Union, 2017], understandability [Society, 2017; UNI Global Union, 2017] or explainability [of Business Ethics, 2018]. Trust plays an important role in these guidelines. For instance, a recent review [Jobin et al., 2019] states that at least 30% of the ethical guidelines for AI name *trust* as one of the main ethical principles. Amershi et al. [2019] present guidelines to help in designing and evaluating AI-embedded systems that users can *trust* and work with efficiently. A framework for building trust in AI proposed by Accenture Federal Services [2019] names Human Centered Design as one of the main tools to instill trust in users.

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

### 2.2.3 *Multidisciplinary Constructs in AI*

Working on Human-AI Interaction (HAI) requires to manipulate several multidisciplinary and complex theoretical constructs such as fairness [Mulligan et al., 2019], explainability [Wang et al., 2019], interpretability [Gilpin et al., 2018] or trust. Loose use of definitions and

conflicting terminology, inherent to multidisciplinary terms, cause misunderstandings in the community. Consequently, multiple projects have been developed with the aim of disentangling these constructs in the HAI community by providing more theoretical foundations. For instance, Mulligan et al. [2019] examine fairness from the perspectives of various fields from law to computer science and create a heuristic tool for more structured interdisciplinary discussions and research collaborations around fairness. Wang et al. [2019] examine explainability as another construct which usually lacks thorough comprehension. Leveraging research on human reasoning and biases, they identify gaps in the existing Explainable AI techniques and propose and validate new ways to facilitate decision-making with AI explanations. Gilpin et al. [2018] discuss the theoretical differences between interpretability and explainability when interacting with an AI.

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

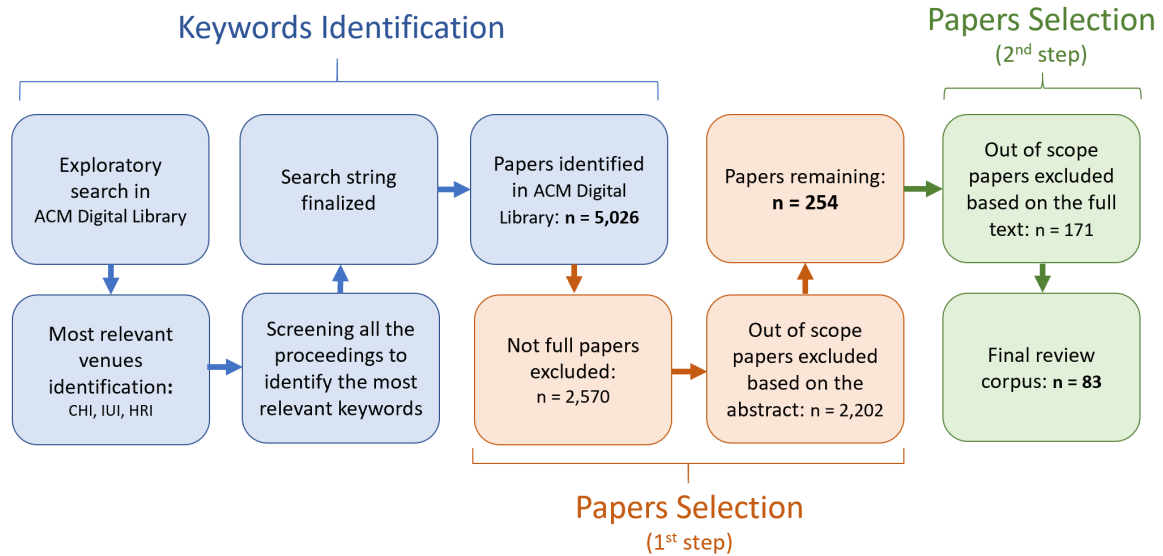
## 2.3 Systematic Review Methodology

We propose a systematic review of previous research in Human-AI trust literature in order to understand how human trust has been empirically investigated in AI-assisted decision-making. In this section, we describe the method used to perform the systematic review. It encompasses three phases: keywords identification, and 2 steps of papers selection (see Figure 2.2).

### 2.3.1 Keywords Identification

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





**Figure 2.2:** Papers search and selection process.

This exploratory search produced 386 results<sup>2</sup>. Out of them, we chose 48 relevant papers (using the same selection procedure as for the final selection of the papers, see *Paper Selection*) to identify the most reoccurring and relevant venues, which are CHI, IUI, and HRI. To find new keywords to be included in our final research string, we manually reviewed every publication in all of their proceedings from 2005 to 2020<sup>3</sup> (8108 papers in total) to find additional relevant papers.

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

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

### 2.3.2 Selection Criteria

The refined search led to 5026<sup>4</sup> papers, and we manually selected the relevant ones for our scope based on five criteria:

1. **Trust.** A paper to be selected should have results discussing Human-AI trust. If there are no results on trust reported in the paper or

<sup>2</sup> This phase was conducted in April 2020. We set no time restriction, the earliest papers found dated 2005. Such year range coincides with the recent rise of interest in Human-AI research [Grudin, 2009].

<sup>3</sup> This phase was conducted in April 2020

<sup>4</sup> All the numbers reported are as of beginning of January 2021.

if there are results on Human-Human trust instead of Human-AI trust, the paper is not included.

2. **Experiments with human participants.** We excluded all the papers that did not have an experiment (e.g., theoretical, guidelines). We also excluded the papers that ran experiments without human participants, for instance, experiments using simulated cognitive models.
3. **AI technology.** We considered the papers involving AI technologies<sup>5</sup>. As AI is a broad term, this criterion was an important selection challenge. To address it, we followed the methodology presented in [Glikson and Woolley, 2020] and included the following systems if a paper did not explicitly mention the system is AI-embedded: robots, virtual agents, and automated vehicles.
4. **Human decision making.** There is a full spectrum of ways in which humans and machines can collaborate to make a final decision with uncertain outcomes, from AI-assisted human decision making (human-centered) to AI system assisted by a human (machine-centered). A paper would be deemed relevant if it is a participant who makes the final decision(s) based on the system's output(s). For example, a hiring system could suggest to accept candidate A, but it is up to a user to take the final decision.
5. **Format.** We included only full papers, so that all the reviewed papers could contain similar level of details about a study. Therefore, posters, late-breaking works, workshops etc., were excluded. We also excluded papers in a language we could not read.

<sup>5</sup>We found several keywords including automated decision aid (e.g., [Huang and Bashir, 2017]), AI-based decision support system (e.g., [Buçinca et al., 2020]), intelligent assistant (e.g., [Ajenghughrure et al., 2019]), intelligent agent (e.g., [Tokushige et al., 2017]), classifier (e.g., [Yang et al., 2020]), etc.

### 2.3.3 Papers Selection

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

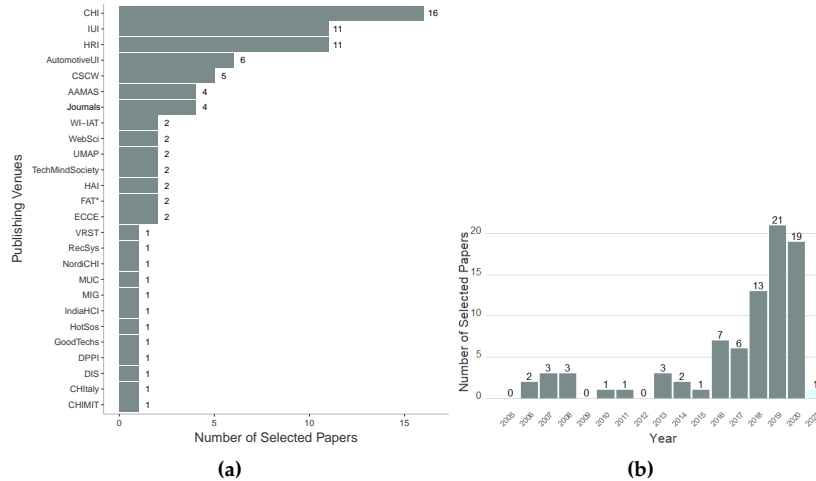
In the second selection step, the principal investigator read the full texts of these 254 papers. All the papers were read twice with a time gap of 2 weeks in a randomly reshuffled order without seeing the previous annotations during these weeks to ensure selection stability. 171 papers were deemed irrelevant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

Human-Human trust (n=4).

### 2.3.4 Corpus Overview

The final corpus consisted of **83 papers**. We did a first analysis on the publication venues and the year of publication, depicted in Figure 2.3. It shows that 79 were published in the conference proceedings and 4 in journals. Papers published at *CHI* (n=16), *IUI*, and *HRI* (n=11 each) account for 45.7%<sup>6</sup> of the selected corpus (Figure 2.3-a). The other venues were centered around socio-technical systems (e.g., CSCW, FAccT), interfaces (e.g., AutomotiveUI, DPPI), and autonomous and intelligent agents (e.g., AAMAS, HAI). 66% of the corpus have been published in the past 5 years (see Figure 2.3-b). Such a trend reflects the increasing number of publications in the venues as well as establishments of new venues (e.g., FAccT, HAI).

<sup>6</sup>Henceforth, all reported percentages are rounded up to the nearest tenth.



**Figure 2.3:** a) Number of the selected papers per publishing venue. b) Number of the selected papers per year from 2005 to 2021. Please note that the data for 2021 is incomplete since the data collection for this survey was conducted in the beginning of January 2021.

### 2.3.5 Corpus Analysis

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

- **Participants:** experience, expertise, and number;
- **Task:** the process of decision-making, task feedback and outcomes;
- **Procedure and experimental design** focusing on instructions and the order of questionnaires;

- **Data collection methods and analysis:** types of measures used, how they are implemented and analyzed.

**2) Papers Analysis.** Once we annotated each paper based on the grid above, we identified similarities in the methods used in each section and grouped them. This resulted in a categorized and complete overview of methods used for studying trust in AI-embedded systems assisting human decisions. This allowed us to identify the common practices in the community and compare them to the ones of Human-Human trust research. To do so, principal investigator, with a background in social and cognitive sciences, relied on handbooks and reviews (e.g., [Castelfranchi and Falcone, 2010; Gillespie, 2011; Lyon et al., 2015]) to identify Human-Human trust community discussions on methodology, raised issues, and proposed solutions pertinent to each element of an experimental protocol. If the common practices between Human-AI trust and Human-Human trust differed or if no common trend was spotted for the former, we explained the limitations of the current approaches stemming from our corpus. Additionally, drawing from social and cognitive science literature on Human-Human trust, we provided guidelines (**G**) aiming at overcoming these methodological limitations. We also proposed research opportunities (**RO**) for investigating further trust factors and trust methodology in the context of AI-assisted decision making.

### 2.3.6 *Review Structure and Summary*

Each section of this review is dedicated to each element of an experimental protocol we used as an annotation grid for our corpus. We start each subsection with a categorized summary of the methods used in the corpus, which we discuss in the light of social and cognitive sciences literature on Human-Human trust, highlighting strengths and limitations. Based on this, we provide guidelines and research opportunities stemming from our discussion. Table 2.1 reports a summary of the guidelines and research opportunities per section.

### 2.3.7 *A Practical Example*

We illustrate a case example to show how to apply our guidelines (and more generally this review) in practice. Consider designers who have been working on an AI-embedded system for college recruitment following principles of trustworthy AI and would like to evaluate users' trust in it. First, thanks to Sections 2.4.1 and to Sections 2.4.2 of this review, they are familiarized with what trust is (**G**<sub>1</sub>). Using section 2.4.3, they can avoid confusing terminology in their literature review search

Sections	Guidelines	Research Opportunities
Definition <i>Section 2.4</i>	(G1) Provide a clear definition of trust  (G2) Prevent any confusion between trust and related constructs	
Participants <i>Section 2.5</i>	(G3) Assess the expertise and prior experience of users (G4) Consider users' self-confidence  (G5) Favour a higher number of participants	(RO1) Investigate individual differences  (RO2) Investigate how groups of users trust an AI-embedded system (RO3) Investigate how AI-embedded systems are perceived by indirectly impacted stakeholders
Task <i>Section 2.6</i>	(G6) Consider alternative interaction flows  (G7) Ensure to involve vulnerability  (G8) Assess participants' likeliness to exhibit realistic behaviors	(RO4) Investigate the impact of the interaction flows, as factors, on trust (RO5) Investigate the impact of delayed feedback on the dynamics of trust (RO6) Investigate to what extent virtual outcomes might replace real ones
Procedure and Design <i>Section 2.7</i>	(G9) Ensure to control initial participants' expectations  (G10) Favour interactions over a long period of time	(RO7) Investigate new methodologies to assess dynamic trust in practice
Quantitative measures <i>Section 2.9</i>	(G11) Favour the use of well-established questionnaires that comprise the key elements of trust (G12) Report psychometric statistics  (G13) Use the term "trust-related behavioral measure" to avoid theoretical confusion (G14) Favour measures relative to the system's precision	(RO8) Investigate whether single-item questionnaires capture trust as well as other measures  (RO9) Explore more fundamental correlates between physiological sensing and trust
Qualitative measures <i>Section 2.10</i>	(G15) Increase empirical rigor when reporting on qualitative methods  (G16) Adopt under-used qualitative methods for studying trust (Critical Incident Technique, Repertory Grid, Hermeneutics)	

and write-up (G2). Reminded that individual differences such as age, gender, cultural background can contribute to trust variance (G3, G4), designers make sure to explore this in their analysis (RO1). Additionally, Section 2.5.2 can bring attention of the system's developers to the fact that college decisions might be made in group, rather than individually, (RO2) and to the fact that university using AI for candidates selection can affect indirect stakeholders - students (RO3). While developing an experimental protocol, designers are reminded that their participants have to have something at stake while doing the task (G7,

**Table 2.1:** Summary of the main guidelines and research opportunities organized according to the constructive elements of an experimental protocol.

G8), and Section 2.6.3 can provide multiple examples of how to do it. Designers also learn about the importance of the first impressions for participants' trust formation (G9), and can find several examples of how to introduce the system in Section 2.7.1. From Section 2.7.3, they can understand that their study should allow for an interaction long enough to record multiple stages of trust development (G10). This would also encourage them to explore which trust measures are more suitable for this (RO7, RO8, RO9). Lastly, Section 2.8.1 will help developers select an appropriate trust questionnaire (G11) and remind them what questionnaire-related statistics should be reported (G12). Section 2.8.2 will familiarize them with other trust-related measures, which do not measure trust directly (G13). If developers decide to conduct qualitative studies with their participants Sections 2.9.2 and 2.9.3 will provide them with some examples of appropriate tools to run, analyze and report one (G15, G16).

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

## 2.4 Trust Definitions

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

### 2.4.1 *Definitions in Human-AI Trust*

Only 26.5% ( $n = 22$ ) papers of our corpus provide a definition of trust resulting in 11 different definitions. 50% of these definitions are adopted directly from the social sciences literature on Human-Human trust [Boon and Holmes, 1991; Mayer et al., 1995; Young and Albaum, 2002] or adapted to Human-Machine trust, based on the grounds of social sciences [Ekman et al., 2016; Lee and Moray, 1992; Lee and See, 2004; Madsen and Gregor, 2000]<sup>7</sup>. Other papers provide definitions based on a review of existing definitions of trust in Human-Machine trust [Rajaonah et al., 2006] or propose their own based on Human-Machine and Human-Human trust definitions [Ajenaghughrure et al., 2019]. Finally, three definitions' origins were not provided [Bridgwater et al., 2020; Salomons et al., 2018; Xie et al., 2019]. We thus observe a va-

<sup>7</sup>We consider Lee and See [2004] and Lee and Moray [1992] as one definition source, because both of them have the same first author and are almost identical.

riety of definitions in the few reviewed papers with definitions. Three of them are most reoccurring<sup>8</sup> (appearing in 15 out of 22 papers; 68%):

1. “An attitude that an agent will achieve an individual’s goal in a situation characterized by uncertainty and vulnerability” [Lee and See, 2004] ( $n = 10$ , 45.5% of the 22 papers with definitions);
2. “The extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid” [Madsen and Gregor, 2000] (adapted from [McAllister, 1995]) ( $n = 3$ , 13.6%);
3. “The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party” [Mayer et al., 1995] ( $n = 3$ , 13.6%).

#### 2.4.2 Elements of Trust

By looking at the common terms in all the trust definitions in the corpus, we identify three most common types of phrases characterizing trust. They mirror three key elements of trust which arise in economics, psychology, and sociology [Rousseau et al., 1998]<sup>9</sup>: trust is linked to a situation of **vulnerability** and **positive expectations**, and is **an attitude**.

All the reviewed definitions (11) define trust as an attitude, with one paper explicitly stating that it is an “unobservable variable” [Xie et al., 2019]. 8 definitions include phrases related to positive expectations [Bridgwater et al., 2020; Ekman et al., 2016; Lee and See, 2004; Madsen and Gregor, 2000; Mayer et al., 1995; Rajaonah et al., 2006; Salomons et al., 2018; Young and Albaum, 2002], but only 3 definitions [Lee and See, 2004; Mayer et al., 1995; Xie et al., 2019] mention vulnerability. Vulnerability and positive expectations emerge from these definitions as they are the condition for trust to exist [Hosmer, 1995; Luhmann, 1979], and the idea that trust is an attitude dictates how it should be investigated and measured.

To better illustrate these elements, let’s imagine a situation where a patient has a serious illness, and their doctor proposes a treatment. The patient is in a situation of *vulnerability*, the first key element of trust, as this situation involves uncertainty of the outcomes of a decision, with potential negative or undesirable consequences [Hosmer, 1995; Luhmann, 1979]. For instance, following a treatment might just not work

<sup>8</sup> See Appendix A for the full list of trust definitions.

<sup>9</sup> It is not surprising as many definitions of our corpus rely on the one proposed by Mayer et al. [Mayer et al., 1995], a slightly modified version of the most widely accepted trust definition in social sciences [Evans and Krueger, 2009; Rousseau et al., 1998]

Definition	Origin	Vulnerability	Positive Expectations	Attitude	Papers
Lee and See [2004] and Lee and Moray [1992]	Automation, adapted from Human-Human Trust	✓	✓	✓	95; 108; 183; 298; 345; 389; 395; 396; 397
Mayer et al. [1995]	Human-Human Trust	✓	✓	✓	109; 278
Ekman et al. [2016], a combination of Lee and See [2004] and Mayer et al. [1995]	Automated Vehicles, adapted from Automation and Human-Human Trust	✗	✓	✓	380
Madsen and Gregor [2000] (adapted from McAllister [1995])	HCI, adapted from Human-Human Trust	✗	✓	✓	84; 181; 387
Young and Albaum [2002]	Human-Human Trust	✗	✓	✓	400
Boon and Holmes [1991]	Human-Human Trust	✗	✗	✓	130
Rajaonah et al. [2006]	Review of Human-Automation and Human-Computer Trust	✗	✓	✓	290
Their own definition	Review of Human-Human and Human-Computer Trust	✗	✗	✓	7
Flawed source stated	–	✗	✓	✓	45
No source stated	–	✗	✓	✓	312
No source stated	–	✓	✗	✓	384
<p>No definition: 2; 10; 17; 46; 49; 53; 54; 63; 65; 71; 76; 88; 99; 107; 117; 123; 127; 137; 139; 143; 145; 149; 164; 179; 186; 195; 203; 205; 214; 230; 249; 253; 261; 273; 274; 279; 287; 288; 294; 302; 321; 323; 329; 332; 343; 346; 350; 353; 354; 358; 359; 361; 368; 370; 371; 382; 383; 388; 392; 394; 402</p>					

**Table 2.2:** List of trust definitions in the reviewed Human-AI papers with highlighted key elements of trust.



or might provoke severe side effects. Uncertainty might be due to the unpredictable nature of the world as well as the lack of human knowledge and capabilities [Castelfranchi and Falcone, 2010]. However, it is necessary to distinguish two natures of uncertainty (sometimes referred as risk vs. ambiguity [Knight, 1921]): the possibility of outcomes can sometimes be estimated (e.g. the treatment has 30% of success with full recovery) or not (e.g. the percentage of success or the side effects of the treatment are not known). In this chapter, the notion of vulnerability relates to both types of uncertainty. Without vulnerability, there is no need for trust to emerge [Castelfranchi and Falcone, 2010; Gambetta, 2000; Lascaux, 2008; Offe, 1999].

Similarly, trust will not emerge if the patient does not have *positive expectations*, the second key element, about the treatment the doctor assigned them. Even if the patient decides to follow it anyway, we cannot claim that the patient trusts the doctor [Hosmer, 1995; Luhmann, 1979]. Trust has grounds to form only when one thinks that negative outcomes associated with trusting do not exist or are very unlikely [Lewis and Weigert, 1985].

The third key element is that trust does not systematically translate into a behavior. For example, the patient's level of trust might not be sufficient enough to follow the doctor's suggestion or the patient trusts the doctor's suggestion enough, but, lacks financial resources to follow it. It is also possible to have actions without trust if the patient has no other option, but to follow the doctor's suggestion. A socio-cognitive approach to defining trust suggests that trust is rather an attitude [Castelfranchi and Falcone, 2010], i.e. a certain way of feeling about the object [Bohner and Dickel, 2011]. Trust then cannot always be fully observable to the third parties (unless clearly and objectively communicated in a verbal or written form), which has an important impact on the choice of the methods to study trust (see sections 2.9 and 2.10).

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

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

In the rest of our paper, we rely on the Lee and See [2004] definition when referring to trust: "*An attitude that an agent will achieve an*

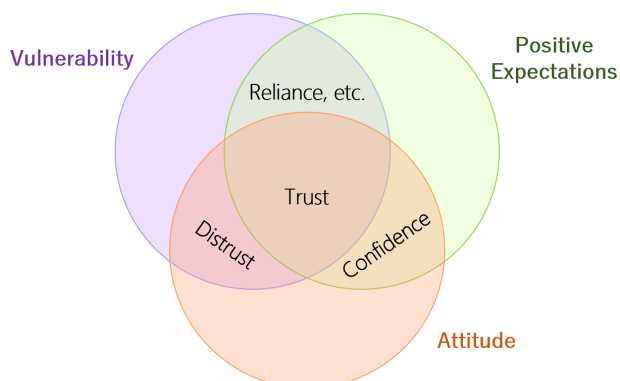
*individual's goal in a situation characterized by uncertainty and vulnerability."* We favour this definition as it comprises the three key elements of trust and is the most used definition in the corpus as well as in the Human-Automation literature. We note, however, that mentioning both "vulnerability" and "uncertainty" in this definition appears to be redundant since vulnerability already comprises uncertainty as explained above. On the other hand, it is the only definition in the corpus that explicitly highlights the notion of uncertainty as impossibility to estimate the likelihood of outcomes, which characterizes the general theme of the scenarios employed in our corpus - decision making with uncertain outcomes.

### 2.4.3 *Constructs Related to Trust*

The three elements presented above are constitutive to trust. Concealing one element leads to consider a different concept, yet related to trust. For instance, without an element of vulnerability in a situation, it would be more appropriate to consider **confidence** instead of trust [Evans and Krueger, 2009; Luhmann, 2000], which is the case in 3 definitions from the corpus [Bridgwater et al., 2020; Madsen and Gregor, 2000; Salomons et al., 2018]. Continuing with the previous example, let's say the illness is not severe and the treatment is unlikely to have serious side effects. When the patient decides to follow the treatment without considering any alternatives and thinks that they will only be better off with it, this suggests that the patient is *confident* in the doctor's suggestion. When there is more of vulnerability for the patient's health, the patient might start looking into alternatives or into not accepting the treatment at all. If in the end they decide to follow the doctor's suggestion despite potential serious side effects of the treatment, this suggests that the patient *trusts* this suggestion [Luhmann, 2000].

Without positive expectations, it is more appropriate to discuss **distrust**. This construct is often confounded with low levels of trust [Mcknight and Chervany, 2001]. While there are some researchers who deem trust and distrust as the opposite ends of one construct [Rotter, 1980], recently the community views them as two separate ones [Lewicki et al., 1998; Sitkin and Roth, 1993]. This means that they can both reach high and low levels and exist simultaneously. Just like for trust, too much of distrust can be harmful as it can lead to inability to identify correct recommendations. Only calibrated levels of trust and distrust are beneficial for decision-making as under this condition users are less likely to blindly follow incorrect recommendations and to override correct ones [Mcknight and Chervany, 2001].

Sometimes trust is confounded with behaviors such as **reliance** and **compliance**. The former is defined as the decision to follow someone's recommendation, and the latter as the decision to ask for a recommendation in the first place. As we have discussed before, trust does not always translate to a behavior, but there is definitely a relation between them [Deutsch, 1958, 1960; Hoff and Bashir, 2015; Holmes and Rempel, 1985; Lee and See, 2004; Meyer and Lee, 2013]. Figure 2.4 summarizes how these constructs are related to trust.



**Figure 2.4:** A simplified representation of some constructs related to trust and how they are connected with the key elements of trust.

Finally, trust is also related to **perceived trustworthiness**. If the patient thinks that the doctor is trustworthy (e.g., has many diplomas, was recommended by someone), this does not mean the patient will trust them. Generally, perceived trustworthiness is not recommended to be used as a proxy for how much another person trusts the counterpart. In addition, having beliefs about the degree of someone's trustworthiness does not involve any vulnerability, a key element of trust [Gillespie, 2003].

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

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

In the next sections, we will summarize and categorize different methodologies used while designing an experimental protocol for studying Human-AI trust in the context of decision making.

## 2.5 Participants

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

### 2.5.1 Experience and Expertise

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

Our past experiences drive our expectations [Pavlov, 2010; Sutton and Barto, 1998], and can affect the way we update our beliefs. For example, if the past experience with a system was negative, a participant is more likely to overreact to an error during an experiment, reconfirming their initial expectations [Fareri et al., 2012]. Therefore, inquiring about participants' prior experience can help in analysing the study's data. It is thus recommended to:

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

*Subjective expertise.* A small subset of papers ( $n = 10$ , 12%) also ask participants about how well they think they understand how to use the system or, in other words, measure their subjective expertise (also called self-confidence or self-efficacy [Perry, 2011]). Subjective expertise is how well participants think they can achieve their goal (e.g., solving a problem). Research suggests that people are generally overconfident in their abilities, which leads to biased judgement [Lichtenstein et al., 1977; Yates, 1990] and in turn might affect trust-related perceptions and decisions [Lee and Moray, 1994].

It is believed that its magnitude depends on gender [Barber and Odean, 2001], age or culture [Hyde, 2005]. In our corpus, only 3 studies consider these individual differences, but they do not link them to self-confidence. The first research opportunity would be:

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

It is, therefore, important to:

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

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

### 2.5.2 *Groups and Stakeholders*

*Number of participants.* The average number of participants per study is 134 ( $SD = 259.9$ , median = 48), which is high in comparison with standard HCI experiments. Such a high number can be explained by the fact that some crowd-sourcing studies ( $n = 29$ , 33.7%) recruited very large number of participants (i.e., 1994, 757, and 1042 in Yin et al. [2019]). Consequently, the average number of participants per crowd-sourcing study is 340 and is much higher than the one of the rest - 51. This could be an indicator that in the corpus, trust has been mostly studied by recruiting a large number of participants in order to compute quantitative correlates (see Section 2.9). It could also be due to the fact that studies related to social sciences and psychology are recommended to recruit a larger number of participants [Brysbaert, 2019; Roscoe, 1969]. As trust is a psychological construct, one should:

G5 **Favour higher numbers of participants** than in standard HCI experiments [Caine, 2016].

*Individual vs. group of participants* The predominant trend in Human-AI trust with decision-making is to investigate trust of an individual ( $n = 81$ , 97.76%). However, a line of literature in social sciences suggests the importance of considering trust of a group (e.g., [Dietz and Hartog, 2006; Fulmer and Gelfand, 2012; Huff and Kelley, 2003]). Indeed, group decisions with an AI-embedded system are part of real-life cases, especially in the medical field (e.g., [Yang et al., 2016]). Moreover, group decision-making and trust processes have been shown to be different from the individual ones [Kim et al., 2013]. For example, repairing trust has been found to be more difficult for groups than for individuals [Kim et al., 2013]. In our corpus, we found only two pa-

pers that investigate trust of a group with decision-making, and they look into groups of 2 users [Shamekhi et al., 2018; Wang et al., 2010]. Thus, there needs to be more research on:

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

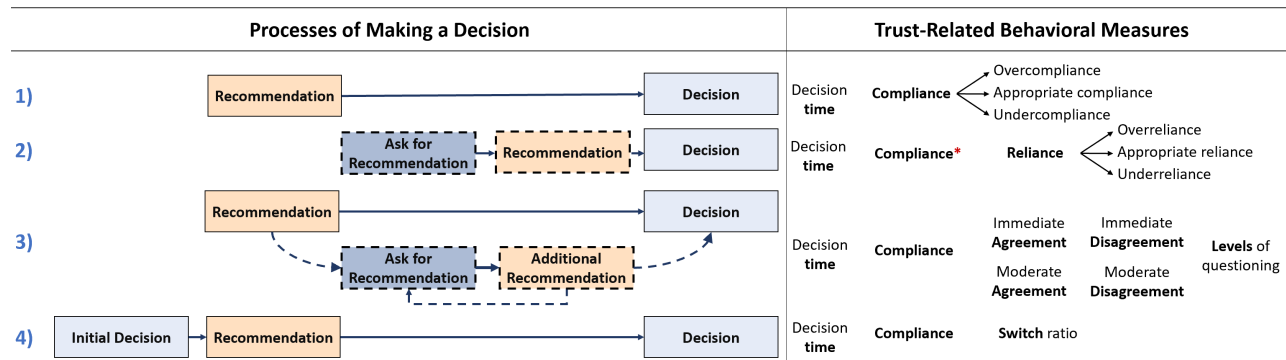
*Direct vs. indirect interaction.* In most experiments ( $n = 75$ , 93.75%), the participant has a role of the user *directly* interacting with the system. However, there are other stakeholders who do not interact with the system directly, yet can be impacted by the decisions made with AI-embedded systems, and it could be insightful to investigate their trust, too. For example, would patients still listen to the doctor if they had known beforehand the doctor is assisted by an AI for diagnosis assessment [Cohen, 2020]? Would citizens be upset to the same extent about a new bus schedule if it had been created manually instead of with the help of an AI [Ito, 2018]?

Discussions around this type of trust, referred to as indirect trust, is predominately found in the research community of reputation systems, mostly with a purpose of software optimization [Gutscher, 2007]. In the reviewed corpus, we found that automated vehicles research starts to focus on studying trust of indirect stakeholders such as pedestrians, because they are also affected by the decisions of direct users [Ackermans et al., 2020; Holländer et al., 2019; Reig et al., 2018]. Besides automated vehicles, the attitudes towards AI-embedded systems of stakeholders, who are affected by the decisions of direct users, is also explored with algorithms [Brown et al., 2019; Woodruff et al., 2018]. This promising line of research should be further expanded by:

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

## 2.6 Task

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



### 2.6.1 Interaction flow in the Decision Making Process

We found a large variability in the process of making decision as illustrated in Figure 2.5. The most common and simplest pattern consists of presenting a recommendation and letting the participant follow, or not, this recommendation (case 1;  $n = 46$ , 60.8% of all the quantitative studies). This setup reflects, for instance, an autonomous system detecting a risk and suggesting an alternative way of working: the operator is then free to accept or reject this recommendation. An alternative process does not automatically provide a recommendation. It gives participants the choice of questioning it if needed (case 2;  $n = 6$ , 9.2%). This process reflects, for instance, when a person is free to choose their source of recommendations if any at all. It also has the advantage to study participants' reliance on the system. These two processes can be combined as illustrated in Figure 2.5. After having provided an initial recommendation, participants can ask for one or several recommendations or additional information (case 3,  $n = 3$ , 4.6%). Asking for several recommendations is common in Human-Human trust [Johnston et al., 2015; Koenig and Jaswal, 2011; Landrum et al., 2013]. For instance, asking the opinion of several doctors before deciding for a specific treatment. Finally, case 4 is interesting ( $n = 14$ , 21.5%) because it is rare to get a recommendation *after* having made a decision in a real-world scenario. For example, in the 7 papers in our corpus that studied functioning prototypes, the recommendation was shown immediately, without a request from participants. The practicality of case 4 in experimental settings makes it possible, however, to compare the decision made before and after receiving a recommendation, and thus to capture the degree of participants' compliance with the system.

The choice of an interaction flow affects what logs researchers can record related to participants' decisions, how and if these decisions evolve. As a consequence, this choice also impacts what other mea-

**Figure 2.5:** A schematic representation of different types of decision making processes and behavioral measures associated with them (discussed further in Section 2.9.2). Dotted lines indicates that the step is optional. \* indicates that the measure is possible only if optional steps are taken.

measures researchers can calculate (right part of Figure 4, learn more about these measures in 8.2). Yet, usually studies in the corpus do not motivate their choice of the interaction flow. The only exceptions are the 14 papers that adopted case 4 to explore whether and when participants changed their decisions. In other cases, the link between an interaction flow and possible measures was not paid attention to. While developing a study, researchers should:

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

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

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

### 2.6.2 *Feedback*

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



indirectly about their performance.

The feedback can be **immediate** ( $n = 39$ , 60%) allowing participants to dynamically update their level of trust. For instance, you will immediately update your trust when you realize you followed a slower route with a lot of traffic due to AI's recommendation. However, in many scenarios, feedback can only be received after some **delay**, e.g. the consequence of the choice of a medical treatment. In our corpus, only 7.6% of the studies provide feedback after a block of decisions ( $n = 4$ ), after one day ( $n = 1$ ) or even within weeks ( $n = 1$ ). This indicates there is a lack of studies with more real-world scenarios. For example, while assigning a treatment or giving out a loan, the decision maker might learn whether they were wrong over a longer period of time - weeks, months or years. As it remains unclear,

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

### 2.6.3 *Task Outcomes*

Vulnerability is one of the pillars of trust (section 2.2). While (im)possibility to predict the likelihood of outcomes is introduced to the experiments through the nature of scenarios AI-embedded systems are used for in the studies (e.g., medical decisions, rescue operations), introducing undesired and regretful outcomes might require more thought. To immerse participants in the state of vulnerability, they must feel that their decisions matter, that is having something at stake. Therefore, task outcomes play an important part in triggering a sense of vulnerability in participants. Researchers have to:

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

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

**G8 Assess participants' likeliness to exhibit realistic behaviors**, that

is how immersive they are and to which extent participants are engaged.

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

**1) Real Incentives.** Only few studies introduce real outcomes ( $n = 20$ , 30.8% of the 65 quantitative studies). One of the strategies includes **temporal** incentives ( $n = 3$ , 4.6%). For every wrong decision, participants have to wait a couple of seconds which can quickly be annoying. Another one is **monetary** incentive ( $n = 12$ , 18.5%) where participants can receive only performance *bonus* ( $n = 8$ , 12.3%) or *bonus and malus* ( $n = 4$ , 6.2%). This strategy is widespread in economics as it has been proven that participants tend to give more optimal answers and avoid random guessing (with an exception for when the task is too complicated) [Camerer and Hogarth, 1999; Hertwig and Ortmann, 2003]. If the bonuses are too small, participants might disregard them and feel unmotivated to perform well [Gneezy and Rustichini, 2000]. If the bonuses are too high, this might put unnecessary pressure on the participants, hindering their motivation [Baumeister, 1984]. Another strategy is **cognitive effort** incentives ( $n = 2$ , 3.1% e.g. solving a puzzle for a long time and losing in the end [Kulms and Kopp, 2019]).

**2) Virtual Incentives.** The majority of studies ( $n = 48$ , 73.85%) use virtual incentives, presumably because they are easier to implement. Among them, we differentiate **virtual penalties** ( $n = 11$ , 16.9% e.g. game points [Natarajan and Gombolay, 2020; Yu et al., 2016, 2017, 2019]), **negative virtual consequences for participants** ( $n = 29$ , 44.6%, e.g. car accident [Maurer et al., 2018; Rajaonah et al., 2006]) or **negative virtual consequences for other stakeholders** ( $n = 9$ , 13.4%, e.g. injury or fatality [Fan et al., 2008; van Maanen et al., 2011]).

However, it is unclear whether virtual outcomes (e.g. virtual car accident) might replace real ones and produce a sense of vulnerability. Recent findings in experimental economics suggest that if the virtual environment is immersive enough (through a presence questionnaire: e.g., [Whelan, 2008]), participants might suppress the feeling of participating in an experiment and consequently demonstrate more realistic behaviors in decision-making tasks with risk [Gürerk et al., 2014]. It remains that participants know that the researchers are not allowed to hurt them. For instance, one study simulates an emergency evacuation but participants rated it 1.5 out of 7 on credibility [Robinette et al.,

2016]. More exploration needs to be done to:

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

## 2.7 Procedure and Design

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

### 2.7.1 *Introducing the System Performance*

Positive expectations are a necessary component of trust (Section 2.4.2). If before or at the initial stages of interaction participants do not have positive expectations about the system, then trust will not start forming and developing. It is thus important to:

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

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

The first one is **instructions**. Two studies [Yang et al., 2020; Yin et al., 2019] directly signal to participants the systems' accuracy percentage (stated accuracy). Four studies [Andrist et al., 2013; Wang et al., 2010; Xiao et al., 2007; Yuksel et al., 2017] follow a less direct approach and introduce their systems claiming they have appropriate expertise for the task, without going into many details. For instance, the systems are simply described to be "reliable" [Yuksel et al., 2017] to do the task. Three other studies [Andrist et al., 2013; Wang et al., 2010; Xiao et al., 2007] mention the system had relevant past experience.

The second is the **initial experience** when interacting with the system. Several studies make the system error-free for the first recommendation [Yang et al., 2020] or in the first group of recommendations [Yu et al., 2016, 2017, 2019] to evoke positive expectations. Indeed, the effect of a mistake occurring during the early stages of an interaction on trust is unlikely to diminish over time [Lee and See, 2004; Marshall,

2003; Robert B. Lount et al., 2008]. A mistake during the last stages of interaction can also negatively distort the trust reports due to a bias in memory [Kahneman, 2000].

The third one is the **behavior** of the system itself, by guarantying a minimum level of accuracy. Indeed, previous studies indicate that 60% - 80% accuracy is considered to be an appropriate window for investigating users' trust in AI-embedded decision-support systems [Baudel et al., 2021; Yin et al., 2019; Yu et al., 2016; Zhang et al., 2020]. Below this threshold, the AI recommendations fail to be helpful for decision makers [Onnasch, 2015], and the study is thus more likely to investigate distrust rather than trust. We would expect this threshold, however, to be context-dependent. For example, 80% in the medical field would be too low.

### 2.7.2 Experimental Design

**Between-subject design** is especially appropriate when the investigated factors can introduce learning effects ( $n = 43$ ) (e.g. the way system communicates) or are related to the profile of the participants ( $n=5$ ) (age, nationality, etc.). However, it requires a large number of participants. In contrast, **within-subject design** requires less participants if it is compatible with the research question. For example, studies can adopt a within-subject design for investigating *accuracy of the system* if they are interested in how different levels of accuracy affect users' trust. If it is not the case and running a between-subject study is not possible, one can to increase the elapsed time between the different conditions<sup>10</sup> (e.g. 2-5 days apart) to reduce learning effects [Yang et al., 2020].

### 2.7.3 Assessing Pre-, Post-, or Dynamic Trust

It is common practice to use questionnaires during the experiment to capture different aspects of trust (see Section 2.9.1). Introducing a questionnaire **before** the interaction with a system ( $n = 3$ , 5.5% of the 55 studies that included questionnaires) captures participants' initial trust, based on their own beliefs and previous experiences if any. **After** the interaction ( $n = 22$ , 40%), it captures participants final trust in a system, affected by the recent interaction. However, these approaches do not capture changes in the participants' trust in the system. Trust is dynamic, it can be increased, decreased, repaired, and maintained [Lewicki and Brinsfield, 2011]. To capture some of these changes, an alternative is to introduce the same questionnaire **before and after** the interaction ( $n = 5$ , 9.1%). In within-subject studies, it is common to

<sup>10</sup> Condition is a level of the independent variable that is manipulated by the researcher in order to assess the effect on a dependent variable (from American Psychological Association Dictionary). For example, system's accuracy as variable can have three conditions - low, average, high.

introduce the trust questionnaire **after each condition** ( $n = 13, 24.1\%$ ) to avoid interference between conditions.

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

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

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

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

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

Ideally, they should not be too long and intrusive in the course of an experiment, and should try to capture trust measures as continuously as possible (more about different types of measures see Sections 2.9 and 2.10).

## 2.8 Summary of Findings for Experimental Protocol

In Sections 2.5, 2.6 and 2.7, we summarize and categorize the methodologies employed in the experimental protocols for studying Human-AI trust in the decision making context. We highlight to which extent key elements of human trust (vulnerability, positive expectations, and attitude; see Chapter 2) are integrated in the design of experimental

protocol and how this can be improved (our guidelines). We also identify some elements of experimental protocol that can be factors that impact Human-AI trust and need further investigations (our research opportunities).

When it comes to participants (Section 2.5), we find that individual differences related to participants' expectations around decision making with AI like prior experiences with AI and subjective and objective expertise of the task at hand should be better controlled (G<sub>3,4</sub>) and further studied (RO<sub>1</sub>). For example, only a small set of papers assesses to which extent participants are good at the decision-making task (objective) rather than relying solely on a self-reported evaluation of expertise (subjective). Additionally, most of the studies investigate trust of individual decision makers who interact directly with the system. It would, therefore, be insightful to investigate trust processes when several people interact with AI-embedded system to make one decision (RO<sub>2</sub>) as well as trust of those who do not interact with the system, but are affected by the Human-AI decision (also referred to as "decision subjects" [Höddinghaus et al., 2021]; RO<sub>3</sub>). An example of such a scenario could be medical decision making, where a team of medical workers interact directly with the system to make a decision regarding a patient.

When deciding on how to design a task (Section 2.6), there are three points of consideration: interaction flow, feedback, and decision-making outcomes. The interaction flow, that is the order in which participants make their decision, see an AI recommendation, and its optionality have an impact on what trust-related behavioral measures one can derive (G<sub>6</sub>). We identified 4 possible interaction flows, represented in Figure 2.5 and the possible measures. We note, however, that recent evidence shows that making AI recommendations optional (scenario 3, Figure 2.5) or have participants make an initial decision before seeing an AI recommendations (scenario 4, Figure 2.5) reduces participants' overreliance on AI recommendations and their acceptance of the system's design [Buçinca et al., 2021]. It is not known yet, however, to which extent these results apply to high-risk scenarios and how different interaction flows affect Human-AI trust, and this remains to be investigated (RO<sub>4</sub>). Secondly, most studies in the reviewed papers that provide feedback about participants' performance or quality of AI recommendations do it immediately. However, in many real-life scenarios, feedback can only be received after some delay, for example, learning about consequences of a medical treatment. While the evidence shows that the delay of AI recommendations reduces reliance on AI recommendations and the design acceptance [Buçinca et al., 2021],

it is not known yet the impact of the feedback delay on trust and related constructs (RO5). Lastly, task outcomes are the way to introduce the element of vulnerability during the study. To make participants feel more involved in the decision making, task outcomes can take form of real incentives (e.g., monetary bonuses and maluses, forced timeout) and virtual ones (e.g., game points, harm to virtual characters) (G7). As it is unclear whether virtual outcomes (e.g. virtual car accident) might replace real ones and produce a sense of vulnerability (RO6), one has to assess to which extent participants feel involved in the experiment (G8).

The considerations around procedure and design (Section 2.7) evolve around the ways to introduce the system to participants and when to include trust measures. The way the system is introduced to the participants contributes to their expectations formation. For trust to be triggered, participants should have positive expectations. This could be achieved through announce stated accuracy and other relevant information in the introduction, controlling the actual accuracy, and making sure that the erroneous AI recommendations are not in the first and last several trials (G9). Finally, the choice when to include trust measures - before, after, before and after, and/or during the interaction - affects to which extent one can capture dynamic changes of trust. Trust levels evolve throughout the interaction, and to capture the full picture of its dynamics, one has to ensure that the interaction is long enough (G10), but there is still a need for less intrusive ways to capture data about trust on a more continuous basis (RO7).

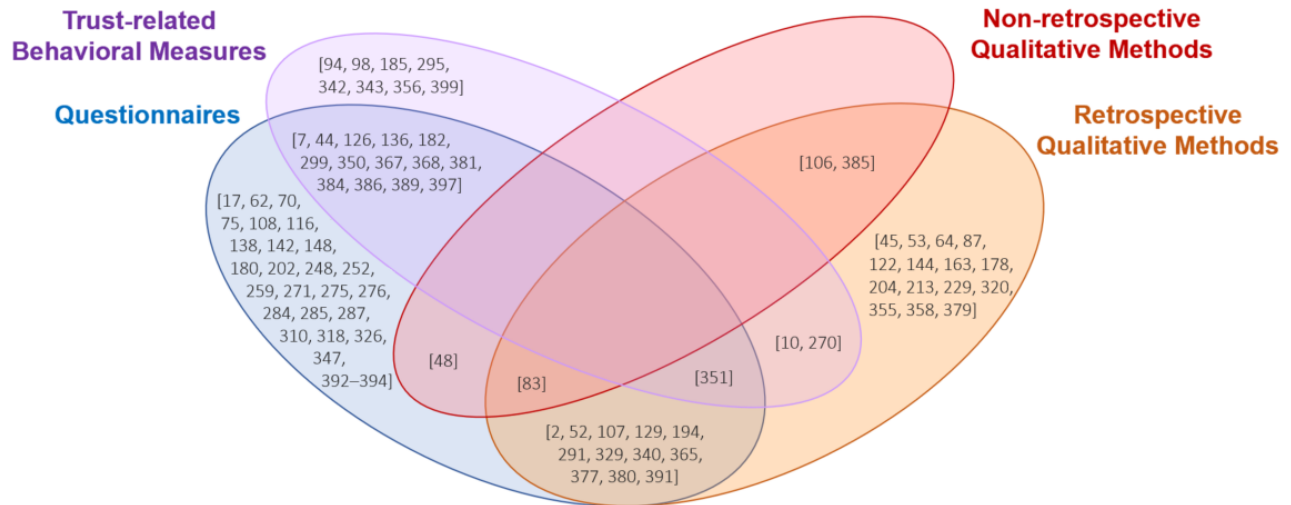
In the next sections, we will summarize and categorize different measures, first quantitative and then qualitative, used for studying Human-AI trust in the context of decision making (Figure 2.6).

## 2.9 Quantitative Measures

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

### 2.9.1 Questionnaires

Questionnaires usually consist of a series of questions, a minimum of three or four [Robinson, 2018]. They are a common method to measure Human-Human trust because [Gillespie, 2011]: (1) they allow to



**Figure 2.6:** Distribution of papers according to which types of quantitative and/or qualitative methods used. There are 66 quantitative trust studies and 34 qualitative ones. Questionnaires ( $n = 56$ ) and Trust-related Behavioral Measures ( $n = 25$ ) belong to quantitative methods. Non-retrospective ( $n = 4$ ) and Retrospective ( $n = 34$ ) belong to qualitative methods.

capture a person’s attitude, what they feel and think [Lewicki and Brinsfield, 2011]; (2) participants might feel more at ease reporting their psychological state as they are not facing another person, but just a screen or a sheet of paper; and (3) they are relatively easy and quick to implement, allowing to collect a bigger quantity of data in a shorter period of time (in comparison with interviews and observations).

**Questionnaires Origins.** Among 32 papers with multi-question questionnaires in our corpus, we identified 21 different questionnaires used to measure participants’ trust in an AI-embedded system (refer to Appendix C for their full text of those we were able to retrieve). Four of the questionnaires originate from Human-Automation literature [Chien et al., 2018; Jian et al., 2000; Merritt, 2011; Muir, 1989] and are cited by almost half ( $n = 15$ ) of the papers with questionnaires. 2 questionnaires were taken from Human-Human trust literature [Mayer, 1999; Ohanian, 1990; Wheelless and Grotz, 1977] ( $n = 5$ , 15.6% of the 32 papers that used questionnaires). The remaining questionnaires originate from Human-Robot trust [Ross, 2008; Schaefer, 2013] ( $n = 4$ , 12.5%), e-Commerce [McKnight et al., 2002], Human-Agent Interaction [Chien et al., 2018], Automated Vehicles [Deb et al., 2017], and HCI [Madsen and Gregor, 2000] ( $n = 1$ , 3.1% each). Finally, 6 questionnaires were not explicitly attributed a source. All the papers, but one [Chien et al., 2018], use an already existing questionnaire. [Chien et al., 2018] introduced their own questionnaire, which accounts for cultural influences on trust. 3 papers [Ghai et al., 2021; Müller et al., 2020; Yu et al., 2020] combine multiple existing and validated trust questionnaires to create a new one for their studies.



Such a variety of questionnaires in the corpus mirrors a general trend of a quite broad choice of questionnaires among the existing trust questionnaires in social sciences [McEvily and Tortoriello, 2011]. This could be explained by the efforts of both of the communities to make the questionnaires more context-specific. However, for now, it has led to the abundance of choice, inhibition understanding of how to appropriately choose and use a questionnaire.

**Link Between Trust Definitions and Trust Questionnaires.** The review of Human-Human trust questionnaires [McEvily and Tortoriello, 2011] suggests that the Mayer [1999] questionnaire equally comprises all the theoretical concepts related to trust without focusing on the related constructs. It also named two other trust questionnaires that reflect well the trust definition: Boundary Role Persons (BRP)<sup>11</sup> [Currall and Judge, 1995], and Behavioral Trust Inventory [Gillespie, 2003].

However, theoretical discrepancies often appear both in many Human-Human trust questionnaires [Dirks and Ferrin, 2002; Graham and N., 2006; McEvily and Tortoriello, 2011] and in the questionnaires from our corpus. Indeed, several questionnaires in our corpus mainly focus on *positive expectations* ( $n = 8$ , 38.1% of the 21 questionnaires) or *vulnerability* ( $n = 8$ , 38.1%). In contrast, the questionnaire by Mayer [1999] ( $n = 3$ ) and the one by McKnight et al. [2002] equally focus on vulnerability and positive expectations [McKnight et al., 1998]<sup>12</sup>. Finally, one of the most reoccurring trust questionnaires [Jian et al., 2000] ( $n = 11$ , 34.4% of the 32 papers with questionnaires) in the corpus also includes vulnerability and positive expectations but still faces some theoretical discrepancy. 5 out of 11 of its questions are reverse coded and thus related to distrust. However, the research community preferably regard trust and distrust as two separate constructs (see section 2.4.3).

**Questionnaires Modifications.** More than half of the papers of our corpus using multi-question questionnaires ( $n = 19$ , 59.4%) introduce modifications to the original, validated questionnaires. It includes changing some words in the questions to better fit the case study ( $n = 4$ ), e.g. replacing the word “system” with “decision aid” [Oduor and Campbell, 2007] or reducing the number of questions ( $n = 8$ ).

However, most of these papers do not report what are the changes ( $n = 8$ , 42.1% of the papers with modifications). Additionally, none of the modifications have been validated in all the 16 papers. This can undermine the questionnaires reliability and the accuracy of the repli-

<sup>11</sup> Boundary Role Persons is an umbrella term for employees who represent their company/group outside the organization and collect and rapport information from external sources to their employers.

<sup>12</sup> See Appendices B and C for the questionnaires' items.

cations as even small changes such as replacing a word or inverting two questions can invalidate the investigation of complex constructs such as trust [Sarlis and Gallhofer, 2007].

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

**G11 Favour well-established questionnaires that equally comprise all the theoretical concepts related to trust** without focusing on the related constructs [Gillespie, 2003].

The examples of such questionnaires are [Aubert and Kelsey, 2003; Currall and Judge, 1995; Gillespie, 2003; Mayer, 1999; McKnight et al., 2002]. If an existing questionnaire needs modifications to better fit in the context, researchers should also ensure that these changes are consistent with the trust definition as per G11.

**Single-item Questionnaires.** Some questionnaires can have a single question ( $n = 24$ ) which asks participants either to rate the trust in the system, e.g. “How much did you trust our machine learning algorithm’s predictions on the first twenty speed dating participants?” [Yin et al., 2019] or to rank the systems, e.g. “Rank the agents in order of trust” [Bridgwater et al., 2020]. Single-item questionnaires have the advantage to be quick for participants to answer [Robinson, 2018], but they are generally less appropriate to study complex constructs like trust [Robert, 2002]. Specifically, when it comes to measuring appropriate trust, there is an issue in determining which score on the Likert scale (e.g., 3 or 4) was exactly appropriate trust. The bigger question is whether “rating” trust is insightful and meaningful enough and if participants can objectively assign a score to their trust levels. There is a need to:

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

**Psychometric Statistics for Replication.** It is a good practice once the participants’ responses are collected, to use *psychometric statistics* to verify whether a reused or modified questionnaire still measures trust accurately in a new, independent study. However, we found that only 14 papers out of 34 (41.2%) reported psychometric statistics in data analysis. This echoes McEvily and Tortoriello’s review

[235] on Human-Human trust, showing that most studies in the field did not report enough information on psychometric statistics in the questionnaires' analysis. Moreover, when it is done, the analysis often uses Cronbach's alpha [Peter, 1979] requiring several conditions to hold true, which are rather strict (e.g. unidimensionality of the construct). Additionally, reported alone, Cronbach's alpha gives little insight about the questionnaire [Sijtsma, 2008]. The  $\omega$  coefficient [Deng and Chan, 2017; Dunn et al., 2014; Trizano-Hermosilla and Alvarado, 2016] might be more appropriate as it has more relaxed requirements, and other statistics such as confirmatory factor analysis (CFA) [Jöreskog, 1967; Moore, 2012; Plucker, 2003] need to be reported, too. See [Peters, 2014; Saris and Gallhofer, 2007] for more information about different types of psychometric statistics and how to implement them. We thus strongly encourage the community to:

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

### 2.9.2 Trust-related Behavioral Measures

In our corpus, 25 papers (37.9% of the 66 quantitative papers) record logs about participants' activity and use them to derive what is often referred to as "behavioral measures" of trust. As trust cannot be always inferred from a behavior (section 2.4.2), it might be misleading to refer to these measures as "behavioral trust measures". To avoid confusion, we preferably:

- G13 Use the term *trust-related behavioral measures* instead of *behavioral trust measures* [Mcknight and Chervany, 2001].

We identify three types of trust-related behavioral measures.

**1) Trust-related Behavioral Measures Based on Following Recommendations.** Figure 2.5 illustrates different processes to make a decision and the associated quantitative measures. Two measures are independent of the process, **Decision Time**, i.e. how fast a recommendation is accepted ( $n = 2$ , 8% of the 25 papers with trust-related behavioral measures, [Feng and Boyd-Graber, 2019; Yuksel et al., 2017]) and **Compliance**. **Compliance** is the number of times participants follow the systems' recommendations ( $n = 18$ , 72%), both correct and incorrect ones. It is then possible to calculate:

- *appropriate compliance*: correct recommendations accepted ( $n = 2$ )

and incorrect recommendations non-accepted ( $n = 2$ );

- *overcompliance*: incorrect recommendations accepted ( $n = 3$ );
- *undercompliance*: correct recommendations rejected ( $n = 1$ ).

When the recommendation is not initially provided (case 2, Figure 2.5), it is also possible to estimate **Reliance** - the number of times participants asked for a recommendation ( $n = 4$ , 16%) and to derive [Sutherland et al., 2015]:

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

Additionally, when participants are free to ask several (typically up to two) recommendations [Bridgwater et al., 2020; Gruber et al., 2018; Yang et al., 2017], the first one being automatically given (case 3, Figure 2.5). We can thus derive the following measures, where how quickly a recommendation is accepted is considered to be indicative of high levels of trust:

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

Finally, when participants indicate an initial decision (before receiving the recommendation (case 4, Figure 2.5), we can estimate the **Switch ratio**, the number of times a participant who initially disagreed with

the system decided to follow its recommendation in the end ( $n = 3$ , 12%). It is assumed the higher the switch ratio is, the higher the levels of trust are.

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

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

**2) Other Trust-related Behavioral Measures.** [Torre et al., 2018] links the amount of money shared with the system as a trust-related behavioral measure, inspired by game theory situations such as *Prisoner's Dilemma* [Deutsch, 1960; Loomis, 1959]. However, such games are criticized for confounding trust with altruism [Cox, 2004] and betrayal aversion [Bohnet et al., 2008; Fehr, 2009], and for the lack of stability [Burnham et al., 2000; Johnson and Mislin, 2011; Sun et al., 2019], and hence, should not be preferred for measuring trust-related behaviors. Finally, measures were related to scenario-specific events such as how long the brakes were hold for and with what intensity [Frison et al., 2019] in an automated vehicle.

**3) Physiological Measures.** In quest of collecting objective trust data, which is not under participants' control and, hence, is less subjective than responses to trust questionnaires, researchers start turning to physiological measures [Novak, 2014]. There is some evidence that higher levels of stress are associated with lower levels of trust. For instance, **Heart Rate Variability** (HRV) ( $n = 2$ , 3% of the 66 papers with quantitative studies) measures the variability of time interval between heartbeats [Shaffer and Ginsberg, 2017]. As elevated levels of stress can be indicated by low HRV, this could also be an indicator of lower levels of trust. Another example is **Galvanic Skin Response** (GSR) ( $n = 2$ , 3%) which measures the intensity of an experienced emotion with the electrical conductance of the skin, which varies with sweat [Rosenzweig, 2015]. High levels of stress can be generally indicated by high GSR, and hence, could be potentially linked to lower levels of trust [Morris et al., 2017]. However, in our corpus, no relationship has been found between these measures (HRV and GSR) and trust [Frison

et al., 2019; Gupta et al., 2019; Wintersberger et al., 2017].

Another was is **Electroencephalography** (EEG) ( $n = 2, 3\%$ ), which records activity of the brain. This approach is more promising as there is some evidence that the predominant brain areas correlated with trust are the frontal and occipital ones [Wang et al., 2018], but the papers in our corpus either did not deeply explore the EEG data [Gupta et al., 2019] or used it primarily for a preliminary model construction [Ajenaghughrure et al., 2019]. Finally, **hand trajectories** [Freeman, 2018], easily captured with a computer mouse has recently been shown to reflect the evolution of decision making as well as hesitations [Freeman, 2018; Maldonado et al., 2019]. While the relationship between hand trajectory and trust is yet to be determined, this measure can provide additional information in comparison with integral measures discussed above. Research community could benefit from:

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

## 2.10 Qualitative Methods

Qualitative methods produce less structured data than the quantitative ones. They might thus aid in discovering new aspects of trust and build new theories [Lewicki and Brinsfield, 2011]. We identified 10 qualitative methods to *collect* data in non-retrospective or retrospective ways among 34 papers with qualitative studies. We also identified 3 methods to analyze the collected data.

### 2.10.1 Non-retrospective Methods

Only a small number of studies use qualitative methods *while* a participant is interacting with the system. **Think-aloud protocol** ( $n = 3$ , [Buçinca et al., 2020; Drozdal et al., 2020; Frison et al., 2018]) can generate authentic and spontaneous reactions of the participants as these ones are not given any prompts to speak up. Moreover, this method avoids memory distortion effects, which sometimes happens with methods used post experiment. The papers use this method to investigate participants' decision-making with a system and what role trust played in the process. **Observations** in the field ( $n = 1$ ; [Yang et al., 2016]) is used to understand doctors' daily routine and decision-making process. However, this method might not be appropriate as

trust does not always translate in a behavior (see section 2.4.2). It remains useful for preparing potential interview questions about trust post experiment.

### 2.10.2 Retrospective Methods

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

**1) Interview-based methods. Semi-structured** interviews are the most common type of interviews ( $n = 24$ , 82.6% of 29 studies with interviews) as they both provide control over the topic while leaving room for unexpected insights [Magaldi and Berler, 2020]. In our corpus, they primary focus on understanding participants' general experience with the system, decision-making process or general perceptions and attitudes towards a system. Only 3 semi-structured interviews primarily focus on Human-AI trust [Maurer et al., 2018; Sultanum et al., 2018; Yang et al., 2016], rather than considering it as one of the multiple factors of users' experience to evaluate.

**Non-structured** interviews ( $n = 1$ , [Gupta et al., 2019]) and **in-depth** interviews ( $n = 1$ , [Jin et al., 2020]) have been used in our corpus to study participants' general experience with AI. They both allow for gathering more personal, sensitive or confidential information, which is especially appropriate for discussing a topic of trust. In particular, in-depth interviews tend to be longer, useful to build a relationship with an interviewee and to ask more detailed questions.

We found one instance of **focus groups** (or group interviews) to study participants' general attitude to AI [Woodruff et al., 2018]. Focus groups are less time-consuming and less expensive than the above types of interviews but there is a risk that the responses of one person bias the rest of the participants. Moreover, some participants might get too shy to express their real opinions, especially for such personal topics as trust [Nyumba et al., 2018]. With this method, the paper in our corpus explore trust of people with a specific background - members of marginalized communities [Woodruff et al., 2018].

The following interview-based methods are especially appropriate to study the *dynamics of trust*, i.e. how trust evolves over time. **Critical incident technique** [Flanagan, 1954] is a set of procedures used to collect data from narrated past experiences (or observations) to identify and

brainstorm about important events related to a pre-defined problem [Andersson and Nilsson, 1964]. When applied to trust, it is especially useful to study real life cases in which trust was established, destroyed or repaired [Münscher and Kühlmann, 2011]. Researchers directly ask participants what aspects of others' behavior was important for trust weakening or strengthening. This information can in turn be applied, for example, towards improving patients' experience during a medical visit [Wendt et al., 2004; Yañez Gallardo and Valenzuela-Suazo, 2012] or enhancing intercultural business negotiations [Münscher and Kühlmann, 2011]. Although this method is an established method, we did not find it in our corpus.

**Repertory grid**<sup>13</sup> is an interview-based method relying on card sorting. During the interview, participants establish links between different elements (words, objects) which is useful to make some concepts emerging. The main advantage of this method is to minimize the researchers' influence on a study by reducing the interviewer's input and maximising the interviewee's output [Ashleigh and Meyer, 2011]. Researchers are thus less prompted to introduce their preconceived assumptions about whether and how an element of the studied environment affects trust. They can then study participants' understanding of trust, its development, breakage and repair processes with a reduced interviewer bias, which enhances validity of the data. This method has also been used in Human-Human trust [Lewicki and Brinsfield, 2011], but not in our corpus, probably because it is quite time consuming. It is more suitable for studying small groups where individual differences play an important role.

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

Our analysis also reveals the lack of information to evaluate or replicate interviews as well as to compare the findings between papers. For instance, only few papers provide question examples ( $n = 5$ , 17.2% of 29 papers with interviews) or describe the general topics of the interviews ( $n = 12$ , 41.4%). Among them, only [Cai et al., 2019; Jin et al., 2020] mention they conducted a pilot study to identify the prominent questions and to refine their wording to study Human-AI trust. It is thus difficult to assess to what extent the questions were really understandable for the participants.

<sup>13</sup> An example of repertory grid based on studying trust in organizational settings. The interviewer presents a random pair of words, *elements*, related to work settings: face-to-face contact, lengthy detailed contracts, frequent emails etc. The participant indicates if these elements are similar/dissimilar with regards to trust, and explains why: "face-to-face contact and lengthy detailed contracts are similar, because they represent 'engagement' (keyword)" or "dissimilar, because the former is associated with 'transparency' (keyword) and the latter with 'bureaucracy' (keyword)". Later on, the participant indicates whether there is a link between each of the combinations of elements and keywords, which later will be translated into a cognitive map with points proximity determined by words similarity [Bachmann, 2011] (see more in Chapters 13 and 14 of [Lewicki and Brinsfield, 2011].)



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

**2) Non Interview-based Methods.** Non interview-based methods are generally less used. However, they might be useful as they are simple and fast to collect data. For instance, some studies just let participants leave any **comment** they wish after the experiment ( $n = 2$ ) or opt for a **open-ended question** ( $n = 3$ ) (i.e., what-how-why questions) to study participants' general attitude to AI [Kolasinska et al., 2019], to understand participants' decision-making [van Huysduyнен et al., 2018], and to directly investigate participants' trust [Glass et al., 2008].

Another method is **UX curve** ( $n = 1$ , [Frison et al., 2018]), used for understanding the reasons behind long-term system use or abandonment (more about it here [Kujala et al., 2011]). Participants draw a line which represents their experience with a system, saying out loud what events changed it and if they affected it positively and negatively and by how much. This method serves to get accurate and chronological insights about what, in what direction and by how much affected participants' trust and experience during an interaction with a system.

Finally, **open-card sorting** ( $n = 1$ , [Drozdal et al., 2020]) identifies what are the most important factors for participants' trust. Participants rank various pre-selected prompts and few ones introduced by them in order of importance for their trust in a system (for more details about the method [Spencer and Warfel, 2004]). Overall, it is not yet established how efficient these methods, marginally used, are to assess trust in our context. More work is needed to more systematically compare these qualitative measures.

### 2.10.3 Analyzing Qualitative Data

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

**Grounded Theory** aims to generate hypotheses based on the themes and categories found in the qualitative data. Consequently, the findings are the presentation of a new theory that includes the core themes [Floersch et al., 2010]. Usually, these themes emerge from the data after it is annotated with *open* and *axial* coding. Open coding is aimed

at summarizing small portions of text with one or two words - codes, and axis coding organizes these codes into groups. Researchers then study how these groups interact with each other to establish a theory or framework [Floersch et al., 2010]. 6 papers in our corpus analyze their data in this manner [Alan et al., 2014; Chromik et al., 2020; Liao et al., 2020; Park et al., 2019; Veale et al., 2018; Yu et al., 2020].

Unlike Grounded Theory, **Thematic Analysis** focuses on identifying the themes most relevant to the research objectives of a paper, without necessarily exploring the relationship between them<sup>14</sup>. Therefore, the main findings are presented as a description of the most important themes. 16 papers in our corpus state using Thematic Analysis as new theory development was not their objective.

Rather than analyzing what participants say, **Discourse Analysis** focuses on how they say it [Potter and Edwards, 1996], i.e., the type of vocabulary, grammar, non-verbal communication used. The advantage of this approach in comparison with the above mentioned ones is that it could supply researchers with insights from the cues which participants are unlikely to voluntarily control. 1 paper in the corpus analyzed the way participants spoke to robots before making a decision, particularly the amount of words used, while studying their trust [Xiao et al., 2007].

In complement to the previous methods identified from the reviewed corpus, we also introduce a method used in Human-Human trust literature. **Hermeneutics**<sup>15</sup> is suitable to analyze not only interviews transcript but also existing stories published in popular media outlets (e.g. [Ito, 2018; Ross and Swetlitz, 2017]). With the rising media coverage and popularity of workshops and webinars related to AI, researchers should **consider Hermeneutics to interpret the current narratives (see G16 below)** as an alternative data source on users' trust.

This method is most widely employed by historians and theologians to interpret human actions and their outcomes [Mantzavinos, 2020]. It offers a toolbox for finding patterns and common threads in texts to justify their interpretation and theories drawn from them. Gerard Breeman, researcher in trust and politics, finds hermeneutics useful for investigating the reasons why people trust and why exactly those reasons were given in that specific context [Breeman, 2011]. Before analyzing the text, a research determines key factors that can influence trust in a certain scenario based on theory. This framework becomes a guiding thread for a researcher while analysing the text to find passages that either confirm or go against the theory. In the final

<sup>14</sup> Though developing a new theory is not a goal of Thematic Analysis, the emerged themes and their relationships can be further studied with the Grounded Theory Approach for a potential theory or framework development [Floersch et al., 2010].

<sup>15</sup> You can find a more detailed description of the method in Chapter 15 of [Lewicki and Brinsfield, 2011]. For more trust studies, using hermeneutic analysis, refer to Breeman [Breeman, 2006] and von Sinner [von Sinner, 2005].

step, a researcher update the framework they established incorporating new insights from the text [Breeman, 2011]. The main limitations of hermeneutics is that this method relies on preselected concepts, which might lead to a biased interpretation and sub-optimal understanding of the case. Plus, it focuses on analyzing a very specific event, which could hinder results' generalization.

To sum up, very few qualitative studies ( $n = 4$ , 11.8% of 34 qualitative papers) consider Human-AI trust as their central focus. The rest of the qualitative studies in our corpus view trust as one of the multiple factors of users' experience. This finding is similar to the one by [Frison et al., 2019], which urges to use qualitative methods for deepening our understanding of Human-AI trust as little is known about its nature and how different it is from Human-Human and Human-machine trust. Some qualitative methods for studying trust stemming from other domains could be found useful in the Human-AI Interaction research, too, being that for studying different aspects of trust (i.e. dynamics of trust) or for having a tool to analyze a different type of data (i.e. media reports). We encourage the community to:

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

## 2.11 Summary of Findings for Trust Measures

In Sections 2.9 and 2.10, we summarize and categorize the quantitative and qualitative measures used in the experimental protocols to evaluate Human-AI trust in the decision making context. Similarly to the sections related to the protocol, we provide guidelines on how to improve the use of measures and highlight the research opportunities that need further investigation, drawing from social and cognitive sciences literature.

Among the quantitative measures, we identify two major groups - questionnaires and trust-related behavioral measures. The reviewed papers contain 21 questionnaires, and as many of them are borrowed from other fields (e.g., Human-Human trust, Human-Robot trust), they are modified to have appropriately adapted terminology. If this is the case, the modifications have to be described and their impact on the integrity of the questionnaires - verified through psychometric statis-

tics (G12). While choosing the questionnaires, one has to look out for well-established ones that equally comprise the theoretical concepts related to trust (see Chapter 2; G11), and section 2.9 provides examples of such. The questionnaires may contain different number of questions, some consisting of 1 question only, but it is not clear yet whether single-item questionnaires capture trust as well as other measures (RO8). Among trust-related behavioral measures (the wording preferred to “behavioral trust measures”, G13), derived from behavioral logs, we identify the ones based on they way people followed recommendations (decision time, reliance, compliance, switch ratio) and on physiology (heart rate variability, galvanic skin response, electroencephalography). Trust-related behavioral measures are useful to assess participants’ decisions vis-à-vis the system’s performance and to measure their appropriate, under- or over-levels (G14). Physiological measures remain a promising approach to evaluate trust, especially its dynamics, but more fundamental correlates need to be established (RO9).

We categorize qualitative methods into non-retrospective and retrospective. The former means that the data is collected while participants interact with the system, and these methods are think-aloud and observations. Retrospective means that the data is gathered after the interaction, and they are notably interview-based: individual interviews (semi and non-structured) and focus groups. Non-interview-based methods are open-ended questions, UX curve, and open-card sorting. We find that reporting on qualitative studies should experience more of empirical rigor (G15); for instance, not providing examples of questions asked inhibits studies’ replication. The qualitative data is analyzed (when the method is reported) with grounded theory, thematic analysis, and discourse analysis. We propose additional methods underlooked in Human-AI trust literature like Critical Incident technique and Repertory grid for data gathering and Hermeneutics for data analysis (G16).

## 2.12 Discussion

Trust has recently emerged as key concept in Human-AI Interaction. While many studies investigated the factors influencing trust, our approach focuses on **how to evaluate trust** in the context of AI-assisted decision making. This survey offers a lens on existing methodologies and highlights the difficulties of properly studying this multi-faceted and dynamic construct. This survey also provides an opportunity to

improve validity and replicability of future experiments by proposing practical guidelines. Finally, it identifies challenges and research opportunities. We now discuss these different contributions.

### 2.12.1 *Main Findings and guidelines*

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

Our first finding (**F1**) is that trust definitions are often incomplete or even not provided. Established definitions exist in related fields [Evans and Krueger, 2009; Mayer et al., 1995; Rousseau et al., 1998] as well as HCI [Lee and See, 2004], but few studies explicitly mention any. However, trust is a multi-dimensional construct and Human-AI interaction is a recent field of research, it is, thus, important to clearly define trust to avoid conflicting terminology and misunderstanding in the community. In particular, we found (**F2**) that the three key elements of trust are not always incorporated in the reviewed studies. The sense of *vulnerability* is often missing or questionable due to the lack of realistic outcomes in the experiments. The system is not always introduced in a way that participants have *positive expectations*. Finally, several methods capture participants' behaviors while trust is an *attitude*. Consequently, several papers investigated constructs related to trust such as distrust, confidence or reliance, rather than trust. It is important that Human-AI community adopts the theoretical evidence establishing the difference between these related constructs [Evans and Krueger, 2009; Meyer and Lee, 2013; Sitkin and Roth, 1993].

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

We also found that (**F3**) there is a large variability among the designs and the measures used to assess trust. For instance, there is no "standard" task, nor procedure, nor questionnaire. While it could be explained by a variety of scenarios in the real life, it also appears that the relevance or validity of existing methods are still under debate. For instance, it is not clear to what extent behavioral, especially physiological, measures can be used as a proxy to capture trust [Alós-Ferrer and Farolfi, 2019; Naef and Schupp, 2009] or whether single-item question-

naires are as robust as multi-question questionnaires [Robert, 2002].

We derived several guidelines from this finding. We suggest to consider the different interaction flows (**G6**) illustrated in Figure 2.5 before choosing the final one to ensure it fits the research question, the envisioned scenarios as well as the target measures. We also recommend using established questionnaires that comprise all the key elements defining trust (**G11**). Lastly, more information should be reported regarding the modifications and analysis performed (e.g. in questionnaires, **G12**) and methods used (e.g. interview questions, **G15**) to foster replicability and increase scientific rigor.

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

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

### 2.12.2 *Challenges and Research Opportunities*

We identified two main classes of research opportunities. The first one is further investigation regarding **methodologies** to study AI-assisted decision making. We already acknowledged one major challenge of studying trust experimentally: the conflict between the importance of the dynamics of trust and the constraints of laboratory experiments. Future work could investigate novel practical methods which do not break the interaction flow and do not make the experiment longer (**R07**, **R08**, **R09**). In particular, several novel measures have been recently introduced, e.g. EEG, mouse trajectory. Future research could investigate to what extent these components have an impact on Human-AI trust and whether there is a relationship between them and trust (**R09**). More generally, it is important to foster connections between the Human-AI and Human-Human trust communities and to investigate how to transpose methods from other fields to the

Human-AI interaction one. We propose several quantitative and qualitative methods used for studying Human-Human trust to apply for Human-AI trust, but it is not an exhaustive list. This work hopes to promote further exploration of other fields studying trust to enrich the pool of trust methods in Human-AI Interaction community. We would also like to note that within the Human-AI Interaction community, we covered the part represented by ACM Digital Library, and hence, further exploration of methods used in this community in AAAI Digital library, IEEE Xplore and HCI journals will be beneficial.

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

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

### 2.12.3 *AI in AI-based decision-making systems*

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

First, this approach allowed us to extract generic guidelines that could be used by designers, developers and HCI practitioners independently of the type of AI involved in the system, as long as the goal of the algorithm is to provide recommendations to users in a decision-making process. That said, we are aware that the study of trust is, to some extent, context-dependent, and certain results may change according to the type of system considered. For instance, we mentioned several studies that indicated that 60% - 70% accuracy could be considered to

be a threshold for investigating users' trust in AI-embedded decision-support systems [Yin et al., 2019; Yu et al., 2016; Zhang et al., 2020]. But this threshold may vary according to the application domain and the task at hand. Nonetheless, we believe that the proposed guidelines capture the fundamental elements that ensure trust to be assessed in this context.

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

### 2.13 Conclusion

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



#### TAKE AWAY MESSAGES

##### *Contributions:*

- A demonstration of Human-AI trust in terms of vulnerability, positive expectations, and attitude to disentangle it from confidence, distrust, reliance, compliance, and trustworthiness;
- An exhaustive presentation of the variety and complexity of the methods to study Human-AI trust in the decision-making context;
- A structured discussion of the current Human-AI trust protocols highlighting flaws in methodologies with a stronger link to the Human-Human Trust community;
- A set of guidelines and research opportunities to improve research quality in Human-AI Trust, highlighting the need for a greater empirical rigor and standardization in the community.

# 3

## *Human-AI Trust in the Context of Decision Making through the Lens of AI Practitioners and Decision Subjects*

As demonstrated in Chapter 2, the predominant approach to study Human-AI trust in the context of decision making relies on controlled lab experiments during which users interact with AI mock-ups [Glikson and Woolley, 2020; Vereschak et al., 2021]. However, in a survey of trust in Human-AI interaction in medical decision making, Browne et al. [2022] show that trust can get impacted in other stages of AI deployment cycle before and after initial users' interaction. Together with a survey on Human-AI decision making [Lai et al., 2021, preprint], they call for a further investigation of Human-AI trust in real world contexts. Additionally, as Jakesch et al. [2022] demonstrate that the ethical values instilled in AI can have different significance and meanings to different groups of people (users, AI practitioners, crowd-workers), it can also be true for trust. For example, Lockey et al. [2021] identify that different types of users do not encounter the same issues related to Human-AI trust. Notably, users who are domain experts are especially sensitive to trust factors that can put their skills and knowledge under questions, while common users are more concerned with unfair or unethical impact [Lockey et al., 2021].

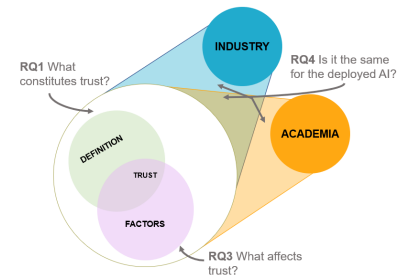
In this chapter, we<sup>1</sup> focus on how people associated with AI-embedded systems deployed in the market understand Human-AI trust and to which extent it differs from the academic findings. Specifically, we explore how they define it with their own words, whether they differentiate trust from other related theoretical constructs and what they think can affect Human-AI trust in the context of decision making. Based on the comparison between the reflections of the people from the industry and the academic findings about Human-AI trust definition and factors, we highlight research opportunities and design implications for academic researchers and people who develop and design AI-embedded systems.

### 3.1 Objectives and Approach

This chapter addresses two research questions of this thesis (Figure 3.1): **RQ4** *Do the academic postulations about trust definition, factors, and evaluation of Human-AI trust reflect the real world considerations?* (in particular, **DEFINITION** and **FACTORS** through the lens of **INDUSTRY**) and **RQ3** *What factors affect Human-AI trust in the context of decision making?* (**FACTORS** through the lens of **ACADEMIA**). This chapter shares 3 objectives. The first one is to complete our understanding of Human-AI trust through the academic lens and provide a landscape of factors studied in the literature on Human-AI trust in the context of decision making (RQ3). This is done through the extension of the systematic literature review conducted in Chapter 2.

The second objective is to investigate how people associated with AI-embedded systems deployed in the market define Human-AI trust and what they think affects it in the context of decision making. Specifically, we interview two groups of people that come before and after users in the decision-making chain: **AI practitioners** and **decision subjects**. **AI practitioners** are the stakeholders involved in different aspects of system design and deployment in the field, their roles range from AI developers to project managers. **Decision subjects** are the stakeholders who do not directly interact with AI-embedded systems, but who are affected by decisions made by users based on the AI's recommendations, e.g. in the medical context doctors are users and patients are decision subjects. Lastly, the third objective is to compare the academic findings and the interviewees' reflections (RQ4). As the experiences of the interviewees can yield contextual nuances about AI systems deployed in the market, comparing their views with the findings of academic literature on Human-AI trust can inform us about

<sup>1</sup> Main portion of this chapter is a cooperative work which led to a submission currently under review. Thus, any use of "we" in this chapter refers to the authors of this work: Oleksandra Vereschak, Fatemeh Alizadeh, Gilles Bailly, and Baptiste Caramiaux.



**Figure 3.1:** Research questions investigated in this chapter.

the extent to which the current state of art is represented in the real world settings and vice versa. We also derive research opportunities and design implications for academia and industry based on this comparison.

This chapter is structured as follows. We first present the methodology behind the semi-structured interviews with AI practitioners and decision subjects. We then present and discuss results alongside with research and design implications. We start each results' subsection with reporting the interviews findings, which we discuss in the light of academic findings from the systematic review, highlighting the similarities and differences between them. Based on this, we provide research and design implications stemming from our discussion targeted at academic researchers, AI practitioners or both.

## 3.2 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. A deeper analysis of the previous work is provided all along the document as we compare our interview findings to the academic literature.

### 3.2.1 *Human-AI Trust*




Human-AI trust literature has two major themes of interest: defining what trust is and what factors affect it. A systematic literature review on Human-AI trust in the decision-making [Vereschak et al., 2021] defines Human-AI trust through three key elements, all encompassed in the trust definition by Lee and See [2004]: 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. These three elements of trust are important as they directly affect the design of experimental protocols in academic research [Vereschak et al., 2021]. For instance, the necessity to make users' decisions have an impact, such as earning or losing money, in order to induce a sense of "vulnerability". This definition of Human-AI trust builds primarily on theoretical works. Little is known about how different stakeholders, such as users, AI practitioners, or decision subjects define trust in practice (i.e. in real use and development cases) and how their opinions reflect the existing academic literature.

The second major theme of interest in the study of Human-AI trust concerns the factors that affect trust in interaction. The predominant approach to study such factors is laboratory experiments (e.g. [Ghai et al., 2021; Jensen et al., 2019; Lai and Tan, 2019; Müller et al., 2020; Park et al., 2019; Suresh et al., 2020; Sutherland et al., 2015; Tokushige et al., 2017; Yin et al., 2019; Zhang et al., 2020]). For instance, [Yin et al., 2019; Yu et al., 2017] study how different levels of accuracy affect Human-AI trust. These experiments generally rely on an AI mock-up rather than a real system deployed in the real-world context [Vereschak et al., 2021]. Moreover, experiments involving quantitative assessment of trust factors control only a limited number of factors (typically 1-3). The qualitative approach is less predominant, but has the advantage to favor external and ecological validity (rather than internal validity). These studies investigate more generally users' experience with AI-embedded systems [Eiband et al., 2019; Frison et al., 2018; Luger and Sellen, 2016; van Huysduynen et al., 2018], their design needs [Cai et al., 2019; Yang et al., 2016], their perception of AI [Kolasinska et al., 2019], and transparency [Jin et al., 2020; Schneider et al., 2019]. One exception, Glass et al. [2008] set users' trust at the center of their research question, and propose that AI explanations can address most of the trust issues the users highlighted. Therefore, there is a need for a broader overview of the factors that affect trust between humans and AI in the context of decision making from real-world use cases as well as from the literature. In addition, most of this work targets a single type of stakeholder, who are the users of the systems under study, even for the domains involving decision subjects [Cai et al., 2019; Jin et al., 2020; Yang et al., 2016], which we review in the following section.

### 3.2.2 Stakeholders of AI-embedded Systems for Decision Making

AI-based systems assisting decision making involve numerous stakeholders. There are several proposed typologies of stakeholders for this context, with different levels of granularity and overlaps in terminology [Ayling and Chapman, 2022; Deshpande and Sharp, 2022; Güngör, 2020; Jakesch et al., 2022; Scott et al., 2021; Yurrita et al., 2022]. We singled out the groups of stakeholders that are most linked to AI-assisted decision making<sup>2</sup> (see Table 3.1): **users** [Ayling and Chapman, 2022; Deshpande and Sharp, 2022; Güngör, 2020; Jakesch et al., 2022; Scott et al., 2021]; **AI practitioners** [Ayling and Chapman, 2022; Deshpande and Sharp, 2022; Güngör, 2020; Jakesch et al., 2022; Scott et al., 2021; Yurrita et al., 2022]; and **decision subjects** [Ayling and Chapman, 2022; Yurrita et al., 2022].

<sup>2</sup>In the context of this chapter, we focused on stakeholders that are related to the development or the use of AI-assisted decision making systems. Additional stakeholders however exist, such as regulators and policy makers, whose contributions, although interesting, are out of the scope of this chapter.

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 3.1:** The different stakeholders involved in the use of AI assisted decision making systems. This chapter focuses on AI practitioners and Decision subjects, two stakeholders who received less attention in the Human-AI trust literature.

The stakeholders that have received the most attention in the literature are the users of the systems in play [Lai et al., 2021, preprint]. Researchers have repeatedly pointed to the need to explore and assess users' trust in these systems to facilitate their adoption (e.g., [Bisantz and Seong, 2001; Schoeffler et al., 2021; Seong et al., 2002]). In contrast, AI practitioners and decision subjects received less attention in the literature. Research looking at decision subjects' trust in AI is still scarce. The literature primarily focuses on general design needs [Lyons et al., 2022; Scott et al., 2022] and experiences with AI [Park et al., 2021; Veale et al., 2018], their understanding of AI fairness [Brown et al., 2019; Gemalmaz and Yin, 2022; Lee et al., 2019; Woodruff et al., 2018] or XAI (explainable AI) [Barocas et al., 2020; Lima et al., 2022; Schoeffler et al., 2022]. However, as far as we know, only few studies looked at their perceptions of trust and the factors they perceive as affecting trust in the context of AI-assisted decision making [Ammitzbøll Flügge et al., 2021; Ferrario and Loi, 2022; Okolo et al., 2021; Ramesh et al., 2022]. Ammitzbøll Flügge et al. [2021] and Okolo et al. [2021] consider the importance of trust between users and decision subjects, but do not examine what factors contribute to decision subjects' trust in AI. Ferrario and Loi [2022] and Ramesh et al. [2022] explicitly focus on decision subjects, but do not compare their findings with the other stakeholders involved in AI-embedded systems.

Similarly to decision subjects, research mainly focuses on human-centered AI values different from trust, e.g. interpretability or explainability [Kaur et al., 2020], fairness and accountability [Hong et al., 2020; Kaur et al., 2020; Liao et al., 2020; Veale et al., 2018]. Only one paper sets trust at core of the research question; yet, it mostly looks on how AI practitioners establish trust among themselves while working with data [Passi and Jackson, 2018]. A more global perspective of AI

practitioners on how they build Human-AI trust is yet to be explored.

To summarize, users are at the center of Human-AI trust research in the decision-making context while AI practitioners and decision subjects are also key stakeholders of such systems. Moreover, most studies are conducted in the lab with AI mock-ups and thus, they do not inform about the current understanding and practices around Human-AI trust in real-world organizational settings [Glikson and Woolley, 2020]. In particular, trust can mean different things for different stakeholders [Yurrita et al., 2022], but currently a holistic overview of Human-AI definitions and factors is lacking. Our article investigates Human-AI trust through the lens of AI practitioners and decision subjects and compare their views with academic findings to fill these gaps.

### 3.3 Methodology

The project spanned over one year starting in 2021. In this section, we illustrate our methodology by starting with our positionality in relation to this research, which motivated it, but which also delimited our analysis.

In the interest of reflexivity, we explicitly position ourselves as four HCI researchers operating in the academia in Western countries, conducting research in the context of Human-centered AI using both quantitative and qualitative methods. Two of us mainly build and use AI models and the others two study how people use AI systems in the context of decision making. In particular, the primary investigator has three years experience in Human-AI trust research. The motivations of this work build upon the recent experiences of this group of researchers and are threefold: 1) we realized that there was no consensus about “what is trust” and “what contributes to trust” when discussing with colleagues in HCI, AI and robotics, while they agreed on its importance; 2) we observed a discrepancy in the academic literature between the theoretical definitions of trust (in AI and social sciences) and the way it is investigated in controlled laboratory experiments (typically, the frequent absence of vulnerability); 3) we noticed that the topic of trust has an increasing impact on the start-up and industrial ecosystem involving AI, in which we are also involved.

Based on these elements, we came to these three research questions:

- How do AI practitioners and decision subjects define Human-AI trust in the decision-making context?
- What do AI practitioners and decision subjects think affects Human-AI trust in the decision-making context?
- What are the differences between the interviewees and the literature in the way they define Human-AI trust and the factors they propose?

### 3.3.1 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 channel. 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), 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<sup>3</sup>, 7 with AI decision subjects).

The AI practitioners are based in Europe and Oceania, and each worked for a different company. The AI decision subjects are all based in Europe and had been affected by AI decision making in three different risk-sensitive areas. The participation in the study was on a voluntary basis. Table 3.2 provides an overview of the decision subjects’ backgrounds and in what context they received a Human-AI decision. Table 3.3 provides an overview of the AI practitioners’ backgrounds, their positions in the company, and the application areas of AI. Three participants work on Explainable AI or XAI (two are responsible for implementation and research, and another is the company’s CEO, chief

<sup>3</sup> We initially contacted 14 AI practitioners, 5 of them did not reply, and 2 did not have availability for an interview

<b>Id</b>	<b>Background</b>	<b>Decision Context</b>
DS1	Software developer	Job application
DS2	Medical student	Phone contract
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 3.2:** Characterization of decision subjects, notably their background and in what context they received a Human-AI decision.



<b>Id</b>	<b>Role</b>	<b>Background</b>	<b>Organization</b>	<b>Type of AI</b>	<b>AI Application</b>
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

executive officer). Three other participants are senior project and product managers, and one more is the company's CEO.

### 3.3.2 Interview Protocol

We conducted semi-structured interviews of the recruited participants. The questions were compiled by two 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 also choose whether or not 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 3.4) evolving around: the context with respect to their interaction with AI, Human-AI trust definitions, trust factors, and trust evaluation. In this chapter, 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 3.4) 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 3.4). There were 8 questions in total as approximate guidance for the interviewers (Appendices D and E). When needed, they deepened the topic with follow-up questions about all the stakeholders involved

**Table 3.3:** 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.

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

in an anecdote, clarifying theoretical terminology, possible solutions to a described challenge, and whether a proposed factor always has effect on Human-AI trust.

### 3.3.3 *Analysis of the Interviews and Comparison with the Academic Literature*

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 read all interviews at least twice and independently identified pertinent phrases and coded them. Based on the thematic analysis method [Clarke and Braun, 2013], the first and second authors compared and finalized the list of selected phrases and fine-tuned the wording of the codes. Codes were organised under a series of sub-themes, which were themselves organised under four main themes: one about the definition of trust and three about its factors (as described in Section 3.3.4).

Once the themes were created, we analysed how they were addressed in the academic literature. To achieve this, we reviewed 113 empirical studies from the academic literature on Human-AI trust in the context of decision making. For each article, we annotated both the elements of trust and the trust factors investigated and discussed in these studies. These studies were selected following the same methodology as [Vereschak et al., 2021]: we include all their articles (83) as well as 30 articles published after their review was conducted (from January 2021 onward). Henceforth, when mentioning “academic literature”, we refer to this scope. The full description of the selection method is in Appendix F and the full table summarizing all the trust factors is in

**Table 3.4:** Structure and examples of questions per each group of participants. Data analysis of this chapter mostly relies on answers around understanding and factors of Human-AI trust. A full list of questions is in Appendices D and E.

## Appendix G.

### 3.3.4 Result presentation

We choose to report the results as distinct sections for clarity. Each section relates to a theme stemming from the analysis. Section 3.4 reports on the key elements of trust and how they differ from trustworthiness. The three following sections report on the Human-AI trust factors. The first author, inspired by the terminology from the existing trust frameworks [Adams et al., 2003; Bindewald et al., 2018; Hancock et al., 2011; Hoff and Bashir, 2015; Schaefer et al., 2014, 2016], named these factor-related themes: “Socio-Technological Context” (Section 3.5), “System’s Development and Design” (Section 3.6), “People’s Preferences and Experiences” (Section 3.7).

For both definitions and factors, we first report the views of the interviewees (*Interview findings*) and then compare them to the academic literature (*Literature comparison*). The aims are 1) to highlight what **AI practitioners** could further consider regarding Human-AI trust in their working processes or to motivate further investigation to understand why they do not discuss some factors, and 2) to bring nuances and identify gaps in the **academic literature** on Human-AI trust in the context of decision making. Some of these research opportunities apply to **both**.

## 3.4 Trust and Trustworthiness in Human-AI Interaction

In this section we discuss the interviewees’ reflections on the key elements of trust in Human-AI interaction and interviewees’ opinions about what makes AI trustworthy and the link with trust. Findings are summarized in Table 3.5.

### 3.4.1 Key Elements of Human-AI Trust

*Interview findings* The interviewees characterize Human-AI trust with 4 keywords: **risk associated with a decision** (P2, P4-P7, DS2, DS4-DS6), **positive expectations** (P6, DS5, DS7), **task complexity** (P2, P4-P6, DS5), and **attitude** (P2-P6).

The interviewees identify **risk associated with a decision** as an element that gives trust in AI foundation to start existing: “*When my physical integrity or money is at risk, it makes sense to consider trust there,*

	Interviewees	Academic Literature
<b>Trust</b>	Reaction to the system Risk Positive expectations <u>Task complexity</u> Attitude (not explicitly)	Reaction to the system Risk Positive expectations Attitude
<b>Trustworthiness</b>	Property of the system Fairness Robustness Transparency Privacy (not explicitly) <u>Human-approved</u> (linked to Accountability) <u>Reputation of the company</u> (linked to Accountability) <u>Human-like</u>	Property of the system Fairness Robustness Transparency and explainability Privacy Accountability <u>Reproducibility</u> <u>Generalization</u>

when there is something important for me [at stake]" (P4). The nature of risk is also called "vulnerability" (P5) or "responsibility" (P2). P4 states that risks such as economic loss or threat to health or life are universal: "... a foundation [for defining risk] would be the physical needs and individual and social integrity from the Maslow's Hierarchy." Decision subjects, notably, refer to risk as something that can impact their health (DS2, DS4, DS5, DS6) or financial stability (DS4, DS5, DS6). Otherwise, what is considered risky is person-dependent, because "not everyone has the same priorities" (P4). For example, DS4 evokes that even Tinder recommendations can still put them in a position of vulnerability. In this scenario, DS4 feels there are moments when "the algorithm says that I am ugly", hence they are faced with "something about myself that I do not want to accept" (DS4).

Then, P6, DS5, and DS7 state that for trust to emerge, people must have **positive expectations** that AI will help them achieve their goal. The goal is defined as "the best answer in the shortest time" (DS5, DS7). P6 highlights that AI recommendations must support the goal of people interacting with or affected by the system, it is important that "the owner [of an AI-embedded system] will not recommend [the user] something in company's interest" (P6). DS5 agrees that it is important that AI recommendations "support humans in their work."

P5 also proposes that when users face a **complex task**, trust emerges as a tool to overcome 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» ". P2 and P6 describe "complex task" as a situation when a user

**Table 3.5:** Keywords presented by the **interviewees** and the **academic literature** when defining trust and trustworthiness. In comparison with the literature, the interviewees additionally mention task complexity for defining trust and reputation of the company, human-like and human-approved for trustworthiness.

cannot determine the quality of AI recommendation and, as a result, has many doubts around the final decision. DS5 echoes P2 and P6 and provides data analysis as an example of a complex task: *“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).

Lastly, the interviewees differentiate between trust and related behaviors. 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], it is very multifactorial”* (P4). Typical trust-related behaviors – reliance (deciding to use AI for decision making) and compliance (deciding to follow an AI recommendation) – do not mean users trust in AI. P3 says that even though *“becoming the client, that’s the sign of trust”*, there can be many reasons beyond trust for one not to use their AI. For example, a user can follow AI recommendations not because of trust, but because they have no other solution (P2). However, considering compliance and reliance is still useful, because they can serve as *“an indicator: as long as there aren’t too many complains, no negative comments, [...] and the user uses the solutions, we can consider that trust is not broken”* (P2).

*Literature comparison* Interviewees’ reflections coincide with the way the academic literature conceptualizes Human-AI trust [Jacovi et al., 2021; Vereschak et al., 2021], which consider three elements of trust (risk associated with a decision, positive expectations and attitude). The only difference is that the academic literature deems task complexity only as a trust factor (we discuss it more as a trust factor in 3.5.3), rather than a key element of trust. If task complexity is a key element required for trust to exist, experimental protocols might need to control it in addition to positive expectations and including risk associated with decisions. Moreover, considering task complexity could give another perspective for legal frameworks categorising AI-embedded systems. For example, in addition to dividing such systems based on risk associated with them (e.g., as in the report by European Commission [92]), complexity of the tasks they deployed for can be another axis of comparison.

There are also nuances in what the interviewees said about the three trust elements they have in common with the academic literature. The experience of DS4 feeling vulnerable when their appearance was judged by AI indicates that vulnerability goes beyond monetary losses or health hazards, which is the ways it is usually presented in the

controlled lab experiments where risk is associated to vulnerability [Vereschak et al., 2021]. Then, decisions to make in trust experiments often affect the user (intrinsic risk) [Vereschak et al., 2021]. However, it is arguable that decision subjects who suffer most considerable decision consequences (extrinsic risk, e.g., health implications, job opportunities). Therefore, differentiating intrinsic and extrinsic risks and including both where applicable in experimental protocols is important.

### 3.4.2 Key elements of AI Trustworthiness

*Interview findings* Four interviewees (P2, P4, P5, P7) explicitly differentiate trust in AI from AI trustworthiness, only P6 does not view these concepts differently. Contrary to trust which is seen as “human reaction” (P5), trustworthiness of AI is more related to a feature of the system (P2, P5), “whether the job has been well done” in designing and developing the system (P7). To be considered trustworthy, AI recommendations have to be fair, “human-like,” (P1), robust and transparent<sup>4</sup> (P1, P2). According to P1, to earn the label of trustworthiness, technical AI certification might not be enough and more evaluations with people have to be done, because trustworthiness “is beyond certification, and it would be something like «human-approved» [...]” (P1).

<sup>4</sup>Note that some of these concepts are also discussed in the following sections about factors that affect Human-AI trust. In this section, they contribute to the definition of trustworthiness, i.e. what makes AI trustworthy.

P4 states that trustworthiness of AI goes beyond the way system is designed and developed, it is rather about the reputation of AI practitioners working on AI: “For me, it’s not so much a question of AI, trustworthiness, it’s more between the individual himself and the entity or the organization that makes the system.” P4 further continues by providing an example that nobody is questioning trustworthiness of Google’s search results until the 2018 Google data breach scandal [Newman, 2018]: “It is the institution [behind the AI] that transmits trustworthiness to me. [...] nobody asks if it is the system behind that is biased, not biased, what we question is the institution.” DS1 echoes this reflection by stating that trustworthiness of AI is determined by the way the company is handling and using their data.

*Literature comparison* Our primary finding is that AI practitioners (and to some extent decision subjects) distinguish well between “trust” (human reaction and attitude) and “trustworthiness” (property of the system). Interestingly, interviewees explicitly mentioned three out of 8 key elements of trustworthiness expressed in the literature [Li et al., 2022]: “fairness”, “robustness” and “transparency”. They implicitly mentioned “privacy” when DS1 talks about the importance of the way a company handle their personal data. The term “accountability” was

not used, but interviewees reported on two related concepts: “human-approved” and “reputation of the company”. It could be that the term “accountability” is not widespread outside of the policy and research worlds. Only two elements from the literature [Li et al., 2022], “reproducibility” and “generalization” were not addressed by the interviewees. This is not very surprising as the literature on trustworthy AI concede that these two elements are less discussed than the others [Li et al., 2022]. One key difference with the literature is the introduction by P1 of the element “human-like” AI, i.e. AI that gives recommendations that a human would give. We found very few studies linking “human-like” and (un-)trustworthiness [Castelo et al., 2019; Patrzyk et al., 2017; Yu and Li, 2022]. Therefore, it is unclear whether human-likeness is a separate key element of AI trustworthiness or a factor that affects it.

	AI Practitioners	Decision Subjects	Academic Literature	
<b>Socio-Technological Context</b>	<b>Human-Human Trust</b>			
	<i>Trust in AI team</i>	P2, P4, P7	DS7	–
	<i>Trust in other users</i>	P3, P6	–	46; 87; 156; 266; 316
	<b>Time Dynamics</b>			
	<i>Long-term interaction</i>	P1, P2, P4, P7	–	68; 214; 266; 343; 396
	<i>Delay of AI recommendation</i>	–	–	50; 181; 273
	<b>Type of Task</b>			
	<i>Subjective evaluations</i>	–	DS4	179; 189
	<i>Task Complexity</i>	P6	DS5	7; 71; 95; 130; 249; 343; 346; 359; 388
	<i>Responsibility / risk</i>	–	–	49; 88; 115; 179; 214; 249; 298; 361
	<b>Marketing</b>	P4	DS1, DS7	
	<i>System terminology</i>			172; 189
<i>Reliability and values signaling</i>			95; 196; 292; 383; 392	
<i>First interaction</i>			351	

**Table 3.6:** Summary of the Human-AI trust factors related to socio-technological context discussed by AI practitioners and decision subjects and studied in academic papers on Human-AI trust in the context of decision making.

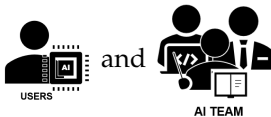
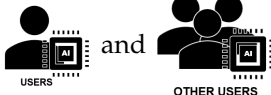


### 3.5 Trust Factors Related to the Socio-Technological Context

A first theme generated from the thematic analysis concerns factors related to the socio-technical context. These factors consider how the en-

vironment in which the interaction between humans and systems can affect trust. The interviewees discuss four of them: Human-Human trust, Time dynamics, Type of task and Marketing. Table 3.6 provides a summary by reporting these factors, the respondents who mention them, and the academic literature that addresses them.

### 3.5.1 Human-Human Trust

*Interview findings* Interviewees suggest that people’s trust in AI is strongly related to how much one trusts other stakeholders (human-human trust) in the socio-technical ecosystem. We distinguish four cases illustrated in Table 3.7. Case 1 is about trust between the **users and the AI team** (P2, P4, P7). If users (i.e. customers) 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). Case 2 is about the trust between the **users and other users** of the same system (P3 and P6). Indeed, 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). The interviews raise only one case (3) involving decision subjects: trust between **decision subjects and AI team**. For instance, DS7 cites the example of Elon Musk and Tesla, 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.

Trust between...	Interviewees	Academic Literature
Case 1 	✓✓	–
Case 2 	✓	✓✓
Case 3 	✓	–
Case 4 	–	✓

**Table 3.7:** 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 **interviews** and **academic literature** in the context of decision making.

*Literature comparison* The main difference between the interviews and the literature is that trust in the AI team (case 1 and 3) is generally absent in the literature. In contrast, respondents believe that this plays an



important role in the trust between humans and AI, particularly with regard to the trust between users and the AI team (case 1). It could be explained by the fact that most controlled lab experiments do not let their participants (“AI users”) interact with the AI team, while it is likely to occur in the real world. Another difference is that users’ trust in other users (case 2) is more emphasised in the academic literature than in the interviews [Brown et al., 2019; Ehsan et al., 2021; Jacobs et al., 2021; Okolo et al., 2021; Saxena et al., 2021]. Research shows that observing other users (especially colleagues) trusting the recommendations of the system can increase one’s own trust in AI [Ehsan et al., 2021; Jacobs et al., 2021]. However, from the interviews, AI practitioners serve as intermediaries between users and convey feedback as product reviews. Finally, the relationship of trust between decision subjects and other users (case 4) is absent from 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 [Okolo et al., 2021] and vice versa [Brown et al., 2019]. Our decision subjects did not discuss this, probably because they have a minimal interaction with the direct users (e.g. credit controllers) than those mentioned in the literature (e.g. medical doctors and caseworkers [Brown et al., 2019; Okolo et al., 2021]). Based on the above findings, we therefore propose the following implications:

- 1) **implications for academic researchers** : investigating the role of **trust between users and the team behind AI, as well as between decision subjects and the team behind AI**;
- 2) **implications for AI practitioners** : taking **trust between direct users and decision subjects** into account;
- 3) **implications for AI practitioners** : establishing **direct exchange of experiences between AI users** so that they can build trust without intermediaries.

### 3.5.2 Time Dynamics

*Interview findings* Interaction over time appears as an important factor affecting Human-AI trust for the interviewees (P1, P2, P4, P7). The development of trust is perceived as “a stimulus-response loop” (P4): through continuous interaction, users learn more about the system and adjust their expectations accordingly. Regarding AI practitioners, trust in AI tends to *decrease* over time as they “give limited credit to AI” (P1) because they become more aware of potential biases and caveats.

Indeed, “*the more I work with [AI], the more doubt creeps into my research*” stated P1. In contrast, AI practitioners seek to ensure that the user’s trust in the system increases over time, which is challenging if trust is low when first used. P7 says it is difficult to get users to understand that even if the AI recommendations are not perfect at first, if they wait and trust the company, the recommendations will improve as more data comes in: “*the hardest thing is to have the first [version of the AI-embedded system] with first results, now that we will be able to continue to collect your data, [...] we will be able to go much faster*” (P7). It results that AI teams mainly focus on gaining initial trust in AI to bring the product to market. Little is done to observe how trust evolves thereafter.

*Literature comparison* Both the interviewees and academic literature agree that Human-AI trust evolves over time because the interaction with the system adjusts one’s expectations [Colley et al., 2022; Okolo et al., 2021; Sultanum et al., 2018; Yu et al., 2017]. Nevertheless, our interviews (especially with P7) emphasize the challenge of maintaining trust in the early stages of interaction [Luger and Sellen, 2016]: how to encourage users with low initial trust to continue interacting with the system until the AI obtains more and richer data to make better recommendations? However, our interviews do not address fine-grained time dynamics: the delay to deliver AI recommendations. In contrast, recommendation delay is investigated in the academic literature [Buçinca et al., 2021; Kraus et al., 2020; Park et al., 2019]. It can reduce users’ over-reliance [Buçinca et al., 2021; Park et al., 2019], yet it might have no effect on self-reported trust [Buçinca et al., 2021]. Therefore, these findings underscore the importance of:

- 1) **implications for academic researchers**: investigating what contributes to **users’ willingness to engage with an AI-embedded system when their initial trust levels are low**;
- 2) **implications for AI practitioners**: focusing on more **granular effects of time on trust such as delays in AI recommendations**.

### 3.5.3 Type of Task

*Interview findings* The nature of the task influences Human-AI trust according to the interviewees (DS4, DS5, P6) and can be organized along two dimensions. First, the magnitude of computational needs (DS5, P6) concerns the amount of data and computational resources needed to achieve the task. Second, the degree of subjective evaluations (DS4) that may not be easily objectified. For example, users are

more likely to trust AI than humans for tasks requiring the processing of large amounts of data “because it is very difficult for a human to perform calculations” (DS5). In contrast, systems that perform subjective evaluation of humans (e.g. matchmaking) inspire less trust (DS4).

*Literature comparison* The academic literature also considers “subjective evaluations” as dimension [Kolasinska et al., 2019; Langer et al., 2022] and provides a more granular analysis of the dimension “magnitude of computational needs”. Indeed, the literature considers “task complexity” [Robinson, 2001; Sasayama, 2016] that encompasses different types of processing demands: computational, attention, memory, reasoning, etc. (e.g. [Fan et al., 2008; Müller et al., 2020; Sutherland et al., 2015]). Moreover, the literature introduces a third dimension, which is the “the degree of responsibility” [Buçinca et al., 2020; Eiband et al., 2019; Gamkrelidze et al., 2021; Kolasinska et al., 2019; Luger and Sellen, 2016; Müller et al., 2020]. Generally, tasks that are socially sensitive (e.g. sending an email) evoke little trust in AI [Eiband et al., 2019; Luger and Sellen, 2016]. Similarly, for life-and-death decisions, people are more likely to trust human advice [Eiband et al., 2019] or a combination of human and AI recommendations [Kolasinska et al., 2019] than AI advice alone. These examples involve high “vulnerability”. While the interviewees mentioned vulnerability as a key element for trust to exist (see Section 3.4.1), they do not consider it as a factor, related to the task, that can increase or decrease Human-AI trust. From these observations, we propose the following:

- 1) **implications for academic researchers**: further identifying the **important axes of decision-making task categorization** are related to Human-AI trust;
- 2) **implications for AI practitioners**: considering the **magnitude of task complexity in a broader meaning as well as degree of responsibility**.

#### 3.5.4 Marketing

*Interview findings* The last factor related to the socio-technological context and discussed by our interviewees (P4, DS1 and DS7) is marketing: “the deciding factor [for trust in AI] is marketing and the way the AI system is presented to users, e.g. in the media” (DS1). DS1 also believes that people’s trust can be influenced by the term “AI” itself: “I think [people] trust [the AI-embedded system], as long as they do not know that AI is in play.” P4 describes the fact that trust can be built on a series of parameters that are not technical but social and related to market-

ing: “there may also be [...] marketing elements, which can be very insidious, which can generate the relationship of trust, which is not at all obvious.”

*Literature comparison* The academic literature and interviews agree that the way an AI-embedded system is presented can affect trust between humans and AI. Then the literature provides ampler details regarding the type of information in marketing that can influence people’s trust in AI. This information includes system’s accuracy metric [Fan et al., 2008; Rechkemmer and Yin, 2022; Yin et al., 2019], the characterization of a system as (in)competent [Xiao et al., 2007], and statements of commitment to fairness and the elimination of racial bias, gender discrimination, and all other forms of discrimination, even if they do not necessarily reflect reality [Lee and Rich, 2021]. Only one aspect was mentioned by one interviewee (DS1): the chosen terminology (e.g. “algorithm”, “robot”, “sophisticated statistical model”, “artificial intelligence”) which affects trust [Langer et al., 2022]. The lack of details provided by the interviewees could be because they have no marketing experience, and as a consequence, can only speak about marketing in general terms. Therefore, these findings have the following **implications for AI practitioners**:

- 1) considering closely the **terminology they choose to describe their system**;
- 2) account for the **amount and type of details communicated about AI** (metrics, instilled ethical values, etc.).

### 3.5.5 Summary

The main Human-AI trust factors related to socio-technological context highlighted by AI practitioners and decision subjects generally coincide with the ones stemming from the academic literature: Human-Human trust, time dynamics, type of task, and marketing. However, each party has different focuses with regard to each trust factor (see Table 3.6). For instance, the interviewees emphasise more the importance of Human-Human trust for Human-AI trust, especially trust of users in the team behind AI development. The interviewees also discuss a challenge related to trust changing over time, not highlighted in the literature: in early stages of deployment, if users’ trust in AI is low, it is difficult to prevent them from abandoning the system and convince them that they will have more reasons to trust AI later on, as the recommendations improve. In return, the academic literature provides a more detailed analysis of the way AI is presented to users and how the type of decision-making task can affect Human-AI trust.

	AI Practitioners	Decision Subjects	Academic Literature	
<b>Systems' Development and Design</b>	<b>Performance and Errors</b>	P2	DS2, DS4, DS6, DS7	17; 63; 94; 95; 108; 115; 127; 130; 137; 143; 161; 183; 186; 213; 214; 220; 261; 271; 273; 292; 351; 370; 371; 384; 389; 392; 395-397; <b>no effect</b> : 10; 45; 109; 302; 321; 350
	<i>Context of errors</i>	P1	DS4	–
	<i>Frequency of errors</i>	P2	DS4	–
	<i>Nature of errors</i>	P1	–	<b>no effect</b> 109
	<i>Relativity of performance: system</i>	P2	DS2, DS4	–
	<i>Relativity of performance: humans</i>	P7	DS2, DS4, DS5	54; 361; 402; 404, <b>relative error tolerance</b> : 278; 279; 298; 312; 352; 382; 384
	<i>Robustness</i>	–	–	196
	<i>Usability</i>	–	–	108
	<i>Errors x design</i>	–	–	400
	<i>Errors x interactivity</i>	–	–	129
	<i>Errors x expectations</i>	–	–	54
	<i>Fairness and Biases</i>	–	DS3, DS6	173; 340; 382
	<b>Transparency</b>			
	<i>Working process</i>	P4, P7	–	343
	<i>Data</i>	P7	–	18; 84; 123; 196; 272
	<i>Explanations</i>	P1, P2, P3, P6; P4, P5, P7 disagree	DS3, DS4, DS7 disagree	49; 97; 99; 117; 145; 164; 186; 205; 321; 372; 373; 394; <b>no effect</b> : 268; 271; 329; 402
	<i>Type of explanations</i>	–	–	345; 372; 373; 387; <b>no effect</b> : 270; 299
	<i>AI confidence score</i>	–	–	30; 76; 139; 163; 288; 292; 323; 371; 389; 402; <b>no effect</b> : 370; 371
	<i>Social transparency</i>	–	–	87
	<b>Interactivity</b>	P1, P2, P3, P4	DS2, DS3, DS4, DS6	45; 53; 129; 266; 316
	<b>AI Certification</b>	P1, P4, P6 (P5 disagrees)	DS1-DS3, DS5, DS7	156; 196
<b>Appearance</b>	–	–	2; 108; 127; 161; 183; 203; 253; 253; 274; 288; 332; 353; 354; 400	
<b>Communication Style</b>	–	–	17; 45; 71; 195; 253; 261; 287; 288; 354; 368; 380; 384; 387; 404	
<b>Privacy</b>	–	–	196	
<b>Behavior and Personality</b>	–	–	45; 353; 384	
<b>Proximity</b>	–	–	253	

**Table 3.8:** Summary of the Human-AI trust factors related to systems' design and development discussed by AI practitioners and decision subjects and studied in academic papers on Human-AI trust in the context of decision making.

### 3.6 Trust Factors Related to the Systems' Development and Design

Another theme generated from the thematic analysis concerns a group of factors related to properties of a system, the interviewees discussed

four of them: Performance, Transparency, Interactivity and AI certification. Table 3.8 provides a summary by reporting these factors, who talk about them, and the academic literature that addresses them.

### 3.6.1 Performance

*Interview findings* AI performance in terms of accuracy is the one of the most discussed factors by the interviewees (P<sub>1</sub>, P<sub>2</sub>, P<sub>7</sub>, DS<sub>2</sub>, DS<sub>4</sub>, DS<sub>6</sub>, DS<sub>7</sub>). They highlight three interesting nuances about it. First, the interviewees distinguish **absolute vs. relative performance**, where the relative performance of a system, as opposed to absolute performance, is the observed performance relative to that of another system or human being considered as baselines. *“Absolute performance doesn’t help much”* (P<sub>7</sub>), the question is whether the system is better than other system (P<sub>2</sub>, DS<sub>2</sub>, DS<sub>4</sub>) or a human (P<sub>7</sub>, DS<sub>2</sub>, DS<sub>4</sub>, DS<sub>5</sub>). For instance, *“I would trust an algorithm that counts molecules for cancer more than a human because I think it can do that task better than the human”* (DS<sub>4</sub>). Second, the interviewees highlight the role of the **context**, i.e. when and where recommendation errors occur. For instance, according to P<sub>1</sub>, users *“do not forgive a slightest error [of AI]”* once AI is deployed, while errors are often tolerated in testing phase (P<sub>1</sub>). Similarly, errors may not be tolerated in some environments, e.g. *“[AI error] in a medical context is bad”* (DS<sub>4</sub>) while they can be tolerated in other environments, e.g. dating decisions (DS<sub>4</sub>). Third, interviewees discuss the **characteristics** of recommendation errors. Interviewees may be more with human-like errors: referring to certain mistakes made by a system P<sub>1</sub> stated *“I do not believe that these are mistakes, I would have answered the same thing.”* (P<sub>1</sub>). In addition to the nature of the error, frequency is important: *“if everyone trusts the system and something weird happens, we’ll say, well, that’s okay. I guess if it happens too often, you start asking questions”* (P<sub>2</sub>). Finally, DS<sub>6</sub> and DS<sub>3</sub> mention another aspect of performance beyond accuracy - **fairness**, e.g. *“when everyone has an equal chance of being selected, regardless of age, gender and experience”* (DS<sub>3</sub>).

*Literature comparison* Both interviewees and academic literature (e.g. [Chien et al., 2018; Fahim et al., 2021; Tolmeijer et al., 2021; Yin et al., 2019], for more see Table 3.8) largely discuss AI accuracy as a factor that influences Human-AI trust. They also agree that fairness and biases affect trust [Kasinidou et al., 2021; Stapleton et al., 2022; Woodruff et al., 2018]. While the interviews highlight the importance of relative performance of an algorithm, few studies in the literature focus on the role of its relativity [Cai et al., 2019; Veale et al., 2018; Zhang et al., 2020; Zheng et al., 2022]. In this case, studies compare to which extent AI errors are tolerated compared to human errors (which is related

to the human-like nature of the error) [Pearson et al., 2016; Perelman et al., 2020; Richter et al., 2019; Salomons et al., 2018; Tolmeijer et al., 2022; Woodruff et al., 2018; Xie et al., 2019] rather than investigating whether taking an AI or a human provides a comparative advantage in certain decisions.

The interviewees highlight people's perceptions of errors depending on the the context and the characteristics (nature and frequency) of the errors. In contrast, the literature focuses more on how system's features such as design aesthetics [Yuksel et al., 2017], interactivity [Gupta et al., 2022] or expectations (mental model of the AI's design objective) [Cai et al., 2019] change the effect of AI errors on trust. Interviewees mentioned the link between interactivity (section 3.6.3) and expectation (section 3.7.2) with human-AI trust, but without reporting on the interaction with performance. Therefore, we propose the following:

- 1) **implications for academic researchers** : evaluating the role of **context, frequency and nature of errors** when studying AI performance and Human-AI trust;
- 2) **implications for academic researchers** : shifting the focus from absolute accuracy to **relative accuracy** when studying AI performance and Human-AI trust;
- 3) **implications for AI practitioners** : taking into account the role of **design aesthetics, interactivity of AI recommendations, mental models of the AI's role** when discussing AI performance and Human-AI trust.

### 3.6.2 Transparency

*Interview findings* AI transparency, alongside with AI performance, is another much discussed and important trust factor (P1-P7, DS3, DS4, DS7). The interviewees define it as understanding the working process of AI development team and as understanding why a specific AI recommendations was shown and its quality. AI practitioners spontaneously report on how to make transparency actionable for the users focusing on explaining the **working process**, e.g. *“when we [...] try to be as transparent as possible on how [the AI-embedded system] works, we try to explain it to [clients], because it can be sometimes quite technical, even mathematical, and then there are no more problems, no problem of trust...”* (P4), explaining the **data**, e.g. *“You have to be very, very transparent about how you prepared the data, because any AI is biased just by the quality of the data (and also the quantity).”* (P7), and explaining the **recommendation**, e.g. *“[ex-*

*plainability is] how we prove that our results are reliable [and] accurate” (P3).*

Interestingly, there are divergent opinions about the relevance of explainability in algorithmic recommendations (Explainable AI, XAI). On the one hand, some AI practitioners think it helps users calibrate their trust in AI recommendations by better understanding how they were derived and by estimating their quality (P1, P2, P3, P6). On the other hand, some practitioners and decision subjects question the role explainability plays for Human-AI trust. For instance, *“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)*. P1 adds *“It is a bad idea to put all the tools we’ve developed in the field of explainable AI directly into the hands of users without them knowing anything”*. P1 believes that AI explanations are more important to the trust that an AI development team will have in its models than to the trust of users in the AI. The two provided reasons are the **complexity** of the explanations (P1, P7, DS3, DS4), e.g. *“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)* and **time pressure** (P7, DS7), e.g. *“If users 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)*. P3 adds that while explainable AI is important for their domain (legal decision making), it is not necessarily the case for all the domains.

*Literature comparison* Transparency is a key factor for Human-AI trust according to both the interviews and the academic literature. It is thus not surprising to observe similarities between these sources of information regarding: its **definition** [Larsson and Heintz, 2020; Lepri et al., 2018]; the **elements to explain** which are the working process [Sultanum et al., 2018], the **data** [Anik and Bunt, 2021; Drozdal et al., 2020; Glass et al., 2008; Lee and Rich, 2021; Park et al., 2021] and the **recommendations** (e.g. [Bućinca et al., 2020; Faulhaber et al., 2021; Schaffer et al., 2019; Wang and Yin, 2021], for more see Table 3.8); the **mitigated role** of transparency due to system complexity [Wang and Yin, 2022], difficulty to make explanations actionable [Glass et al., 2008; Sultanum et al., 2018], and context [Xin et al., 2021].

However, the interviews and the academic literature do not put the emphasis on the same aspects. First, the interviewees focus more than the literature on the working process as an element to explain (rather than data and recommendations). In contrast, the academic literature includes AI confidence score associated with a recommendation



as part of transparency [Bansal et al., 2021; Desai et al., 2013; Helldin et al., 2013; Jiang et al., 2021; Pynadath et al., 2018; Rechkemmer and Yin, 2022; Schneider et al., 2019; Wang et al., 2016; Yang et al., 2017], absent from the interviewees' discussions. Moreover, the academic literature introduced the term "social transparency" [Ehsan et al., 2021]. Rather than explaining the inner workings of an AI-embedded system to build user trust, the authors advocate providing more information about how other users have historically incorporated AI recommendations into their decision making.

Finally, the interviewees highlight the different roles of transparency for AI practitioners and decision subjects. While AI practitioners tend to raise the importance of explainability (both for themselves and other stakeholders), decision subjects do not see how transparency can affect their trust in AI since explanations might be difficult to understand and the additional information is usually not actionable. Meanwhile, the academic literature focuses only on users with only one study suggests that transparency has effect on trust of decision subjects [Park et al., 2021]. Based on these findings, we therefore propose the following:

- 1) **implications for academic researchers** : evaluating the role of **time pressure** regarding the impact of transparency on Human-AI trust;
- 2) **implications for AI practitioners** : taking into account the role of **AI recommendation's accuracy and type of explanations** regarding the impact of transparency on Human-AI trust;
- 3) **implications both for academic researchers and AI practitioners** : exploring the role of **social transparency of AI** for Human-AI trust;
- 4) **implications both for academic researchers and AI practitioners** : further investigating to which extent **AI transparency contributes to decision subjects' trust in AI**.

### 3.6.3 Interactivity

*Interview findings* Interactivity is another factor impacting Human-AI trust in the context of decision making (P1-P4, DS2, DS3, DS4, DS6). The interaction with AI is often perceived as limited to "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). However, interactivity "allows users to question [AI recommendations]" (P2), and that is the "way [...] to gain trust" (P2). P4 echoes this reflection by saying that "trust [in

AI] is not established because the button is red or green”, but because there is “a dialogue” (P4). Interestingly, interactivity is mentioned through different terms and expressions related to human-human interaction such as “a dialogue” (P4), “cooperation” (P1), “ask for more explanations” (P3), “negotiate” (DS4). This comparison with human-like interaction is sometimes more explicit. For instance, DS3 would prefer a decision made directly by a human, because “a human being is flexible” (DS3) and could re-examine their case and give them a second chance: “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). This mechanism seems to be currently missing in their interactions with AI-based decision making process. They report that AI recommendations lack flexibility and room for negotiation. They feel excluded from the decision loop, and they see interactivity as a solution to this problem. Decision subjects’ impressions are to be a “part of the statistics” (DS2) or simply “filtered out” by AI (DS3).

*Literature comparison* The academic literature on Human-AI trust examining the interactivity of AI recommendations remains scarce [Bridgewater et al., 2020; Cai et al., 2019; Gupta et al., 2022; Okolo et al., 2021; Saxena et al., 2021]. The limited evidence they provide supports the interviewees’ claims: interactivity affects Human-AI trust, because it contributes to the refinement of the mental model about AI [Cai et al., 2019], gives a sense of striving to improve decision making [Okolo et al., 2021], allows to explore to which extent nuances are accounted for AI recommendations [Saxena et al., 2021]. As in the discussions of AI practitioners, these studies are primarily about users rather than decision subjects. However, it seems that interactivity for users and decision subjects serves different purposes. For the former, interactivity means data exploration and mental model refinement. For the latter, it is an empowerment over Human-AI decisions so as not to feel solely “part of the statistics” (DS2). Therefore, these findings both for academic researchers and AI practitioners highlight the importance of:

- 1) further investigating the role of **interactivity of AI recommendations, including decision subjects** in the scope for Human-AI trust.

### 3.6.4 AI Certification

*Interview findings* 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."* However, respondents do not agree on whether certification is sufficient. P4 says that *"the certification alone should be enough [for Human-AI trust] if it is done well."*, while others highlight the importance of the institution behind the certification (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). This also holds for the process of certification (DS1 and DS5): it should be done, for example, *"based on research studies"* (DS1). P5 and DS7 are more suspicious about certification because 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).

*Literature comparison* AI certification has not been largely studied as a factor for Human-AI trust. Two qualitative studies, from medical decision domain, report that direct users [Jacobs et al., 2021] and decision subjects [Lee and Rich, 2021] would trust an AI-embedded system more if it went through a certification. However, knowing that an AI-embedded system is certified could potentially lead to overtrust: if there is AI certification, doctors would more systematically follow AI recommendations rather than evaluating them each time they are faced with them [Jacobs et al., 2021]. Given the paucity of empirical evidence on how AI certification affects Human-AI trust and the importance it takes in the industry, these findings provide the following implication for academic researchers :

- 1) further investigating to which extent **AI certification contributes to Human-AI trust** in the context of decision making.

### 3.6.5 Summary

The interviews and the academic literature bring light onto four Human-AI trust factors related to the system's development and design: AI's performance, transparency of AI, degree of the system's interactivity, and AI certification. AI's performance and transparency are the factors that both the interviewees and the literature discuss the most, but each party highlights different aspects of the factors. Regarding AI performance, interviewees place less emphasis on the existence of AI errors than on the context in which they occur and their nature. With regard to AI transparency, the literature highlights more contextual aspects that vary the degree of impact of transparency on human-AI trust than the interviewees. Lastly, the interviewees focus more on the

role of AI certification for Human-AI trust. This could be explained by the fact that certification plays a more important role in industry than in academia, as it offers a competitive advantage in the market, especially for critical systems.

	AI Practitioners	Decision Subjects	Academic Literature	
People's Preferences and Experiences	<b>Agency</b>			
	<i>Direct users</i>	P4	–	10; 123; 172; 181; 230; 290; 346; 359; 383; 385; no effect: 50; 104; 321; 352
	<i>Decision subjects</i>	P7	DS1, DS4, DS5, DS6	–
	<i>Agency awareness</i>	–	DS7	–
	<b>Element of Unexpected and Expected</b>			P6
	<i>Good surprise</i>	P1, P2	DS1, DS2, DS4, DS7	–
	<i>Bad surprise</i>	P1, P2	–	–
	<i>Prior experiences</i>	–	DS3, DS6	46; 196
	<b>AI Literacy</b>			
	<i>Direct users</i>	P5, P7	–	54; 65; 117; 172; 295; no effect: 351; 387
	<i>Decision subjects</i>	–	DS2 (DS4, DS6, DS7 –disagree)	–
	<b>Domain Expertise</b>			
	<i>Actual expertise</i>	P6	–	95; 99; 115; 117; 163; 316; 321; 359; 402; 404; no effect: 387
	<i>Self-confidence</i>	–	–	163; 203; 321; no effect: 290
	<b>Individual differences</b>			
	<i>Age</i>	–	–	107; 107; 271; 351
	<i>Education</i>	–	–	372; 373; no effect: 351
	<i>Personality</i>	–	–	149
	<i>Propensity to trust</i>	–	–	97; 149; no effect: 351; 387
	<i>Culture</i>	–	–	63; 179; 368; no effect: 351
	<i>Emotional state</i>	–	–	94
	<i>Gender</i>	–	–	no effect: 271; 351
	<i>Work style</i>	–	–	117; 149; 358
	<i>Attitudes towards robots</i>	–	–	351
	<b>Operator Workload</b>			–
				278; inconclusive effects: 63; 130

**Table 3.9:** Summary of the Human-AI trust factors related to people's preferences and experiences discussed by AI practitioners and decision subjects and studied in academic papers on Human-AI trust in the context of decision making.

### 3.7 Trust Factor Related to People's Preferences and Experiences

In this section we report findings on trust factors related to people's preferences and experiences, which are: Agency, Expectations, AI literacy and Domain expertise. They are summarized in Table 3.9.

### 3.7.1 Agency

*Interview findings* Decision subjects mostly led the discussions about agency and Human-AI trust (P4, P7, DS1, DS4-DS7). They highlighted the lack of flexibility of these systems which undermines trust. This lack of flexibility is illustrated with 1) the limited number of options regarding the recommendations (DS7, P7): *"We need to build trust by making [decision subjects] understand that we do not always claim to be right [...] we're going to introduce four levels [of AI recommendations] instead of the binary [...]"* (P7) and 2) the rigidity of the dialogue (DS5, DS7, P7) because humans have to adapt to the system's constraints: *"this is just defined so that a machine can evaluate it"* (DS6) while the system does not adapt to the humans: *"these [AI] systems should be more flexible for human error. Right now, it's so strict."*

Sometimes the system adapts to the users' behavior, but still some agency problems can happen: *"we never know how the system will evolve over time, and this raises a lot of questions"* (P4). These changes may occur without people being aware of it, which undermines their sense of agency. For instance, the task management application on DS7' phone *"recognized some of my behaviors and made some decisions for me based on that."* The app would turn off notifications or mobile data without DS7's approval, and because of this, they missed an important phone call.

*Literature comparison* Among the interviewees, it was predominantly decision subjects (5 out of 7 respondents) who highlighted the relationship between agency and trust in AI, while the literature focuses exclusively on the agency of direct users [Alan et al., 2014; Buçinca et al., 2021; Fogliato et al., 2021; Glass et al., 2008; Kapania et al., 2022; Kraus et al., 2020; Maurer et al., 2018; Rajaonah et al., 2006; Schaffer et al., 2019; Sutherland et al., 2015; Tolmeijer et al., 2022; van Maanen et al., 2011; Xiao et al., 2007; Xin et al., 2021]. 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. Furthermore, most of the interviewees see the agency over AI as binary, full control or none (apart from P7), while the literature oversees three levels of agency: full - AI recommendations are optional and appear on demand [Buçinca et al., 2021; Kraus et al., 2020,?; Schaffer et al., 2019; Sutherland et al., 2015; van Maanen et al., 2011], limited - mandatory AI recommendations that appear immediately [Buçinca et al., 2021; Fogliato et al., 2021; Kraus

et al., 2020; Rajaonah et al., 2006; Schaffer et al., 2019; Sutherland et al., 2015; Tolmeijer et al., 2022; van Maanen et al., 2011] or only after users' initial decision [Buçinca et al., 2021; Fogliato et al., 2021], and none- AI recommendation executed autonomously [Kraus et al., 2020; Maurer et al., 2018; Rajaonah et al., 2006; Tolmeijer et al., 2022; van Maanen et al., 2011]. Lastly, the literature mentions that choosing what AI-embedded system to use instead of being assigned one affects users' trust [Xiao et al., 2007], but do not address how showing multiple AI recommendations versus one can affect Human-AI trust.

Considering these elements, we propose the following **implications for academic researchers** :

- 1) further investigating what contributes to **the sense of agency of decision subjects**;
- 2) understanding how people's trust in AI (both users' and decision subjects') changes as a result of **varying levels of control over AI recommendations**;
- 3) studying how **binary AI recommendations versus multiple ones** affects Human-AI trust.

### 3.7.2 Expectations about AI Recommendations

*Interview findings* The extent to which an AI recommendation meets one's expectation (or not) affects Human-AI trust, especially if one finds it surprising, which means a strong disconfirmation of expectation (P1, P2, P6, DS1, DS2, DS3, DS6, DS7), moderated by past experiences (DS3, DS6). The interviewees distinguish between a positive and a negative surprising recommendations. A "good surprise, it is AI that teaches us [humans] things we did not know" (P1). In contrast, bad surprise is when an AI recommendation does not meet our expectations and is mostly likely wrong. The interviewees do not agree about the influence of good/bad surprise on Human-AI trust. While P1 believes that a good surprise can have a positive impact on Human-AI trust, P2 thinks that any kind of surprising recommendation - good or bad - undermines trust, but with a different magnitude: "a good surprise is always perceived positively, it damages trust [in AI] a bit less [than a bad surprise]."

*Literature comparison* Expectation and Surprise as factors affecting Human-AI trust are largely discussed by the interviewees, but not largely studied in the academic literature. While Kawakami et al. [2022] shows

that unexpected recommendations generally decrease users' trust in AI, surprise had an inconclusive effect on users' trust [Tokushige et al., 2017]. However, these two articles do not distinguish positive surprise from negative surprise as described by the interviewees. Lastly, two studies [Brown et al., 2019; Lee and Rich, 2021] also show that past experiences affect expectations, supporting the claims of DS<sub>3</sub> and DS<sub>6</sub>. These findings have the following **implications for academic researchers** :

- 1) further investigating the effect of **good and bad surprises in AI recommendations** on Human-AI trust;

### 3.7.3 AI Literacy

*Interview findings* AI literacy appears an important factor for the interviewees for trust in AI. They distinguish public education and specific training. First, P<sub>5</sub> 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" (P<sub>5</sub>). Similarly, P<sub>7</sub> believes that users should understand what AI can do and cannot do; DS<sub>2</sub> believes that "educating people about the difference between AI programs and a simple algorithm" would be beneficial. However, other decision subjects do not share this sentiment (DS<sub>4</sub>, DS<sub>6</sub>, DS<sub>7</sub>). For instance, "educational events [about AI] do not really make sense to me, because often nobody knows how the system really works" (DS<sub>4</sub>) or "the educational sessions [about AI] do not make sense to me, how can they help?.." (DS<sub>6</sub>).

Some interviewees propose **training specific** to certain needs. For instance, when it comes to AI literacy and patients (decision subjects), P<sub>7</sub> is ready to "create materials, [...] flyers, [...] content, for patients so that they are informed, that they are not afraid of this new technology" (P<sub>7</sub>). They also regularly recall basic concepts of deep learning models to the doctors they work with (P<sub>7</sub>). Rather than considering the whole AI system, P<sub>1</sub> focuses more on how to help their clients to understand AI recommendations and the risks related to explainability (section 3.6.2): "We start by introducing what the methods [of explainability] do, and also introduce what they don't do... I often start with a little trap, i.e. I point out the little mistakes one can fall into. When I have done that, I accompany [clients] in making an interpretation [of the explanations] together" (P<sub>1</sub>).

*Literature comparison* The academic literature does not widely investigate how AI literacy affects Human-AI trust and their focus is largely on direct users [Cai et al., 2019; Chromik et al., 2020; Ghai et al., 2021; Kapania et al., 2022; Reig et al., 2018; Tolmeijer et al., 2021; Yang et al.,

2020]. Generally, the findings are in line with the reflections of our AI practitioners: AI literacy can be a tool to decrease undertrust caused by a lack of knowledge about the data or inner processes [Cai et al., 2019; Reig et al., 2018] or made-up misleading folk theories about AI [Reig et al., 2018]. AI literacy can also lead to decrease overtrust in AI recommendations in case of blind trust [Chromik et al., 2020; Ghai et al., 2021; Kapania et al., 2022]. However, these studies rely on participants' self-reported familiarity with AI, rather than introducing educational sessions about AI for their participants. Therefore, the question of whether specific educational events have an impact on human-AI trust and how to design them to meet the needs and roles of different stakeholders (users, decision subjects, investors, etc.) remains open. For example, decision subjects stated that they would like the knowledge received in the educational sessions about AI to be actionable. Therefore, these findings inspire the following:

- 1) **implications for academic researchers** : investigating the effect of **educational sessions about AI** on Human-AI trust in the context of decision making;
- 2) **implications both for academic researchers and AI practitioners** : designing **educational materials about AI, considering the needs of various stakeholders**.

#### 3.7.4 Domain Expertise

*Interview findings* Besides AI literacy, domain expertise is another factor, which relates to a person's knowledge about the task they are performing, e.g. medical decision making. P6 is the only interviewee to express that task expertise can influence trust in AI: “[users] do not have the knowledge to challenge our recommendations and analyses” (P6).

*Literature comparison* Academic literature confirms what P6 says: users with no or limited knowledge about the task at hand generally exhibit higher rates of overtrust than those who are experts in the domain [Fan et al., 2008; Feng and Boyd-Graber, 2019; Gamkrelidze et al., 2021; Ghai et al., 2021; Jiang et al., 2021; Saxena et al., 2021; Schaffer et al., 2019; van Maanen et al., 2011; Zhang et al., 2020; Zheng et al., 2022]. However, the empirical studies offer another angle on domain expertise: Both objective and subjective domain expertise have an effect on trust in AI. Indeed, even if users are not domain experts, high self-confidence, or simply considering themselves good at completing the task, might impact their trust in AI recommendations [Jiang et al., 2021; Li et al., 2007; Schaffer et al., 2019]. AI practitioners do not seem



to largely account for this individual difference in their practice. Therefore, these findings highlight for **AI practitioners** the importance of:

- 1) accounting for **domain expertise (objective & subjective)** for Human-AI trust in the decision-making context.

### 3.7.5 Summary

The AI practitioners and decision subjects talk about 4 major factors related to people's preferences and experiences - degree of agency, expectations regarding AI recommendations, AI literacy, and domain knowledge (see Table 3.9). The reviewed academic literature also investigates these factors, but the studies usually focus more on users' trust rather than decision subjects' one. This gap is particularly striking for the sense of agency as this factor is predominantly discussed by decision subjects. Another difference is that the interviewees put more importance on the unmet expectations regarding AI recommendations, which can result in a good or bad surprise. The literature at the same time focuses more on domain knowledge. Furthermore, the interviewees and the literature could address the need for designing educational materials about AI, considering the needs of the various stakeholders.

## 3.8 General Discussion

In the context of AI-assisted decision making, trust is often studied through users' perspectives, with controlled lab experiments involving AI mock-ups [Glikson and Woolley, 2020; Vereschak et al., 2021]. We argue that studying Human-AI trust through the lens of stakeholders other than users is crucial for advancing our understanding of existing practices with AI-assisted decision making systems. In this chapter, we reported on the interview findings about the definition and factors affecting Human-AI trust in the context of decision making through the lens of 7 AI practitioners and 7 decisions subjects. Moreover, the comparison of the interviewees' reflections with the academic literature on trust in AI-assisted decision making allowed us to identify implications for academic research and AI practitioners. In light of the reported findings, this section discusses our three research questions and proposes some directions for future work.

### 3.8.1 *Discussing Our Research Questions*

#### **1) How do AI practitioners and decision subjects define Human-AI trust in decision making in comparison to the academic literature?**

We find that the interviewees and the academic literature share similar views on the Human-AI trust definition. It was unexpected, because trust is a complex and abstract theoretical concept [Lewicki and Brinsfield, 2011; Lyon et al., 2015], which leads to frequent theoretical confusions [Jacovi et al., 2021; Liao and Sundar, 2022; Vereschak et al., 2021]. Interestingly, between the two groups of interviewees, we also do not find differences in defining key elements of trust. It remains that the interviewees provided a more nuanced outlook on the key elements of trust in comparison with the academic literature [Jacovi et al., 2021; Vereschak et al., 2021]. First, their discussions highlight that vulnerability and positive expectations cannot be boiled down to monetary losses and high levels of accuracy as they are presented in the empirical studies [Vereschak et al., 2021]. The interviewees point out a new possible key element of trust, task complexity, which could have experimental and policy implications. All in all, these findings show that discussing theoretical concepts with laypeople of different backgrounds not only can validate the academic theories, but also potentially contribute to theoretical advancement. Similar approach has been used to better define contestability of AI recommendations [Lyons et al., 2021], AI fairness [Saxena et al., 2019], and responsible AI in general [Jakesch et al., 2022], and we contribute to this line of literature with the findings about Human-AI trust.

#### **2) What do AI practitioners and decision subjects think affects Human-AI trust in the context of decision making?**

AI practitioners and decision subjects mostly name the same factors that they believe can affect Human-AI trust. However, they prioritize them differently. For example, AI practitioners put a lot of importance on AI explanations as a way to affect Human-AI trust. At the same, decision subjects who discussed AI explanations and AI literacy mostly disagree with AI practitioners with regards to the importance of these factors, claiming that the information provided is usually of no use to them as they cannot act upon it. Additionally, both AI practitioners and decision subjects highlight the importance of AI interactivity for trust, but AI practitioners see it as a means to refine mental model of AI and decision subjects see it as a way to get involved in the decision making loop. Lastly, decision subjects discuss the sense of agency and its relationship to Human-AI trust more than AI practitioners, which is expected considering the above 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. These findings align with the prior work [Jakesch et al., 2022] 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 [Yurrita et al., 2022] for Human-AI trust.

**3) What are the differences and similarities between the interviewees and the literature in the factors they propose?** There is much discussion in the interviewees and the academic literature about the role of AI performance and transparency for human-AI trust. The interviewees, put more importance on contextual (e.g. frequency of errors) and social (e.g. relativity of AI performance in comparison with humans) nuances of AI performance, while the literature views it in absolute terms (the higher the accuracy, the better) and focuses on the interaction between performance and AI design. Considering that the interviewees also provide more discussions about Human-Human trust and surprising AI recommendations, we can see that the factors of trust they prioritize the most are related to human interactions and perceptions rather than purely technological and design ones (e.g., absolute performance, explanations). These observations are in-line with the recent, under-explored for AI-assisted decision making concept of social transparency [Ehsan et al., 2021]. Through highlighting the history of other users' interactions with AI recommendations rather than the inner workings of AI, it embraces the interviewees' emphasis on trust factors related to social interactions, information actionability, and expectations as a part of the system's design. Our findings can motivate further investigation regarding the incorporation of these socio-oriented trust factors into system's design.

### 3.8.2 *Future Work Directions*

In this chapter, we interviewed representatives from a large panel of decision domains (e.g. medicine, finance, recruitment). Considering that type of task and level of risk 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. Moreover, the literature studies two groups of factors, interpersonal differences (e.g. age, gender, culture) and systems' embodiment (e.g. appearance, communication style), which have not been mentioned in the interviews. It is possible that these factors would have been discussed by the interviewees in another task domain. Understanding the

differences and similarities between various task domains can inform researchers and policy makers on higher level classification of domains [Lai et al., 2021, preprint].

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 [Ayling and Chapman, 2022] 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 investigate how they define Human-AI trust and to what extent they consider it in their working practices.

Finally, the scope of our study is Human-AI trust in the context of decision making. A promising direction is to investigate whether our approach and findings can be generalized to other trust contexts such as Human-Automation trust. This will be an important step to understand what sets Human-AI trust apart.

#### TAKE AWAY MESSAGES

##### *Contributions:*

- A comprehensive overview of Human-AI trust factors in the context of decision making;
- An advanced understanding of Human-AI trust and its factors completed with the views of AI practitioners and decision subjects;
- A set of research and design opportunities aimed to account for the particularities of the socio-technical context the AI-embedded systems are deployed at and the needs of the stakeholders other than users.

# 4

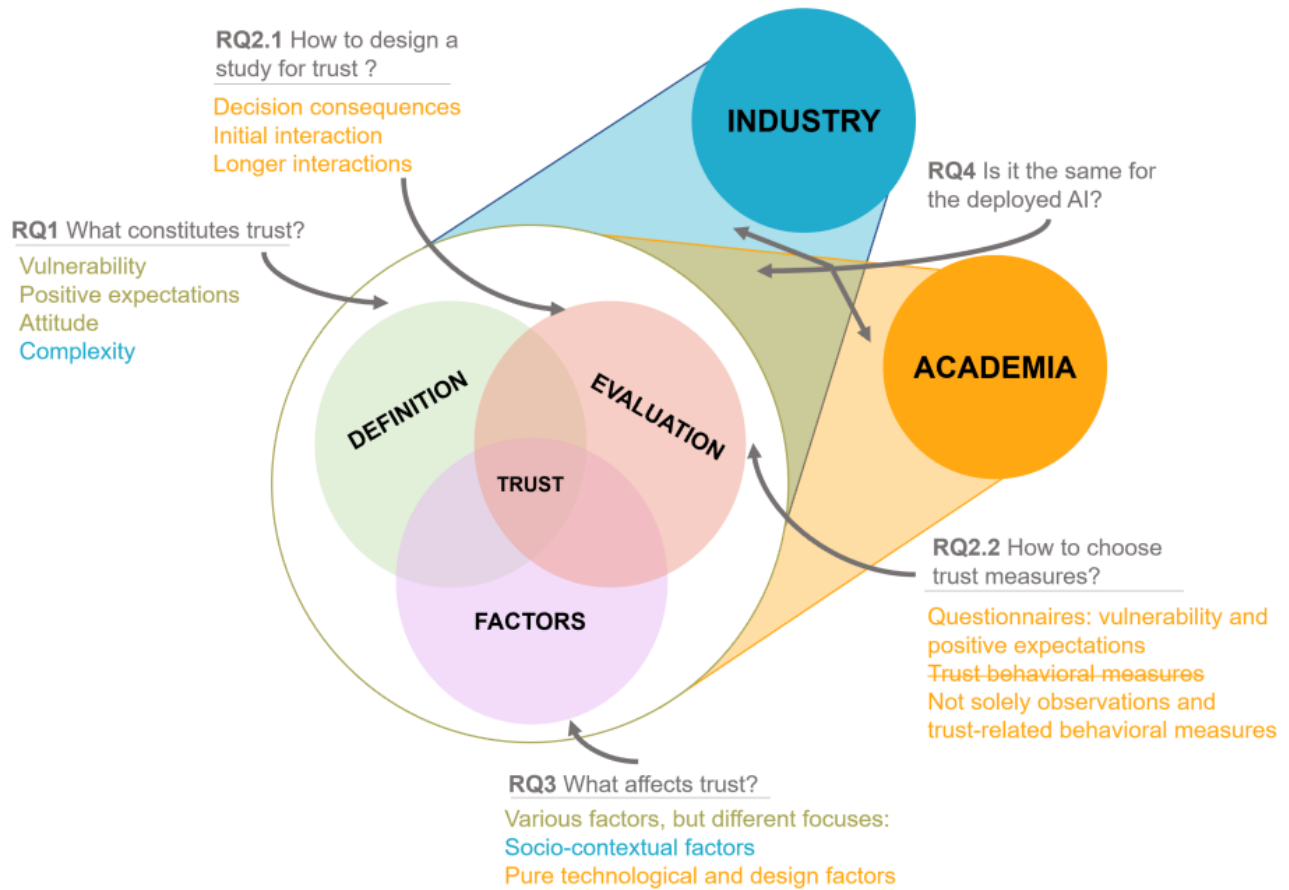
## *Discussion, Future Perspectives, and Conclusions*

### 4.1 Progress on Research Problems

This thesis was originally motivated by the question “*What is Human-AI trust in the context of decision making?*” I further split it up in four research questions related to Human-AI trust definition (RQ1), evaluation (RQ2), factors (RQ3), and their comparison between the academia and industry (RQ4). In the subsequent subsections, I summarize and discuss the findings regarding these questions.

#### *4.1.1 RQ1 What differentiates Human-AI trust from other related constructs, such as reliance, compliance, trustworthiness, etc.?*

I addressed this question by conducting a systematic literature review of the studies on Human-AI trust in the context of decision making and analyzing the trust definitions they proposed (*Chapter 2*). I found three common elements among these definitions: vulnerability, positive expectations, and attitude. Building on social and cognitive sciences literature, I reinforced for the Human-AI interaction research community that these key elements of trust are what differentiates trust from other concepts such as confidence, distrust, behaviors like reliance and compliance, and trustworthiness. While this distinction remains true for human trust in the entities other than AI, this was the first time it was articulated for the Human-AI trust community on the



**Figure 4.1:** Principal findings related to each research question. Note that RQ4 does not have findings next to it as they are already incorporated in the answers to the other RQs. Hence, the findings are color-coded: from academia, industry, and both.

example of decision making. Highlighting the key elements of trust could also serve as a guidance for researchers to what look out for in a trust definition while selecting one among the numerous existing ones. Moreover, the key elements of trust have implications on trust evaluation, which I will discuss in the next subsection. However, vulnerability, positive expectations, and attitude might not be the only key elements of trust that contribute to its formation and differentiate it from other related concepts, which I will address in Section 4.1.4.

#### 4.1.2 RQ2 How to evaluate trust in the context of decision making?

To conduct an evaluation, one needs an appropriate protocol to follow (RQ 2.1) and adequate measures (RQ 2.2). Thus, the answer to this question contains two parts. To tackle it, during the systematic literature review mentioned above, I annotated and analyzed the empirical protocols of the studies and the methods they used to study Human-AI trust (Chapter 2). Building on the comparison between the trends

in the analyzed corpus and the existing evaluation trends in the social and cognitive sciences literature, I proposed 14 guidelines (8 for the empirical protocol and 6 for the trust measures) aimed to support a more standardized approach to evaluating trust. Notably, almost half of these guidelines are related to the key elements of trust, which highlights that a clear theoretical definition of a concept is beneficial for its evaluation. For example, vulnerability implies that decisions that participants make based on AI recommendations should have real or virtual consequences and feel realistic. Positive expectation means that in the first trials AI recommendations should be correct. Attitude signifies that trust cannot be evaluated solely through observations and behaviors as it is mostly questionnaires and interviews that can measure attitudes. Hence, the term “behavioral trust measures” should be substituted with “trust-related behavioral measures.” Lastly, trust is not static, it evolves over time, thus to study its dynamics one has to ensure that the interaction between participants and AI has been long enough to capture various stages of trust development.

The proposed guidelines are a step towards standardization of the studies’ protocols and measures that can achieve three goals: 1) ensuring that the data collected is about Human-AI trust rather than trust-related concepts such as confidence, distrust, behaviors, and trustworthiness; 2) facilitating the studies replication; 3) supporting the cross-study comparison of the findings. While most of the guidelines on assessment of AI and its trustworthiness focus on examining technical properties of AI and performance metrics [National Commission on Informatics and Liberty, 2022], I put AI evaluation with humans at the center of my guidelines.

However, the effect of some experimental protocol decisions on Human-AI trust and data collection remains unclear. For example, it is not known whether virtual decision consequences are as effective for trust formation as the real ones. Another example would be uncertainty about to which extent 1-item questionnaires capture trust in comparison with other measures. I addressed these and other aspects of the empirical protocol and trust measure decisions in 9 research opportunities in Chapter 2.

#### 4.1.3 *RQ3 What factors affect Human-AI trust in the context of decision making?*

To address this question, I expanded the previous systematic literature review, annotated, and summarized the factors the studies on Human-AI trust in the context of decision making. Following the structure of



some existing trust frameworks, I grouped them in 3 categories: factors related to the socio-technical context, to system's development and design, and to people's preferences and experiences. In total, I found 47 trust factors in the literature, and the most studied ones in each category respectively are type of task (the degree of subjectivity, complexity, and responsibility or risk), AI performance (mostly viewed through absolute AI accuracy and existence of errors) and transparency (mostly represented with AI explanations and confidence score), and individual differences (e.g., age, gender, etc.). This is the first overview of Human-AI trust factors in the general context of decision making.

However, I have not explored which of these trust factors are unique to the decision making context nor how they would vary depending on the decision-making domain (e.g. medical, financial, juridical). Additionally, I have only considered ACM Digital Library, and the review scope can be enlarged to IEEE Xplore Library and Scopus. Pursuing this further exploration can enhance the research community's understanding of what trust aspects are particular to Human-AI trust and to the decision-making context.

#### 4.1.4 *RQ4 Do the academic postulations about trust definition, factors, and evaluation of Human-AI trust reflect the real world considerations?*

To address this question, I conducted semi-structured interviews with two groups of people - AI practitioners and decision subjects. I considered specifically these two populations, because besides users, the most common focus of the empirical studies on Human-AI trust, they are the most tightly linked to Human-AI decision making. I asked them about how they define Human-AI trust and what factors they think would affect people's trust in AI. Then I compared their reflections to the findings from the systematic literature reviews conducted for RQ1-3. From the differences in the interview and review findings, I provided two sets of implications: research opportunities for the academia and design implications for the industry.

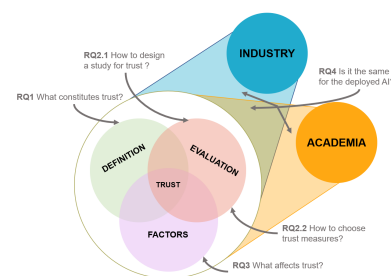
For the trust definition, I found that despite trust being a complex and abstract theoretical concept, the interviewees define it with the same key elements as the academia - vulnerability, positive expectations, and attitude. They also proposed a new trust element not considered in the academia as such - task complexity. Since key elements of trust have direct impact on the choices around experimental protocol and trust measures, it is beneficial for the academia to further investigate the role of complexity for trust formation.

For the trust factors, the interviewees brought up 24 different factors that can affect Human-AI trust. For the socio-technological context, the most discussed factor is Human-Human trust, while type of task is not largely regarded. For system's development and design, the interviewees almost equally discussed AI performance, transparency, interactivity, and AI certification. However, contrary to the academia, AI performance is not talked in absolute terms and always compared to another person or system. Another difference that the role of AI explanations is more disputed by decision subjects in particular due to the lack of interactivity and actionability. For the people's preferences and experiences, (un)expectedness of AI recommendations and sense of agency were discussed the most, rather than individual differences. Therefore, the interviewees put importance on the trust factors that are related to human interaction, perception and context, while the academia focuses more on the factors related to technology and design. Thus, for AI-embedded systems to be a supportive tool for decision makers, we should aim to promote the socio-contextual factors in AI design and development. I note, however, that I recruited the interviewees from different decision-making backgrounds, and it is possible that the importance of a group of factors might vary depending on the application domain.

#### 4.1.5 Pieces of Trust Puzzle Brought Together

The last point I would like to address is the meaning behind the representation of my thesis project as Venn diagram. In fact, each research question is a standalone piece, yet helped me to inform the results of another one, which is represented by the circles' intersections:


- Understanding what constitutes trust and differentiates it from the related theoretical concepts contributed to guidelines for designing more standardized empirical protocols and navigating the choice of trust measures;
- Standardization of protocols and understanding what to look out for in trust measures contributes to better understanding what factors affect Human-AI trust through ensuring that the collected data is related to trust and facilitating the cross-study comparison of the findings;
- A complete overview of what factors can affect Human-AI trust in the decision making context and to which extent contributes to our understanding of Human-AI trust, and specifically how it is different from Human-AI trust in other contexts and human trust in



**Figure 4.2:** Reminder of the graphical representation of the problematics investigated in the thesis.

other entities. Additionally, the intersection between definition and factors underlines potentially double nature of vulnerability (risk) and task complexity as key elements for trust formation and also factors that affect the levels of trust later on;

- Lastly, comparing the findings about Human-AI trust through the lenses of academia and industry allows for the validation of the theoretical concepts as well as evaluation of the prominence of the literature findings for the systems deployed in the market.

Therefore, tackling the initial question “*What is Human-AI trust in the context of decision making?*” through the reflections of academia and industry about trust definition, evaluation, and factors is like bringing the complementary pieces of trust puzzle together .

## 4.2 Scientific Contributions

In summary, the contributions of this thesis can be grouped in the following manner: *methodological, survey, theoretical, and empirical* as per categorization by Wobbrock and Kientz [2016].

- Methodological<sup>1</sup>
  - *Guidelines for Empirical Protocols and Trust Measures*. I provided 8 guidelines to standardize the design of empirical protocol for studies on Human-AI trust in the context of decision making, emphasising the importance of including vulnerability, positive expectations, and longer interactions. I have also provided 6 guidelines to facilitate the choice and use of trust measures, emphasising that trust is an attitude and, hence, cannot solely be derived from the behaviors.
- Survey<sup>2</sup>
  - *Landscape of Current Trends in Protocols and Measures for Human-AI Trust Evaluation*. As an outcome of the systematic literature review, I have summarized and categorized the existing protocol choices per each standard section of an empirical protocol in the current studies on Human-AI trust in the context of decision making. I did the same for the existing qualitative and quantitative trust measures. This way the community has a compact overview of all the possible ways to design their studies and a

<sup>1</sup> “Methodological research contributions create new knowledge that informs how we carry out our work” [Wobbrock and Kientz, 2016].

<sup>2</sup> “Survey research contributions [...] review and synthesize work done on a research topic with the goal of exposing trends and gaps” [Wobbrock and Kientz, 2016].

straightforward access to the repertoire of trust measures used in the research community.

- *Landscape of Factors that Affect Human-AI Trust in the Context of Decision Making.* As an outcome of the systematic literature review, I have provided a structured overview of all the trust factors considered in the studies on Human-AI trust in the context of decision making, organizing them in three groups.
- Theoretical<sup>3</sup>
  - *Difference between Trust and Related Concepts.* I have highlighted the key elements that differentiate trust from confidence, distrust, reliance, compliance, and trustworthiness: vulnerability, attitude, and positive expectations. Additionally, I also proposed to further investigate the role of task complexity.
  - *Research and Design Implications around Human-AI Trust Factors.* I have identified the research and design opportunities for academic researchers and AI practitioners. For academic researchers, they are mostly related to the investigation of socio-contextual trust factors, e.g. trust between AI users and AI team, relative performance, social transparency, and surprising AI recommendations. For AI practitioners, they are related to the practitioners' design, development, and deployment practices, e.g. the way they communicate about their AI, types of AI explanations, and individual differences of users.
- Empirical<sup>4</sup>
  - *Human-AI Trust Definition and Factors as Seen in the Industry.* As an outcome of the semi-structured interviews, I have described how AI practitioners and decision subjects define Human-AI trust with their own words. I also discovered what they think can affect Human-AI trust in the context of decision making and which of these factors they consider the most important.

<sup>3</sup> “Theoretical research contributions consist of new or improved concepts, definitions, models, principles, or frameworks” [Wobbrock and Kientz, 2016].

<sup>4</sup> “Empirical research contributions [...] provide new knowledge through findings based on observation and data-gathering” [Wobbrock and Kientz, 2016].

### 4.3 Further Perspectives

In this thesis, I set to understand what Human-AI trust in the context of decision making is through the lenses of academia and industry. Building on the comparison between these two perspectives, I present

the Short-, Med-, and Long-Term Perspectives for this work.

#### 4.3.1 *Short-term*

I identify two projects that are an immediate continuation of this thesis:

1) studying the effect of surprising AI recommendations on Human-AI trust, 2) *real-world* evaluation of Human-AI trust.

##### 1) Human-AI Trust and Surprising AI Recommendations

Among all the trust factors discussed from the perspectives of academia and industry, unexpected, or surprising, AI recommendations have one of the biggest gaps in terms of the attention paid by the academia (6 studies out of 113, predominantly qualitative) and the interviewees (half of the respondents). In general terms, surprise is a conscious feeling, that is triggered after experiencing *unexpected* events, that is the ones that disconfirm one's expectations [Reisenzein, 2000]. Besides being linked to expectations similarly to trust, surprise can also offer another perspective on why absolute AI accuracy might not always matter as per my call to the academia to shift towards relative and contextual AI performance.

To explore the effect of surprising AI recommendations on Human-AI trust in the context of decision making, we<sup>5</sup> designed an online experiment. Participants had to assume the role of a real estate agent who revises their rentals portfolio and needs to estimate the monthly prices of their rentals based on 8 criteria. They were assisted by AI recommendations, derived by a linear regression model we trained on a database of real rentals. However, investigating surprise in the experimental settings and its effect on trust imposes two challenges: theoretical and methodological.

The first challenge is linked to controlling surprise. As surprise is linked to the degree of unexpectedness (difference between initial expectations and actual experience) [Reisenzein and Studtmann, 2007; Teigen and Keren, 2003], we need to control participants' expectations. While it could be done with a training before the main experiment, it is not clear whether it is the expectations about AI recommendations, the task itself or both that have to be controlled. Additionally, the literature suggests that there are other prerequisites for surprise such as the ability to provide explanations about the unexpected event [Foster and Keane, 2013; Maguire et al., 2011], novelty of the event, and the confidence about an alternative outcome [Reisenzein, 2000].

<sup>5</sup>This study was designed and conducted in a collaborative effort, and thus, any use of "we" here refers to: Katerina Batziakoudi (intern at our research group), Oleksandra Vereschak, Gilles Bailly, Baptiste Caramiaux.

Surprise evaluation is the second challenge, as there exist 4 categories of surprise measures [Reisenzein, 2000]: subjective surprise [Foster and Keane, 2013; Meyer et al., 1991; Reisenzein, 2000; Teigen and Keren, 2003]; expressions [Reisenzein, 2000; Reisenzein and Studtmann, 2007]; behavioral: reaction Time (delay between the event and the user's reaction) [Foster and Keane, 2013; Meyer et al., 1991; Niepel et al., 1994; Reisenzein, 2000], error (increased error rate) [Reisenzein and Studtmann, 2007], attention shift [Horstmann and Herwig, 2015; Itti and Baldi, 2009; Meyer et al., 1991; Niepel et al., 1994]; and cognitive: expectations [Itti and Baldi, 2009; Kahneman et al., 1982; Maguire et al., 2011; Teigen and Keren, 2003], confidence [Kahneman et al., 1982; Reisenzein, 2000] and explanations [Foster and Keane, 2013; Maguire et al., 2011]. It remains unclear which of these measures is the most appropriate for measuring surprise.

Therefore, before investigating the effect of surprising AI recommendations on Human-AI trust, one has to understand what causes surprise and how to evaluate it. We ran the online study with 70 participants, and our preliminary findings and limitations are the following:

- The greater the disconfirmation of expectations, i.e. the greater the difference between the participants' initial price estimation and the AI recommendation, the greater the surprise. Other variables such as ground truth, confidence in the answer, and the timing of the surprise have no significant correlation with surprise. However, this could be attributed to the nature of the experiment and the task. For example, in online experiment, we cannot determine whether a long reaction time is due to participants' reflection or distraction. Additionally, participants reported their confidence did not fluctuate much throughout the entirety of the task. Lastly, our experimental set-up did not account for the novelty effect nor for the ability to explain an unexpected event. We argue that a more controlled experiment with the physical presence of participants is needed with tasks of varying difficulty and novelty, pre-determined in a pilot study. Think aloud protocol might be more appropriate to understand to which extent participants are able to find an explanation in the event of surprise.
- The greater the distance between the AI prediction and the real price (AI error), the lower the trust in AI, but the error alone cannot account for the changes on trust levels. Another plausible explanation is high levels of surprise, but its effect might depend at what moment of the experiment the surprise occurs. Further analysis is required to see if this relationship holds true across varying levels

of expertise of the participants. Additionally our set-up does not account for the distinction between bad and good surprises as defined by AI practitioners and decision subjects on Chapter 3. Lastly, we collected data about trust with less frequent periodicity than surprise to make them more engaged with decision making and not to make the experiment feel like an elaborated questionnaire. Consequently, it is difficult to assess the precise changes of trust following right after high levels of surprise are reported. Think aloud protocol might also be helpful for collecting the data about good and bad surprises during the experiment as well as noting immediate trust changes after a surprising AI recommendation.

## 2) Real-world Evaluation of Human-AI Trust

In this thesis, I used both the findings from the academia and reflections from the industry to investigate Human-AI trust definition and factors, which allowed me to have a global overview about Human-AI trust as well as understand to which extent the academic findings are reflected in the real world. One aspect of Human-AI trust I have looked so far only through the academic perspective is Human-AI trust evaluation, and I believe it would be insightful to also understand the industry's approach to it. If the societal goal is to build AI-embedded systems one can trust, it is necessary to know how to evaluate people's trust in AI to understand to which extent the goal was achieved, and I would like to know whether the industry possesses the appropriate protocols and tools to do so.

Specifically, I propose to investigate the following questions through revisiting the interviews conducted with AI practitioners:

1. *What role does the evaluation of Human-AI trust play in the industry?*  
The Human-AI interaction community has experienced a rise in interest in understanding Human-AI trust and the policy makers put it as one of the design and development priorities, but it is unclear whether the industry matches them in their efforts to incorporate Human-AI trust in their practices.
2. *How do AI practitioners evaluate Human-AI trust?* As seen in Chapter 2, academia proposes a large spectrum of empirical protocols' design and trust measures, and choosing one among them is not trivial. It is not known how AI practitioners approach Human-AI trust evaluation, whether consult academic sources or develop their own procedures and measures.

3. *What are the barriers for Human-AI trust evaluation?* For those AI practitioners who do not evaluate Human-AI trust, I would like to explore the reasons behind this to see what kind of support and incentives they require. For those AI practitioners who evaluate Human-AI trust, this would be an opportunity to understand their challenges as well as to validate and complete the evaluation guidelines proposed in Chapter 2 in the real-world scenarios. It is possible the proposed guidelines are not as actionable for the people who know little about trust and evaluations with users.

So far I have mainly envisioned the research questions targeted at only AI practitioners, because they would be in charge of conducting AI evaluations with users. It would also be beneficial to interview decision subjects about trust evaluation, because their feedback can be useful for validating certain trust measures, e.g. ensuring that wording of the trust questionnaires are understandable. Decision subjects could also explain whether and how they would like to participate in Human-AI trust evaluation as they do not directly interact with AI. Thus, the guidelines, which for now are mostly focused on trust evaluation with users, can be extended to include decision subjects in the evaluation process, too.

#### 4.3.2 *Mid- and Long-term*

- *Expanding the stakeholders ecosystem.* In this thesis, I focused on three stakeholders, the most directly linked to Human-AI interaction: users (the main focus of the academic articles) and AI practitioners and decision subjects (the focus of the conducted interviews). As mentioned in Chapter 3, it would be interesting to investigate understanding of Human-AI trust as seen by other stakeholders. While there is no established categorization of stakeholders involved in governance, development, and use of AI-embedded systems, some researchers suggest up to 11 groups of people whose decisions can affect how Human-AI interaction will go down [Ayling and Chapman, 2022]. For instance, while I considered AI practitioners as a whole, some categorizations differentiate AI managers, AI developers, and AI UX designers [Ayling and Chapman, 2022; Yurrita et al., 2022]. As their business goals differ and might not be aligned, they might have different approaches to defining and establishing Human-AI trust. Another example could be policy makers who provide a global strategy and directives on how to regulate AI-embedded systems. The way they understand Human-AI trust might impact the legislation and national policies.



- *Contextualizing Human-AI trust factors.* In this thesis, I provided a global overview of Human-AI trust factors studied in the academia and considered in the industry in the context of decision making. As briefly suggested in Chapter 3, it could be interesting to conduct a cross-domain comparison to see to which extent Human-AI trust factors play importance for various types of decision making. Besides the application domain, such comparison can be expanded to different type of AI, task complexity, and the magnitude of stakes. Additionally, while I talked about the importance of Human-AI trust factors, I have not discussed how exactly they affect trust: whether they increase, decrease or have no effect on Human-AI trust. I have also mostly treated the factors separately, not exploring their interaction effect on trust. A more granular understanding of Human-AI trust factors in the context of decision making would also enable their comparison with more general trust factors, e.g. in Human-Automation interaction or Human-Human interaction, to identify the similarities and differences between these types of trust.
- *Hierarchy of AI recommendations.* When we receive a piece of advice from a human, we might view it differently depending on who it comes from [Tajfel and Turner, 2004]. For example, if it comes from someone superior like a boss, we might be more inclined to trust it because of their experience or even authority. On the other hand, if we receive a suggestion from a new intern, we might be more inclined to dismiss it due to their inexperience in comparison with us. Similarly, people might assign the same hierarchical structures and rule to the AI-embedded systems. Consequently, the perceived hierarchical role of AI can affect users expectations about its capabilities and role of their AI recommendations [Cai et al., 2019]. For example, the medical personnel who viewed AI more of an assistant or an intern did not expect it to be accurate, but rather to provide an additional opinion on the issue and, hence, was tolerant to AI errors [Cai et al., 2019]. I believe further investigating the perceived hierarchy of AI recommendations could shed more light on the mechanisms of algorithm aversion, overtrust in AI, and differences in Human-AI trust repair.

#### 4.4 Conclusion

In summary, this thesis contributes to understanding of Human-AI trust in the context of decision making with a broad family of AI-

embedded systems that make predictions, can be opaque and adaptable. I explore Human-AI trust definitions, evaluation, factors through the lenses of academia, that is the empirical studies, and industry, that is AI practitioners and decision subjects. I illustrate what differentiates trust from other related concepts, such as confidence, distrust, reliance, compliance, and trustworthiness on an example of decision making. This understanding helped me to derive the guidelines aiming to standardize the empirical protocols and trust measures used in the studies on Human-AI trust in the context of decision making. I also provide an overview of trust factors related to Human-AI trust while making decisions, and demonstrate that while academia and industry share numerous factors in common, they focus on different aspects. Investigating Human-AI trust through two lenses also informed a series of research opportunities and design implications on how to further support Human-AI trust in the context of decision making. This work also calls for a deeper understanding of Human-AI trust factors and their interaction with the context and other factors. In addition, it argues for integrating the views on Human-AI trust of other stakeholders linked to regulation, development and use of AI-embedded systems assisting decision making.

Human-AI trust is a recent research domain, and the work of this thesis provides an overview on the current state of art in terms of definitions, evaluation, and factors in the context of decision making. It provides several pointers on how the research community can continue building on the existing knowledge in a way that allows for easier cross-study comparison and that is reflected in the real world needs and challenges.



# A

## *Trust Definitions*

1. *“an attitude that an agent will achieve an individual’s goals in a situation characterized by uncertainty and vulnerability”* by **Lee and See** [Lee and See, 2004] and **Lee and Moray** [Lee and Moray, 1992] (n=9, 11.25%);
2. *“the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party”* by **Mayer** [Mayer et al., 1995] (n=5, 6.25%);
3. *“evolving affective state including both cognitive and affective elements and emerges from the perceptions of competence and a positive, caring motivation in the relationship partner to be trusted”* by **Ekman** [Ekman et al., 2016], which is stated to be a combination of the definitions by Lee and See [Lee and See, 2004] and Mayer et al. [Mayer et al., 1995] (n=1, 1.25%);
4. *“the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid”* by **Madsen** [Madsen and Gregor, 2000] (adapted from McAllister [McAllister, 1995]) (n=2, 2.5%);
5. *“an evolving, affective state including both cognitive and affective elements and emerges from the perceptions of competence and a positive, caring motivation in the relationship partner to be trusted”* by **Young and Albaum** [Yuksel et al., 2017] (n=1, 1.25%);

6. *“a psycho-physiological state that involves a firm belief about another’s intention and one’s willingness to act by following their words, expressions, decisions, or actions”* by **Bonn and Holmes** [Boon and Holmes, 1991] (n=1, 1.25%);
7. *“a psychological state, resulting from knowledge, beliefs, and assessments related to the decision-making situation, which creates confident expectations for human-machine system performance and guides operator reliance on automation”* by **Rajaonah et al.** [Rajaonah et al., 2006] (n=1, 1.25%);
8. proposed **their own definition**: *“a relationship between two entities(trustor: users and trustee: AI technologies) guided by compound cognitive processes (mental deliberation, reasoning and mental processing involving memory, learning and accumulated knowledge) during the evaluation of the trustworthiness of a trustee(AI technology) by a trustor (user) based on the accumulation of the following: trustor’s (user’s) intentions, beliefs, and anticipated behaviors”* inspired by [Braynov, 2013], [Cho et al., 2015] (n=1, 1.25%);
9. *“confidence in a robot’s decision-making capabilities and therefore the likelihood to follow those decisions”* with a **flawed source** stated (n=1, 1.25%);
10. *“a latent (hidden, unobservable) variable that summarizes (mental model) past experience with an agent/robot, which is useful for predicting future behavior of the trustee and making a decision to put oneself in a position of vulnerability”* with **no source** stated (n=1, 1.25%);
11. *“how confident an individual is in the abilities of the other members of the group”* with **no source** stated (n=1, 1.25%).

# *B*

## *Selected Human-Human Trust Questionnaires*

See the next page.

## B.1 Behavioral Trust Inventory [Gillespie, 2003]

Note: Items 1–5 tap reliance-based trust and items 6–10 tap disclosure-based trust.

Please indicate how willing you are to engage in each of the following behaviors with *your Leader/Team Member/Follower*, by circling a number from 1 to 7.

	<i>Not at all willing</i>			<i>Completely willing</i>			
1. Rely on your leader's task related skills and abilities.	1	2	3	4	5	6	7
2. Depend on your leader to handle an important issue on your behalf.	1	2	3	4	5	6	7
3. Rely on your leader to represent your work accurately to others.	1	2	3	4	5	6	7
4. Depend on your leader to back you up in difficult situations.	1	2	3	4	5	6	7
5. Rely on your leader's work-related judgments.	1	2	3	4	5	6	7
6. Share your personal feelings with your leader.	1	2	3	4	5	6	7
7. Discuss work-related problems or difficulties with your leader that could potentially be used to disadvantage you.	1	2	3	4	5	6	7
8. Confide in your leader about personal issues that are affecting your work.	1	2	3	4	5	6	7
9. Discuss how you honestly feel about your work, even negative feelings and frustration.	1	2	3	4	5	6	7
10. Share your personal beliefs with your leader.	1	2	3	4	5	6	7

## B.2 Trust Questionnaire [Currall and Judge, 1995]

168

CURRALL AND JUDGE

that Task Coordination is a central feature of all BRP relationships, involves use of all four dimensions of trust including Task Coordination. This option would be attractive to those who wish to study whether the overall level of BRP trust impacts interorganizational collaboration. To develop Task Coordination items for a particular organizational context, researchers should use the procedure we outlined in the Method section. To develop Task Coordination items, preliminary interviews must include open-ended questions asking how BRPs engage in trusting behavior ("reliance") in the context of Task Coordination. Prior to use in a final survey, these initial items should be refined by subjecting them to Ghiselli *et al.*'s (1981) process analysis.

### APPENDIX: TRUST ITEMS

Instructions for the items read: "Answer the questions in terms of what you would actually do in dealing with the (counterpart BRP) . . ." The response format was: 1 = *extremely unlikely*, 2 = *quite unlikely*, 3 = *slightly unlikely*, 4 = *neither*, 5 = *slightly likely*, 6 = *quite likely*, and 7 = *extremely likely*. Item numbers correspond to their order in the surveys. Asterisks indicate reversed items.

#### Communication Dimension Items

1. Think carefully before telling the (counterpart BRP) my opinions.\*

7. Give the (counterpart BRP) all known and relevant information about important issues even if there is a possibility that it might jeopardize the (respondent's organization).

8. Give the (counterpart BRP) all known and relevant information about important issues even if there is a possibility that it might jeopardize my job as the (respondent's job).

12. Minimize the information I give to the (counter part BRP).\*

18. Deliberately withhold some information when communicating with the (counterpart BRP).\*

#### Informal Agreement Dimension Items

3. Enter into an agreement with the (counterpart BRP) even if his/her future obligations concerning the agreement are not explicitly stated.

5. Enter into an agreement with the (counterpart BRP) even if I think other people might try to persuade him/her to break it.

10. Enter into an agreement with the (counterpart BRP) even if it is unclear whether he/she would suffer any negative consequences for breaking it.

17. Decline the (counterpart BRP's) offer to enter into an unwritten agreement.\*

20. Suggest that the (counterpart BRP) and I enter into an unwritten agreement.

#### Surveillance Dimension Items

2. Watch the (counterpart BRP) attentively in order to make sure he/she doesn't do something detrimental to the (respondent's organization).\*

6. Keep surveillance over the (counterpart BRP) (i.e., "look over his/her shoulder") after asking him/her to do something.\*

9. Feel confident after asking the (counterpart BRP) to do something.

14. Check with other people about the activities of the (counterpart BRP) to make sure he/she is not trying to "get away" with something.\*

15. In situations other than contract negotiations, check records to verify facts stated by the (counterpart BRP).\*

#### Task Coordination Dimension Items for Superintendents

4. Ask the president to convince the membership of the local teacher's union to give support to a newly initiated cooperative program between teachers and school administrators.

11. Ask the president to convince several incompetent teachers to take early retirement.

13. Ask the president to stop false rumors about personnel decisions that are circulating among the teachers.

16. Ask the president to convince the teachers to file grievances only in extreme cases.

19. Rely on the president to convince the membership of the teachers' local to have realistic expectations about what contract changes will be made in the next negotiation.

#### Task Coordination Dimension Items for Presidents

4. Ask the superintendent to try to persuade the district's administrators to lend their support to a newly initiated cooperative program between teachers and administrators.

11. Rely on the superintendent to make decisions about teacher transfers and assignments with a genuine concern for teacher job preferences.

13. Rely on the superintendent to dismiss teachers only in cases when poor performance has been clearly and impartially demonstrated.

16. Rely on the superintendent to solve a grievance through informal and cooperative discussions.

19. Rely on the superintendent to adhere to the collective bargaining contract.



### B.3 Trust for Management Questionnaire [Mayer, 1999]

#### Measures of Trust, Trustworthiness, and Performance Appraisal Perceptions

The following instructions prefaced the scales. The anchors shown below were consistent throughout. Headings of construct names are for clarity of exposition, and were not included in the surveys.

Indicate the degree to which you agree with each statement by using the following scale:

1	2	3	4	5
Disagree strongly	Disagree	Neither agree nor disagree	Agree	Agree strongly

Think about [company name]'s top management team [names listed in parentheses for clarity]. For each statement, write the number that best describes how much you agree or disagree with each statement.

#### Ability

Top management is very capable of performing its job.  
 Top management is known to be successful at the things it tries to do.  
 Top management has much knowledge about the work that needs done.  
 I feel very confident about top management's skills.  
 Top management has specialized capabilities that can increase our performance.  
 Top management is well qualified.

#### Benevolence

Top management is very concerned about my welfare.  
 My needs and desires are very important to top management.  
 Top management would not knowingly do anything to hurt me.  
 Top management really looks out for what is important to me.  
 Top management will go out of its way to help me.

#### Integrity

Top management has a strong sense of justice.  
 I never have to wonder whether top management will stick to its word.  
 Top management tries hard to be fair in dealings with others.  
 Top management's actions and behaviors are not very consistent.\*  
 I like top management's values.  
 Sound principles seem to guide top management's behavior.

#### Propensity

One should be very cautious with strangers.  
 Most experts tell the truth about the limits of their knowledge.  
 Most people can be counted on to do what they say they will do.  
 These days, you must be alert or someone is likely to take advantage of you.  
 Most salespeople are honest in describing their products.  
 Most repair people will not overcharge people who are ignorant of their specialty.  
 Most people answer public opinion polls honestly.  
 Most adults are competent at their jobs.

#### Trust

If I had my way, I wouldn't let top management have any influence over issues that are important to me.\*  
 I would be willing to let top management have complete control over my future in this company.  
 I really wish I had a good way to keep an eye on top management.\*  
 I would be comfortable giving top management a task or problem which was critical to me, even if I could not monitor their actions.

Think about the performance review system at [company name], and answer the following questions.

#### Accuracy

The evaluation of what skills I have is pretty accurate.  
 How much work I get done is important to my performance review.  
 How many mistakes I make in my work is important to my performance review.  
 Whether or not my supervisor likes me is important to my performance review.\*  
 How much effort I put into my job is important to my performance review.  
 How many "extra" things I do is important to my performance review.  
 Finding ways for the company to save money is important to my performance review.  
 Coming up with good ideas for the company improves my performance review.

#### Outcome instrumentality

Whether or not I get a raise depends on my performance.  
 If you are one of the better performers in this company, you will get one of the better raises.  
 If I perform well, my chances of moving up are improved.  
 \*-Reverse-scored item.

Received February 27, 1997

Revision received June 15, 1998

Accepted June 16, 1998 ■

# C

## *Trust Questionnaires Used in Human-AI Literature*

### C.1 Human Trust in Automation Scale [Jian et al., 2000]

Instructions: Below is a list of statements for evaluating trust between people and automation. There are several scales for you to rate intensity of your feeling of trust, or your impression of the system while operating a machine. Please select the option which best describes your feeling or your impression using the 7-point scale ranging from 1 (not at all) to 7 (extremely).

- The system is deceptive. (R)<sup>1</sup>
- The system behaves in an underhanded (concealed) manner. (R)
- I am suspicious of the system's intent, action, or outputs. (R)
- I am wary of the system. (R)
- The system's actions will have a harmful or injurious outcome. (R)
- I am confident in the system.
- The system provides security.
- The system has integrity.
- The system is dependable.
- The system is reliable.
- I can trust the system.
- I am familiar with the system.

<sup>1</sup> The R represents reverse coded items for scoring.

## C.2 Human-Robot Trust Questionnaire [Schaefer, 2013]

What % of the time will this robot...	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Act consistently*											
Protect people											
Act as part of the team											
Function successfully*											
Malfunction (R)											
Clearly communicate											
Require frequent maintenance (R)											
Openly communicate											
Have errors * (R)											
Perform a task better than a novice human user											
Know the difference between friend and foe											
Provide Feedback*											
Possess adequate decision-making capability											
Warn people of potential risks in the environment											
Meet the needs of the mission*											
Provide appropriate information*											
Communicate with people*											
Work best with a team											
Keep classified information secure											
Perform exactly as instructed*											
Make sensible decisions											
Work in close proximity with people											
Tell the truth											
Perform many functions at one time											
Follow directions*											
Be considered part of the team											
Be responsible											
Be supportive											
Be incompetent (R)											
Be dependable *											
Be friendly											
Be reliable *											
Be pleasant											
Be unresponsive* (R)											
Be autonomous											
Be predictable *											
Be conscious											
Be lifelike											
Be a good teammate											
Be led astray by unexpected changes in the environment											

\* marks the questions that can be used for a shorter version of the questionnaire. The R represents reverse coded items for scoring.

### C.3 Trust in Management Questionnaire [Mayer, 1999]

See Appendix B.3

### C.4 Trust in Automation [Muir, 1989]

Please select a value from 1 to 10, where 1 = Not at all and 10 = Completely.

- To what extent can the system's behavior be predicted from moment to moment?
- To what extent can you count on the system to do its job?
- What degree of faith do you have that the system will be able to cope with all systems "states in the future"?
- Overall how much do you trust the system?

### C.5 Trust in Teammate [Ross, 2008]

1. To what extent does Teammate A perform this search-and-rescue task effectively?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

5. To what extent can you anticipate Teammate A's behavior with some degree of confidence?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

3. To what extent is the Teammate A free of errors?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

4. To what extent do you have a strong belief and trust in Teammate A to do the search-and-rescue task in the future without being monitored?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

5. How much did you trust the decisions of Teammate A overall?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

6. What percentage of responses by Teammate A do you think were correct?

\_\_\_\_\_ (enter a value between 0% to 100%)

7. How often did you notice an error made by Teammate A?

Not At All 1 2 3 4 5 6 7 8 9 Many Times

8. To what extent did you lose trust in Teammate A when you noticed it made an error?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

Questions 9 and 10 of this survey seem not to be included.

## C.6 Human-Computer Trust Scale (HCT) [Madsen and Gregor, 2000]

### 1. Perceived Reliability

- R1) The system always provides the advice I require to make my decision.
- R2) The system performs reliably.
- R3) The system responds the same way under the same conditions at different times.
- R4) I can rely on the system to function properly.
- R5) The system analyzes problems consistently.

### 2. Perceived Technical Competence

- T1) The system uses appropriate methods to reach decisions.
- T2) The system has sound knowledge about this type of problem built into it.
- T3) The advice the system produces is as good as that which a highly competent person could produce.
- T4) The system correctly uses the information I enter.
- T5) The system makes use of all the knowledge and information available to it to produce its solution to the problem.

### 3. Perceived Understandability

- U1) I know what will happen the next time I use the system because I understand how it behaves.
- U2) I understand how the system will assist me with decisions I have to make.
- U3) Although I may not know exactly how the system works, I know how to use it to make decisions about the problem.
- U4) It is easy to follow what the system does.
- U5) I recognize what I should do to get the advice I need from the system the next time I use it.

### 4. Faith

- F1) I believe advice from the system even when I don't know for certain that it is correct.
- F2) When I am uncertain about a decision I believe the system rather than myself.

- F3) If I am not sure about a decision, I have faith that the system will provide the best solution.
- F4) When the system gives unusual advice I am confident that the advice is correct.
- F5) Even if I have no reason to expect the system will be able to solve a difficult problem, I still feel certain that it will.

#### 5. Personal Attachment

- P1) I would feel a sense of loss if the system was unavailable and I could no longer use it.
- P2) I feel a sense of attachment to using the system.
- P3) I find the system suitable to my style of decision making.
- P4) I like using the system for decision making.
- P5) I have a personal preference for making decisions with the system.

### C.7 Trust in Automation [Chien et al., 2018]

Encompassing the cultural aspects.

Dimension	Survey Items	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
<i>General: Automation, Performance,</i>	Using a smart phone increases my effectiveness on my jobs.	1	2	3	4	5
<i>Expectancy</i>	Using a smart phone will improve my output quality.	1	2	3	4	5
	Using a smart phone will increase my chances of achieving a higher level of performance.	1	2	3	4	5
<i>General: Automation, Process,</i>	The information that a smart phone provides is of high quality.	1	2	3	4	5
<i>Transparency</i>	A smart phone provides sufficient information.	1	2	3	4	5
	I am satisfied with the information that a smart phone provides.	1	2	3	4	5
<i>General: Automation, Cultural-Technological</i>	I prefer to use a smart phone to make decisions under high workload situations.	1	2	3	4	5
<i>Context</i>	Using a smart phone helps me to expend less effort to accomplish tasks.	1	2	3	4	5
	Using a smart phone helps me accomplish tasks with lower risks.	1	2	3	4	5
<i>Specific: Automation, Performance,</i>	GPS improves my performance.	1	2	3	4	5
<i>Expectancy</i>	GPS enables me to accomplish tasks more quickly.	1	2	3	4	5
	GPS increases my productivity.	1	2	3	4	5
<i>Specific: Automation, Process,</i>	My interaction with GPS is clearly understandable.	1	2	3	4	5
<i>Transparency</i>	GPS is user-friendly.	1	2	3	4	5
	GPS uses appropriate methods to reach decisions.	1	2	3	4	5
<i>Specific: Automation, Purpose,</i>	I am confident about the performance of GPS	1	2	3	4	5
<i>Influence</i>	When an emergent issue or problem arises, I would feel comfortable depending on the information provided by GPS.	1	2	3	4	5
	I can always rely on GPS to ensure my performance.	1	2	3	4	5

## C.8 Semantic Pairs for Credibility [Ohanian, 1990]

### **Attractiveness**

Attractive-Unattractive

Classy-Not Classy

Beautiful-Ugly

Elegant-Plain

Sexy-Not sexy

### **Trustworthiness**

Dependable-Undependable

Honest-Dishonest

Reliable-Unreliable

Sincere-Insincere

Trustworthy-Untrustworthy

### **Expertise**

Expert-Not an expert

Experienced-Inexperienced

Knowledgeable-Unknowledgeable

Qualified-Unqualified

Skilled-Unskilled

## C.9 Trust in Automation Questionnaire [Merritt, 2011]

- I believe the AWD is a competent performer
- I trust the AWD
- I have confidence in the advice given by the AWD
- I can depend on the AWD
- I can rely on the AWD to behave in consistent ways
- I can rely on the AWD to do its best every time I take its advice

### C.10 Pedestrian Receptivity Questionnaire [Deb et al., 2017]

A fully autonomous vehicle (FAV) is driven by technology instead of by a human. A FAV is equipped with radars, cameras, and sensors which can detect the presence, position, and speed of other vehicles or road-users. With this information, the FAV can then respond as needed by stopping, decelerating and/or changing direction. A driverless vehicle has the potential to reduce pedestrian-motor vehicle crashes and to decrease the possibility of severe injuries by controlling the driving task effectively. You have recently learned that there will be fully autonomous vehicles on the road in your area. As you consider this, how much would you agree or disagree with the following statements.

All items will be measured on the following 7-point Likert scale:

- 1 = strongly disagree
- 2 = moderately disagree
- 3 = somewhat disagree
- 4 = neutral (neither disagree nor agree)
- 5 = somewhat agree
- 6 = moderately agree
- 7 = strongly agree

1. (A) FAVs will enhance the overall transportation system.
2. (A) FAVs will make the roads safer.
3. (A) I would feel safe to cross roads in front of FAVs.
4. (A) It would take less effort from me to observe the surroundings and cross roads if there are FAVs involved.
5. (A) I would find it pleasant to cross the road in front of FAVS.
6. (S) People who influence my behavior would think that I should cross roads in front of FAVs.
7. (S) People who are important to me would not think that I should cross roads in front of FAVs.[reverse-scaled]
8. (S) People who are important to me and/or influence my behavior trusts FAVs (or has a positive attitude towards FAVs).



9. (E) Interacting with the system would not require a lot of mental effort.
10. (E) FAV can correctly detect pedestrians on streets.
11. (T) I would feel comfortable if my child, spouse, parents – or other loved ones – cross roads in the presence of FAVs.
12. (T) I would recommend my family and friends to be comfortable while crossing roads in front of FAVs.
13. (T) I would feel more comfortable doing other things (e.g., checking emails on my smartphone, talking to my companions) while crossing the road in front of FAVs.
14. (C) The traffic infrastructure supports the launch of FAVs.
15. (C) FAV is compatible with all aspects of transportation system in my area.
16. (E, C) FAVs will be able to effectively interact with other vehicles and pedestrians.

Note: A-Attitude, S-Social norms, E-Effectiveness, T-Trust, C-Compatibility. Higher scores indicate higher receptivity toward FAV.

#### C.11 Trust in E-Commerce [McKnight et al., 2002]

See the next page.

	<b>Disposition to Trust</b>
<i>Benevolence</i>	<ol style="list-style-type: none"> <li>1. In general, people really do care about the well-being of others.</li> <li>2. The typical person is sincerely concerned about the problems of others.</li> <li>3. Most of the time, people care enough to try to be helpful, rather than just looking out for themselves.</li> </ol>
<i>Integrity</i>	<ol style="list-style-type: none"> <li>1. In general, most folks keep their promises.</li> <li>2. I think people generally try to back up their words with their actions.</li> <li>3. Most people are honest in their dealings with others.</li> </ol>
<i>Competence</i>	<ol style="list-style-type: none"> <li>1. I believe that most professional people do a very good job at their work.</li> <li>2. Most professionals are very knowledgeable in their chosen field.</li> <li>3. A large majority of professional people are competent in their area of expertise.</li> </ol>
<i>Trusting Stance</i>	<ol style="list-style-type: none"> <li>1. I usually trust people until they give me a reason not to trust them.</li> <li>2. I generally give people the benefit of the doubt when I first meet them.</li> <li>3. My typical approach is to trust new acquaintances until they prove I should not trust them.</li> </ol>
	<b>Institution-Based Trust</b>
<i>Situational Normality-General</i>	<ol style="list-style-type: none"> <li>1. I feel good about how things go when I do purchasing or other activities on the Internet.</li> <li>2. I am comfortable making purchases on the Internet.</li> </ol>
<i>Situational Normality-Benevolence</i>	<ol style="list-style-type: none"> <li>1. I feel that most Internet vendors would act in a customers' best interest.</li> <li>2. If a customer required help, most Internet vendors would do their best to help.</li> <li>3. Most Internet vendors are interested in customer well-being, not just their own wellbeing.</li> </ol>
<i>Situational Normality-Integrity</i>	<ol style="list-style-type: none"> <li>1. I am comfortable relying on Internet vendors to meet their obligations.</li> <li>2. I feel fine doing business on the Internet since Internet vendors generally fulfill their agreements.</li> <li>3. I always feel confident that I can rely on Internet vendors to do their part when I interact with them.</li> </ol>
<i>Situational Normality-Competence</i>	<ol style="list-style-type: none"> <li>1. In general, most Internet vendors are competent at serving their customers.</li> <li>2. Most Internet vendors do a capable job at meeting customer needs.</li> <li>3. I feel that most Internet vendors are good at what they do.</li> </ol>
<i>Structural Assurance</i>	<ol style="list-style-type: none"> <li>1. The Internet has enough safeguards to make me feel comfortable using it to transact personal business.</li> <li>2. I feel assured that legal and technological structures adequately protect me from problems on the Inthernet.</li> <li>3. I feel confident that encryption and other technological advances on the Internet make it safe for me to do business there.</li> <li>4. In general, the Internet is now a robust and safe environment in which to transact business.</li> </ol>
	<b>Trusting Beliefs</b>
<i>Benevolence</i>	<ol style="list-style-type: none"> <li>1. I believe that LegalAdvice.com would act in my best interest.</li> <li>2. If I required help, LegalAdvice.com would do its best to help me.</li> <li>3. LegalAdvice.com is interested in my well-being, not just its own.</li> </ol>
<i>Integrity</i>	<ol style="list-style-type: none"> <li>1. LegalAdvice.com is truthful in its dealings with me.</li> <li>2. I would characterize LegalAdvice.com as honest.</li> <li>3. LegalAdvice.com would keep its commitments.</li> <li>4. LegalAdvice.com is sincere and genuine.</li> </ol>
<i>Competence</i>	<ol style="list-style-type: none"> <li>1. LegalAdvice.com is competent and effective in providing legal advice.</li> <li>2. LegalAdvice.com performs its role of giving legal advice very well.</li> <li>3. Overall, LegalAdvice.com is a capable and proficient Internet legal advice provider.</li> <li>4. In general, LegalAdvice.com is very knowledgeable about the law.</li> </ol>
	<b>Trusting Intentions</b>
<i>Willingness to Depend</i>	<ol style="list-style-type: none"> <li>1. When an important legal issue or problem arises, I would feel comfortable depending on the information provided by LegalAdvice.com</li> <li>2. I can always rely on LegalAdvice.com in a tough legal situation.</li> <li>3. I feel that I could count on LegalAdvice.com to help with a crucial legal problem.</li> <li>4. Faced with a difficult legal situation that required me to hire a lawyer (for a fee), I would use the firm backing LegalAdvice.com</li> </ol>
<i>Subjective Probability of Depending : Follow Advice</i>	<ol style="list-style-type: none"> <li>1. If I had a challenging legal problem, I would want to use LegalAdvice.com again.*</li> <li>2. I would feel comfortable acting on the landlord/tenant information given to me by LegalAdvice.com</li> <li>3. I would not hesitate to use the landlord/tenant information LegalAdvice.com supplied me</li> <li>4. I would confidently act on the legal advice I was given by LegalAdvice.com.</li> <li>5. I would feel secure in using the landlord/tenant information from LegalAdvice.com.</li> <li>6. Based on the advice I just read, I would serve notice, wait, go ahead and get the repair done, and then deduct the cost of the repair from my rent.</li> </ol>
<i>Subjective Probability of Depending : Give Information</i>	<p>Suppose you wanted more specific information about landlord/tenant relationships and you could consult (one time only) by telephone with one of the LegalAdvice.com lawyers for 15-30 minutes (free of charge). For this service, please answer the following:</p> <ol style="list-style-type: none"> <li>1. I would be willing to provide information like my name, address, and phone number to LegalAdvice.com.</li> <li>2. I would be willing to provide my social security number to LegalAdvice.com.</li> <li>3. I would be willing to share the specifics of my legal issue with LegalAdvice.com.</li> </ol>
<i>Subjective Probability of Depending : Make Purchases</i>	<p>Suppose the LegalAdvice.com site was not free, but charged to access information on the site. Answer the following questions:</p> <ol style="list-style-type: none"> <li>1. Faced with a difficult legal situation, I would be willing to pay to access information on the LegalAdvice.com Web site.</li> <li>2. I would be willing to provide credit card information on the LegalAdvice.com Web site.</li> <li>3. Given a tough legal issue, I would be willing to pay for a 30-minute phone consultation with a LegalAdvice.com lawyer.</li> </ol>



# D

## *Interview Questions for AI Practitioners*

The main list of questions asked to AI practitioners during the semi-structured interviews. These interviews are a larger research project, and for this study's analysis, we focused on the data mainly from the first two parts of the protocol. Questions not considered in this study are in lightgray.

### *D.0.1 Background*

1. How would you describe your role in the company? What does your role imply?
2. What type of AI-embedded system you are working on?
  - What is its main objective and context of use?
  - What kind of users interact with the system?
  - Who are other stakeholders involved around the system?

### *D.0.2 Understanding of Human-AI Trust*

1. How would you define Human-AI trust in your own words? what comes to your mind when I say it?
2. How would you define trustworthy AI in your own words? what comes to your mind when I say it?
3. What role does Human-AI trust play for your company/team?

- What is your company's/team's strategy to establish trust in your AI?
- At what moment of project development do you play out these strategies?
- Do you read any literature related to Human-AI trust and trust-worthy AI? What kind?

### *D.0.3 Evaluation of Human-AI Trust*

- How would you know to which extent someone trusts your AI?
- Has your team/company ever evaluated someone's trust in your AI?
  - [yes] What methods did you use to obtain this information?
- no Why do you think your team/company has never evaluated someone's trust in your AI?
  - [no] Imagine for a project you need to estimate trust in your AI. What would you do to obtain this information?
- I will now present three guidelines aiming to support evaluation of Human-AI trust. After I read each of them, could you please tell me whether you have ever heard about the statement and whether it might be useful for your work?
  - Trust in AI and compliance with its recommendations are two different concepts: it is not because a person has decided to follow an AI recommendation that he or she trusts it;
  - Trust is something that evolves over time, so it must be evaluated over a long period of time;
  - If, during the evaluation, the decisions made by users have no consequences (they should not necessarily be real), the data collected might not be related to trust.

# *E*

## *Interview Questions for Decision Subjects*

The main list of questions asked to decision subjects during the semi-structured interviews. These interviews are a larger research project, and for this study's analysis, we focused on the data mainly from the second part of the protocol. Questions not considered in this study are in lightgray.

### *E.0.1 Experiences with Human-AI Decision Making*

- Could you please tell me about your case and your experience with AI decision making?
  - What was the system?
  - When was it?
  - What happened exactly?
- How did you find out that you were subject of an AI decision making?
  - Did you try to find out how the decision was made? How?
  - Could you understand the decision? Why? Why not?
  - Did you try to change/influence the decision? How? Did it work?
  - Were you satisfied with the decision? Why? Why not?

- Would you prefer that a human being had made the decision? Why? Why not?

### *E.o.2 Understanding of Human-AI Trust*

1. How would you define Human-AI trust in your own words? what comes to your mind when I say it?
2. How would you define trustworthy AI in your own words? what comes to your mind when I say it?
3. How important is trust when it comes to decision making support systems?
  - Can you name some AI-embedded systems that you trust/do not trust? Why do you trust them/not trust them?
  - Has it ever happened to you that you trusted an AI-embedded system too much/too little? What happened? How did you feel in that situation?
  - Are there use cases in which trust is more important than in others? In which ones?
  - Based on the competence, where do you think decision making support system stand in comparison with you on a hierarchical level (i.e., same level – colleague, co-worker; lower level – intern, assistant, right hand etc., higher level – senior advisor, boss etc.)?

### *E.o.3 Opinions about Industry's Efforts towards Trustworthy AI*

- Do you think the developers of decision making support systems have considered users' trust when developing their system?
- What do you think are the biggest challenges in developing trustworthy AI?
- I will now present several practices related to development of trustworthy AI. After I read each of them, could you please tell me whether you have ever heard about the statement and what do you think about this practice as a way to learn your trust in decision making support system? Are they sufficient or not?
  - AI certification (technical robustness and safety);
  - Providing explanations and other transparency elements (access to data, being able to manipulate the data/other features, etc.)
  - Educational sessions about how the system works;
  - Conducting evaluations of users' trust in the system

# F

## *Systematic selection of empirical studies of Human-AI trust in the decision making context*

To select the empirical papers that study Human-AI trust in the context of decision making, we followed the methodology of the systematic review on evaluation of Human-AI trust in the context of decision making [Vereschak et al., 2021]. We included all the studies they reviewed (n = 83) from ACM Digital Library, all of them were published prior to January 2021 (their search phase). To find new studies that were published between January 2021 and early July 2022 (our search phase), we used the search string as reported in [Vereschak et al., 2021]:

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

This resulted in 788 papers published in ACM Digital Library over that period. The main selection criteria were 1) results section discussing Human-AI trust, 2) presence of an empirical study with human participants, 3) interaction with or discussion about AI-embedded system,



4) interaction or discussion should include humans making a decision,  
5) full paper. For more detailed description of each criterion, refer to Section 3.2 of [Vereschak et al., 2021].

We selected the papers manually in two rounds. First, we only read the papers' titles, abstracts, and checked their formats. We excluded 139 papers, because their format was other than a full paper (i.e. abstract, poster) and 532 papers, because they were out of the scope defined above. This left us with 117 papers, and the first author read their full texts twice in a randomly shuffled order to make the final selection. The second read took 10 days after the first one, without seeing the notes taken during the first pass. In this step, the author excluded 87 papers, according to the selection criteria defined above. This resulted in 30 newly selected papers, and thus the final corpus to systematically review of 113 papers.

# G

*Summary of all the discussed  
Human-AI trust factors in the  
context of decision making*

		AI Practitioners	Decision Subjects	Academic Literature
<b>Socio-Technological Context</b>	<b>Human-Human Trust</b>			
	<i>Trust in AI team</i>	P2, P4, P7	DS7	-
	<i>Trust in other users</i>	P3, P6	-	46; 87; 156; 266; 316
	<b>Time Dynamics</b>			
	<i>Long-term interaction</i>	P1, P2, P4, P7	-	68; 214; 266; 343; 396
	<i>Delay of AI recommendation</i>	-	-	50; 181; 273
	<b>Type of Task</b>			
	<i>Subjective evaluations</i>	-	DS4	179; 189
	<i>Task Complexity</i>	P6	DS5	7; 71; 95; 130; 249; 343; 346; 359; 388
	<i>Responsibility / risk</i>	-	-	49; 88; 115; 179; 214; 249; 298; 361
<b>Marketing</b>				
<i>System terminology</i>	P4	DS1, DS7	172; 189	

Table G.1 continued from previous page

	AI Practitioners	Decision Subjects	Academic Studies on Trust in DM Context
<i>Reliability and values signaling</i>			95; 196; 292; 383; 392
<i>First interaction</i>			351
<b>Performance and Errors</b>	P2	DS2, DS4, DS6, DS7	17; 63; 94; 95; 108; 115; 127; 130; 137; 143; 161; 183; 186; 213; 214; 220; 261; 271; 273; 292; 351; 370; 371; 384; 389; 392; 395-397; no effect: 10; 45; 109; 302; 321; 350
<i>Context of errors</i>	P1	DS4	-
<i>Frequency of errors</i>	P2	DS4	-
<i>Nature of errors</i>	-	DS3, DS6	no effect 109
<i>Relativity of performance: system</i>	P2	DS2, DS4	-
<i>Relativity of performance: humans</i>	P7	DS2, DS4, DS5	54; 361; 402; 404, relative error tolerance: 278; 279; 298; 312; 352; 382; 384
<i>Robustness</i>	-	-	196
<i>Usability</i>	-	-	108
<i>Errors x design</i>	-	-	400

Table G.1 continued from previous page

		AI Practitioners	Decision Subjects	Academic Studies on Trust in DM Context
Systems' Development and Design	<i>Errors x interactivity</i>	-	-	129
	<i>Errors x expectations</i>	-	-	54
	<i>Fairness and Biases</i>	-	-	173; 340; 382
	<b>Interactivity</b>	P1, P2, P3, P4	DS2, DS3, DS4, DS6	45; 53; 129; 266; 316
	<b>Transparency</b>			
	<i>Working process</i>	P4, P7	-	343
	<i>Data</i>	P7	-	18; 84; 123; 196; 272
	<i>Explanations</i>	P1, P2, P3, P6; P4, P5, P7 disagree	DS3, DS4, DS7 disagree	49; 97; 99; 117; 145; 164; 186; 205; 321; 372; 373; 394; no effect: 268; 271; 329; 402
	<i>Type of explanations</i>	-	-	345; 372; 373; 387; no effect: 270; 299
	<i>AI confidence score</i>	-	-	30; 76; 139; 163; 288; 292; 323; 371; 389; 402; no effect:
	<i>Social transparency</i>	-	-	370; 371 87

Table G.1 continued from previous page

	AI Practitioners	Decision Subjects	Academic Studies on Trust in DM Context
<b>AI Certification</b>	P1, P4, P6 (P5 disagrees)	DS1-DS3, DS5, DS7	156; 196
<b>Appearance</b>	-	-	2; 108; 127; 161; 183; 203; 253; 253; 274; 288; 332; 353; 354; 400
<b>Communication Style</b>	-	-	17; 45; 71; 195; 253; 261; 287; 288; 354; 368; 380; 384; 387;
<b>Privacy</b>	-	-	404 196
<b>Agency</b>			
<i>Direct users</i>	P4	-	10; 123; 172; 181; 230; 290; 346; 359; 383; 385; no effect; 50; 104; 321; 352
<i>Decision subjects</i>	P7	DS1, DS4, DS5, DS6	-
<i>Agency awareness</i>	-	DS7	-
<b>Element of Unexpected and Expected</b>	P6		68; 175; 220; 350
<i>Good surprise</i>	P1, P2	DS1, DS2, DS4, DS7	-

Table G.1 continued from previous page

		AI Practitioners	Decision Subjects	Academic Studies on Trust in DM Context
People's Preferences and Experiences	<i>Bad surprise</i>	P1, P2	-	-
	<i>Prior experiences</i>	-	DS3, DS6	46; 196
	<b>AI Literacy</b>			
	<i>Direct users</i>	P5, P7	-	54; 65; 117; 172; 295; no effect: 351; 387
	<i>Decision subjects</i>	-	DS2 (DS4, DS6, DS7 - disagree)	-
	<b>Domain Expertise</b>			
	<i>Actual expertise</i>	P6	-	95; 99; 115; 117; 163; 316; 321; 359; 402; 404; no effect: 387
	<i>Self-confidence</i>	-	-	163; 203; 321; no effect: 290
	<b>Individual differences</b>			
	<i>Age</i>	-	-	107; 107; 271; 351
	<i>Education</i>	-	-	372; 373; no effect: 351
	<i>Personality</i>	-	-	149
	<i>Propensity to trust</i>	-	-	97; 149; no effect: 351; 387
<i>Culture</i>	-	-	63; 179; 368; no effect: 351	
<i>Emotional state</i>	-	-	94	

Table G.1 continued from previous page

	AI Practitioners	Decision Subjects	Academic Studies on Trust in DM Context
<i>Gender</i>	-	-	no effect: 271; 351
<i>Work style</i>	-	-	117; 149; 358
<i>Attitudes towards robots</i>	-	-	351
<b>Operator Workload</b>	-	-	278; inconclusive effects: 63; 130





## Bibliography

- [1] Accenture Federal Services. Responsible AI: A framework for building trust in your AI solutions. Technical report, Accenture, January 2019. URL <https://www.accenture.com/us-en/insights/us-federal-government/ai-is-ready-are-we>.
- [2] Sander Ackermans, Debargha Dey, Peter Ruijten, Raymond H. Cuijpers, and Bastian Pfleging. The effects of explicit intention communication, conspicuous sensors, and pedestrian attitude in interactions with automated vehicles. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20*, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376197. URL <https://doi.org/10.1145/3313831.3376197>.
- [3] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018. doi: 10.1109/ACCESS.2018.2870052.
- [4] Barbara D. Adams, Lora E. Bruyn, Sébastien Houde, and Paul Angelopoulos. Trust in automated systems literature review. Technical report, Department Of National Defence of Canada, June 2003.
- [5] William C. Adams. *Conducting Semi-Structured Interviews*, chapter 19, pages 492–505. John Wiley Sons, Ltd, 2015. ISBN 9781119171386. doi: <https://doi.org/10.1002/9781119171386.ch19>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119171386.ch19>.
- [6] AI Taskforce. Report of Estonia’s AI taskforce. Technical report, Republic of Estonia Government Office and Republic of Estonia Ministry of Economic Affairs and Communications, Esto-

nia, May 2019. URL <https://ec.europa.eu/knowledge4policy/ai-watch/estonia-ai-strategy-report>.

- [7] Ighoyota Ben. Ajenaghughrure, Sonia C. Sousa, Ilkka Johannes Kosunen, and David Lamas. Predictive model to assess user trust: A psycho-physiological approach. In *Proceedings of the 10th Indian Conference on Human-Computer Interaction, IndiaHCI '19*, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450377164. doi: 10.1145/3364183.3364195. URL <https://doi.org/10.1145/3364183.3364195>.
- [8] Ifeoma Ajunwa, Sorelle Friedler, Carlos E Scheidegger, and Suresh Venkatasubramanian. Hiring by algorithm: predicting and preventing disparate impact. *Available at SSRN*, 2746078, 2016.
- [9] George A. Akerlof. The Market for “Lemons”: Quality Uncertainty and the Market Mechanism\*. *The Quarterly Journal of Economics*, 84(3):488–500, 08 1970. ISSN 0033-5533. doi: 10.2307/1879431. URL <https://doi.org/10.2307/1879431>.
- [10] Alper Alan, Enrico Costanza, Joel Fischer, Sarvapali D. Ramchurn, Tom Rodden, and Nicholas R. Jennings. A field study of Human-Agent interaction for electricity tariff switching. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS '14*, page 965–972, New York, NY, USA, 2014. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450327381.
- [11] Ahmed J. Aljaaf, Dhiya Al-Jumeily, Abir J. Hussain, Paul Fergus, Mohammed Al-Jumaily, and Khaled Abdel-Aziz. Toward an optimal use of artificial intelligence techniques within a clinical decision support system. In *2015 Science and Information Conference (SAI)*, pages 548–554, 2015. doi: 10.1109/SAI.2015.7237196.
- [12] S. L. Alter. *Computer aided decision making in organizations : a decision support systems typology*. M.I.T. Center for Information Systems Research, 03 1976.
- [13] Carlos Alós-Ferrer and Federica Farolfi. Trust games and beyond. *Frontiers in Neuroscience*, 13(887):1–14, September 2019. doi: 10.3389/fnins.2019.00887. URL <https://www.frontiersin.org/article/10.3389/fnins.2019.00887>.
- [14] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, and et al. Guidelines for Human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human*

- Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300233. URL <https://doi.org/10.1145/3290605.3300233>.
- [15] Asbjørn Ammitzbøll Flügge, Thomas Hildebrandt, and Naja Holten Møller. Street-level algorithms and AI in bureaucratic decision-making: A caseworker perspective. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021. doi: 10.1145/3449114. URL <https://doi.org/10.1145/3449114>.
- [16] Bengt-Erik Andersson and Stig-Göran Nilsson. Studies in the reliability and validity of the critical incident technique. *Journal of Applied Psychology*, 48:398–403, 1964. ISSN 1939-1854. URL <https://doi.org/10.1037/h0042025>.
- [17] Sean Andrist, Erin Spannan, and Bilge Mutlu. Rhetorical robots: Making robots more effective speakers using linguistic cues of expertise. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '13, page 341–348, New York, NY, USA, 2013. IEEE Press. ISBN 9781467330558.
- [18] Ariful Islam Anik and Andrea Bunt. Data-centric explanations: Explaining training data of machine learning systems to promote transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445736. URL <https://doi.org/10.1145/3411764.3445736>.
- [19] Anonymous. Incident 37: Woman down-ranked by amazon recruiting tool, Aug 2016. URL <https://incidentdatabase.ai/cite/37>.
- [20] Musbah Aqel, Omar Nakshabandi, and Ayodeji Adeniyi. Decision support systems classification in industry. *Periodicals of Engineering and Natural Sciences (PEN)*, 7:774–785, 08 2019. doi: 10.21533/pen.v7i2.550.
- [21] Lise Arena, Nathalie Oriol, and Iryna Veryzhenko. Too fast, too furious? algorithmic trading and financial instability. *Systèmes d'Information et Management*, 23:81–106, 01 2018. doi: 10.3917/sim.182.0081.
- [22] Melanie J. Ashleigh and Edgar Meyer. Deepening the understanding of trust: combining repertory grid and narrative to explore the uniqueness of trust. In Fergus Lyon, Guido Möllering, and Mark Saunders, editors, *Handbook of Research Methods*

on *Trust*, chapter 14, pages 138–148. Edward Elgar, Cheltenham, UK; Northampton, MA, USA, 01 2011.

- [23] Maryam Ashoori and Justin D. Weisz. In AI we trust? factors that influence trustworthiness of AI-infused decision-making processes, 2019.
- [24] Benoit A. Aubert and Barbara L. Kelsey. Further understanding of trust and performance in virtual teams. *Small Group Research*, 34(5):575–618, 2003. doi: 10.1177/1046496403256011. URL <https://doi.org/10.1177/1046496403256011>.
- [25] AXA Research Fund. Artificial intelligence: Fostering trust. Technical report, AXA, March 2019. URL <https://www.axa-research.org/en/news/AI-research-guide>.
- [26] Jacqui Ayling and Adriane Chapman. Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics*, 2, 08 2022. doi: 10.1007/s43681-021-00084-x.
- [27] Reinhard Bachmann. Utilising repertory grids in macro-level comparative studies. In Fergus Lyon, Guido Möllering, and Mark Saunders, editors, *Handbook of Research Methods on Trust*, chapter 13, pages 130–137. Edward Elgar, Cheltenham, UK; Northampton, MA, USA, 01 2011.
- [28] Annette Baier. Trust and antitrust. *Ethics*, 96(2):231–260, January 1986.
- [29] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 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, AAAI’19/IAAI’19/EAAI’19*. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33012429. URL <https://doi.org/10.1609/aaai.v33i01.33012429>.
- [30] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI ’21*, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445717. URL <https://doi.org/10.1145/3411764.3445717>.

- [31] Brad M. Barber and Terrance Odean. Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, 116(1):261–292, 2001. ISSN 00335533, 15314650. URL <http://www.jstor.org/stable/2696449>.
- [32] Solon Barocas, Andrew D. Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT\* '20*, page 80–89, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372830. URL <https://doi.org/10.1145/3351095.3372830>.
- [33] Thomas Baudel, Manon Verbockhaven, Victoire Cousergue, Guillaume Roy, and Rida Laarach. Objectivaize: Measuring performance and biases in augmented business decision systems. In Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen, editors, *Human-Computer Interaction – INTERACT 2021*, pages 300–320, Cham, 2021. Springer International Publishing. ISBN 978-3-030-85613-7.
- [34] Roy F. Baumeister. Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology*, 46(3):610–620, 1984. doi: 10.1037/0022-3514.46.3.610. URL <https://doi.org/10.1037/0022-3514.46.3.610>.
- [35] Markus Bertl, Janek Metsallik, and Peeter Ross. A systematic literature review of AI-based digital decision support systems for post-traumatic stress disorder. *Front Psychiatry*, 13:923613, August 2022.
- [36] Jason Bindewald, Christina Rusnock, and Michael Miller. Measuring human trust behavior in Human-Machine teams. pages 47–58, 06 2018. ISBN 978-3-319-60590-6. doi: 10.1007/978-3-319-60591-3\_5.
- [37] Ann M Bisantz and Younho Seong. Assessment of operator trust in and utilization of automated decision-aids under different framing conditions. *International Journal of Industrial Ergonomics*, 28(2):85–97, 2001.
- [38] Gerd Bohner and Nina Dickel. Attitudes and attitude change. *Annual Review of Psychology*, 62(1):391–417, 2011. doi: 10.1146/annurev.psych.121208.131609. URL <https://doi.org/10.1146/annurev.psych.121208.131609>. PMID: 20809791.

- [39] Iris Bohnet, Fiona Greig, Benedikt Herrmann, and Richard Zeckhauser. Betrayal aversion: Evidence from brazil, china, oman, switzerland, turkey, and the united states. *American Economic Review*, 98(1):294–310, 2008. URL <http://dx.doi.org/10.1257/aer.98.1.294>.
- [40] S. Boon and J. Holmes. The dynamics of interpersonal trust: resolving uncertainty in the ace of risk. In R. Hinde and J. Gorebel, editors, *Cooperation and Prosocial Behaviour*, pages 190–211, Cambridge, 1991. Cambridge University Press.
- [41] Sviatoslav Braynov. Contracting with uncertain level of trust. *Computational Intelligence*, 18:501–514, 2002.
- [42] Sviatoslav Braynov. What human trust is and is not: On the biology of human trust. *AAAI Spring Symposium: Trust and Autonomous Systems*, pages 10–15, 2013.
- [43] Gerard Breeman. Hermeneutic methods in trust research. In Fergus Lyon, Guido Möllering, and Mark Saunders, editors, *Handbook of Research Methods on Trust*, chapter 15, pages 149–160. Edward Elgar, Cheltenham, UK; Northampton, MA, USA, 01 2011.
- [44] Gerard Engelbert Breeman. *Cultivating trust : how do public policies become trusted*. PhD thesis, Dept. of Public Administration, Faculty of Social and Behavioural Sciences, Leiden University, 2006.
- [45] Tom Bridgwater, Manuel Giuliani, Anouk van Maris, Greg Baker, Alan Winfield, and Tony Pipe. 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*, page 23–31, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367462. doi: 10.1145/3319502.3374804. URL <https://doi.org/10.1145/3319502.3374804>.
- [46] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 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*, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300271. URL <https://doi.org/10.1145/3290605.3300271>.

- [47] Jacob Browne, Saskia Bakker, Bin Yu, Peter Lloyd, and S. Ben Allouch. Trust in clinical AI: Expanding the unit of analysis. In *The First International Conference on Hybrid-Human AI*, 2022.
- [48] Marc Brysbaert. How many participants do we have to include in properly powered experiments? a tutorial of power analysis with reference tables. volume 2, USA, July 2019. doi: 10.5334/joc.72. URL <https://www.journalofcognition.org/articles/10.5334/joc.72/>.
- [49] Zana Bućinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI '20*, page 454–464, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371186. doi: 10.1145/3377325.3377498. URL <https://doi.org/10.1145/3377325.3377498>.
- [50] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 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), apr 2021. doi: 10.1145/3449287. URL <https://doi.org/10.1145/3449287>.
- [51] Jeremy Burke and Angela A. Hung. Trust and financial advice. *Journal of Pension Economics and Finance*, 20(1):9–26, 2021. doi: 10.1017/S147474721900026X.
- [52] Terence Burnham, Kevin McCabe, and Vernon Smith. Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavior & Organization*, 43(1):57–73, 2000. URL <https://EconPapers.repec.org/RePEc:eee:jeborg:v:43:y:2000:i:1:p:57-73>.
- [53] 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. 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*, page 1–14, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300234. URL <https://doi.org/10.1145/3290605.3300234>.
- [54] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. “hello AI”: Uncovering the onboarding needs of medical practitioners for Human-AI collaborative



- decision-making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), 2019. doi: 10.1145/3359206. URL <https://doi.org/10.1145/3359206>.
- [55] Kelly Caine. Local standards for sample size at chi. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 981–992, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858498. URL <https://doi-org.accesdistant.sorbonne-universite.fr/10.1145/2858036.2858498>.
- [56] Colin F. Camerer and Robin Hogarth. The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19(1/3):7–42, 1999. ISSN 08955646, 15730476. URL <http://www.jstor.org/stable/41760945>.
- [57] Janis A Cannon-Bowers and Eduardo Salas. Individual and team decision making under stress: Theoretical underpinnings. In *Making decisions under stress: Implications for individual and team training*, pages 17–38. American Psychological Association, Washington, 1998.
- [58] Cristiano Castelfranchi and Rino Falcone. *Socio-Cognitive Model of Trust: Basic Ingredients*, chapter 2, pages 35–94. John Wiley and Sons, Ltd, Chichester, United Kingdom, 2010. doi: 10.1002/9780470519851.ch2. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470519851.ch2>.
- [59] Noah Castelo, Maarten Bos, and Donald Lehmann. Task-dependent algorithm aversion. *Journal of Marketing Research*, 56:002224371985178, 07 2019. doi: 10.1177/0022243719851788.
- [60] William L. Cats-Baril and George P. Huber. Decision support systems for ill-structured problems: An empirical study. *Decision Sciences*, 18(3):350–372, 1987. doi: <https://doi.org/10.1111/j.1540-5915.1987.tb01530.x>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5915.1987.tb01530.x>.
- [61] Raja Chatila, Virginia Dignum, Michael Fisher, Fosca Giannotti, Katharina Morik, Stuart Russell, and Karen Yeung. *Trustworthy AI*, pages 13–39. Springer International Publishing, Cham, 2021. ISBN 978-3-030-69128-8. doi: 10.1007/978-3-030-69128-8\_2. URL [https://doi.org/10.1007/978-3-030-69128-8\\_2](https://doi.org/10.1007/978-3-030-69128-8_2).
- [62] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren. The state of the art in enhancing trust in machine learning models with the use of visualizations. *Computer*

- Graphics Forum*, 39(3):713–756, 2020. doi: <https://doi.org/10.1111/cgf.14034>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14034>.
- [63] Shih-Yi Chien, Michael Lewis, Katia Sycara, Jyi-Shane Liu, and Asiye Kumru. The effect of culture on trust in automation: Reliability and workload. *ACM Trans. Interact. Intell. Syst.*, 8(4), November 2018. ISSN 2160-6455. doi: 10.1145/3230736. URL <https://doi.org/10.1145/3230736>.
- [64] Jin-Hee Cho, Kevin Chan, and Sibel Adali. A survey on trust modeling. *ACM Comput. Surv.*, 48(2), October 2015. ISSN 0360-0300. doi: 10.1145/2815595. URL <https://doi.org/10.1145/2815595>.
- [65] Michael Chromik, Florian Lachner, and Andreas Butz. *ML for UX? - An Inventory and Predictions on the Use of Machine Learning Techniques for UX Research*. NordiCHI '20. Association for Computing Machinery, New York, NY, USA, 2020. ISBN 9781450375795. URL <https://doi.org/10.1145/3419249.3420163>.
- [66] Victoria Clarke and Virginia Braun. Teaching thematic analysis: Overcoming challenges and developing strategies for effective learning. *The Psychologist*, 26:120–123, 02 2013.
- [67] I. Glenn Cohen. Informed consent and medical artificial intelligence: What to tell the patient? *Georgetown Law Journal*, 108:1425–1469, 2020. doi: 10.2139/ssrn.3529576. URL <https://doi.org/10.2139/ssrn.3529576>.
- [68] Mark Colley, Elvedin Bajrovic, and Enrico Rukzio. Effects of pedestrian behavior, time pressure, and repeated exposure on crossing decisions in front of automated vehicles equipped with external communication. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517571. URL <https://doi.org/10.1145/3491102.3517571>.
- [69] Jason Colquitt, Jeffery Lepine, Ronald Piccolo, Cindy Zapata, and Bruce Rich. Explaining the justice-performance relationship: Trust as exchange deepener or trust as uncertainty reducer? *The Journal of applied psychology*, 97:1–15, 09 2011. doi: 10.1037/a0025208.

- [70] James Cox. How to identify trust and reciprocity. *Games and Economic Behavior*, 46(2):260–281, 2004. URL <https://EconPapers.repec.org/RePEc:eee:gamebe:v:46:y:2004:i:2:p:260-281>.
- [71] Henriette Cramer, Vanessa Evers, Nicander Kemper, and Bob Wielinga. Effects of autonomy, traffic conditions and driver personality traits on attitudes and trust towards in-vehicle agents. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 03, WI-IAT '08*, page 477–482, New York, NY, USA, 2008. IEEE Computer Society. ISBN 9780769534961. doi: 10.1109/WIIAT.2008.326. URL <https://doi.org/10.1109/WIIAT.2008.326>.
- [72] Steven C. Currall and Timothy A. Judge. Measuring trust between organizational boundary role persons. *Organizational Behavior and Human Decision Processes*, 64(2):151–170, 1995. URL <https://doi.org/10.1006/obhd.1995.1097>.
- [73] Shuchisnigdha Deb, Lesley Strawderman, Daniel W. Carruth, Janice DuBien, Brian Smith, and Teena M. Garrison. Development and validation of a questionnaire to assess pedestrian receptivity toward fully autonomous vehicles. *Transportation Research Part C: Emerging Technologies*, 84:178 – 195, 2017. ISSN 0968-090X. doi: <https://doi.org/10.1016/j.trc.2017.08.029>. URL <http://www.sciencedirect.com/science/article/pii/S0968090X17302383>.
- [74] Defense Innovation Board. 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, October 2019. URL [https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB\\$\\_\\$AI\\$\\_\\$PRINCIPLES\\$\\_\\$PRIMARY\\$\\_\\$DOCUMENT.PDF](https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB$_$AI$_$PRINCIPLES$_$PRIMARY$_$DOCUMENT.PDF).
- [75] Lifang Deng and Wai Chan. Testing the difference between reliability coefficients alpha and omega. *Educational and psychological measurement*, 77(2):185–203, Apr 2017. ISSN 1552-3888. doi: 10.1177/0013164416658325. URL <https://pubmed.ncbi.nlm.nih.gov/29795909>.
- [76] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction, HRI '13*, page 251–258, New York, NY, USA, 2013. IEEE Press. ISBN 9781467330558.

- [77] Advait Deshpande and Helen Sharp. Responsible AI systems: Who are the stakeholders? pages 227–236, 07 2022. doi: 10.1145/3514094.3534187.
- [78] M Deutsch. Trust, trustworthiness, and the f scale. *Journal of abnormal and social psychology*, 61:138–140, July 1960. ISSN 0096-851X. doi: 10.1037/h0046501. URL <https://doi.org/10.1037/h0046501>.
- [79] Morton Deutsch. Trust and suspicion. *Journal of Conflict Resolution*, 2(4):265–279, 1958. doi: 10.1177/002200275800200401. URL <https://doi.org/10.1177/002200275800200401>.
- [80] Morton Deutsch. The effect of motivational orientation upon trust and suspicion. *Human Relations*, 13(2):123–139, 1960. doi: 10.1177/001872676001300202. URL <https://doi.org/10.1177/001872676001300202>.
- [81] Graham Dietz and Deanne N. Den Hartog. Measuring trust inside organisations. *Personnel Review*, 35:557–588, 2006. doi: 10.1108/00483480610682299.
- [82] Kurt Dirks and Donald Ferrin. Trust in leadership: Meta-analytic findings and implications for research and practice. *The Journal of applied psychology*, 87:611–28, 09 2002. doi: 10.1037/0021-9010.87.4.611.
- [83] John Donovan and Stuart Madnick. Institutional and ad hoc decision support systems and their effective use. *DataBase*, 8, 01 1977.
- [84] Jaimie Drozdal, Justin Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. Trust in automl: Exploring information needs for establishing trust in automated machine learning systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI '20*, page 297–307, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371186. doi: 10.1145/3377325.3377501. URL <https://doi.org/10.1145/3377325.3377501>.
- [85] Yanqing Duan, John S. Edwards, and Yogesh K Dwivedi. Artificial intelligence for decision making in the era of big data – evolution, challenges and research agenda. *International Journal of Information Management*, 48:63–71, 2019. ISSN 0268-4012. doi: <https://doi.org/10.1016/j.ijinfomgt.2019.01.021>. URL <https://www.sciencedirect.com/science/article/pii/S0268401219300581>.

- [86] Thomas J. Dunn, Thom Baguley, and Vivienne Brunsten. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105(3):399–412, 2014. doi: 10.1111/bjop.12046. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/bjop.12046>.
- [87] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. Expanding explainability: Towards social transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445188. URL <https://doi.org/10.1145/3411764.3445188>.
- [88] Malin Eiband, Sarah Theres Völkel, Daniel Buschek, Sophia Cook, and Heinrich Hussmann. When people and algorithms meet: User-reported problems in intelligent everyday applications. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 96–106, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362726. doi: 10.1145/3301275.3302262. URL <https://doi.org/10.1145/3301275.3302262>.
- [89] Fredrick Ekman, Mikael Johansson, and Jana Sochor. Creating appropriate trust for autonomous vehicle systems: A framework for HMI design. *IEEE Transactions on Human-Machine Systems*, 48(1):95–101, 01 2016.
- [90] M.C. Er. Decision support systems: A summary, problems, and future trends. *Decision Support Systems*, 4(3):355–363, 1988. ISSN 0167-9236. doi: [https://doi.org/10.1016/0167-9236\(88\)90022-X](https://doi.org/10.1016/0167-9236(88)90022-X). URL <https://www.sciencedirect.com/science/article/pii/016792368890022X>.
- [91] European Commission. On artificial intelligence - a european approach to excellence and trust. Technical report, European Commission, Brussels, Belgium, February 2020. URL [https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020\\$.en.pdf](https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020$.en.pdf).
- [92] European Commission-Directorate General for Communications Networks-Content and Technology. Regulation laying down harmonised rules on artificial intelligence (artificial intelligence act). Technical Report 2021/0106/COD, European Commission, Brussels, Belgium, April 2021. URL <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52021PC0206>.

- [93] Anthony M. Evans and Joachim I. Krueger. The psychology (and economics) of trust. *Social and Personality Psychology Compass*, 3(6):1003–1017, 2009. doi: 10.1111/j.1751-9004.2009.00232.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-9004.2009.00232.x>.
- [94] Md Abdullah Al Fahim, Mohammad Maifi Hasan Khan, Theodore Jensen, Yusuf Albayram, and Emil Coman. Do integral emotions affect trust? the mediating effect of emotions on trust in the context of Human-Agent interaction. In *Designing Interactive Systems Conference 2021, DIS '21*, page 1492–1503, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384766. doi: 10.1145/3461778.3461997. URL <https://doi.org/10.1145/3461778.3461997>.
- [95] Xiacong Fan, Sooyoung Oh, Michael McNeese, John Yen, Haydee Cuevas, Laura Strater, and Mica R. Endsley. The influence of agent reliability on trust in Human-Agent collaboration. In *Proceedings of the 15th European Conference on Cognitive Ergonomics: The Ergonomics of Cool Interaction, ECCE '08*, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605583990. doi: 10.1145/1473018.1473028. URL <https://doi.org/10.1145/1473018.1473028>.
- [96] D. S. Fareri, L. J. Chang, and M. R. Delgado. Effects of direct social experience on trust decisions and neural reward circuitry. *Front Neurosci*, 6:148, 2012.
- [97] Anja K. Faulhaber, Ina Ni, and Ludger Schmidt. The effect of explanations on trust in an assistance system for public transport users and the role of the propensity to trust. In *Mensch Und Computer 2021, MuC '21*, page 303–310, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386456. doi: 10.1145/3473856.3473886. URL <https://doi.org/10.1145/3473856.3473886>.
- [98] Ernst Fehr. On the economics and biology of trust. *Journal of the European Economic Association*, 7(2-3):235–266, 2009. doi: 10.1162/JEEA.2009.7.2-3.235. URL <https://onlinelibrary.wiley.com/doi/abs/10.1162/JEEA.2009.7.2-3.235>.
- [99] Shi Feng and Jordan Boyd-Graber. What can AI do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19*, page 229–239, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362726.

doi: 10.1145/3301275.3302265. URL <https://doi.org/10.1145/3301275.3302265>.

- [100] Andrea Ferrario and Michele Loi. How explainability contributes to trust in AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, page 1457–1466, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533202. URL <https://doi.org/10.1145/3531146.3533202>.
- [101] Maria Figueroa-Armijos, Brent B. Clark, and Serge P. da Motta Veiga. Ethical perceptions of AI in hiring and organizational trust: The role of performance expectancy and social influence. *Journal of Business Ethics*, Jun 2022. ISSN 1573-0697. doi: 10.1007/s10551-022-05166-2. URL <https://doi.org/10.1007/s10551-022-05166-2>.
- [102] J. C. Flanagan. The critical incident technique. *The Psychological Bulletin*, 51(4):327–358, 1954.
- [103] Jerry Floersch, Jeffrey L. Longhofer, Derrick Kranke, and Lisa Townsend. Integrating thematic, grounded theory and narrative analysis: A case study of adolescent psychotropic treatment. *Qualitative Social Work*, 9(3):407–425, 2010. doi: 10.1177/1473325010362330. URL <https://doi.org/10.1177/1473325010362330>.
- [104] Riccardo Fogliato, Alexandra Chouldechova, and Zachary Lipton. The impact of algorithmic risk assessments on human predictions and its analysis via crowdsourcing studies. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021. doi: 10.1145/3479572. URL <https://doi.org/10.1145/3479572>.
- [105] Meadhbh Foster and Mark T. Keane. Surprise: Youve got some explaining to do. pages 2321–2326, 2013. URL <http://arxiv.org/abs/1308.2236>.
- [106] Jonathan B. Freeman. Doing psychological science by hand. *Current Directions in Psychological Science*, 27(5):315–323, 2018. doi: 10.1177/0963721417746793. URL <https://doi.org/10.1177/0963721417746793>.
- [107] Anna-Katharina Frison, Laura Aigner, Philipp Wintersberger, and Andreas Riener. Who is generation a? investigating the experience of automated driving for different age groups. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI '18*,

- page 94–104, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359467. doi: 10.1145/3239060.3239087. URL <https://doi.org/10.1145/3239060.3239087>.
- [108] Anna-Katharina Frison, Philipp Wintersberger, Andreas Riener, Clemens Schartmüller, Linda Ng Boyle, Erika Miller, and Klemens Weigl. In ux we trust: Investigation of aesthetics and usability of driver-vehicle interfaces and their impact on the perception of automated driving. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300374. URL <https://doi.org/10.1145/3290605.3300374>.
- [109] Ernestine Fu, Mishel Johns, David A. B. Hyde, Srinath Sibi, Martin Fischer, and David Sirkin. Is too much system caution counterproductive? effects of varying sensitivity and automation levels in vehicle collision avoidance systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20*, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376300. URL <https://doi.org/10.1145/3313831.3376300>.
- [110] C. Ashley Fulmer and Michele J. Gelfand. At what level (and in whom) we trust: Trust across multiple organizational levels. *Journal of Management*, 38(4):1167–1230, 2012. doi: 10.1177/0149206312439327. URL <https://doi.org/10.1177/0149206312439327>.
- [111] C. Ashley Fulmer and Michele J. Gelfand. At what level (and in whom) we trust: Trust across multiple organizational levels. *Journal of Management*, 38(4):1167–1230, 2012. doi: 10.1177/0149206312439327. URL <https://doi.org/10.1177/0149206312439327>.
- [112] G20. G20 ministerial statement on trade and digital economy. Technical report, G20, Brussels, Belgium, June 2019. URL <http://trade.ec.europa.eu/doclib/press/index.cfm?id=2027>.
- [113] Diego Gambetta. *Can We Trust Trust?*, chapter 13, pages 213–237. Department of Sociology, University of Oxford, Oxford, United Kingdom, 08 2000. URL <http://www.sociology.ox.ac.uk/papers/gambetta213-237.pdf>.
- [114] Diego Gambetta and Diego Gambetta. *Can We Trust Trust?*, page 213–237. Department of Sociology, University of Oxford, 2000.



- [115] Tamari Gamkrelidze, Moustafa Zouinar, and Flore Barcellini. Working with machine learning/artificial intelligence systems: Workers' viewpoints and experiences. In *European Conference on Cognitive Ergonomics 2021, ECCE 2021*, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450387576. doi: 10.1145/3452853.3452876. URL <https://doi.org/10.1145/3452853.3452876>.
- [116] Meric Altug Gemalmaz and Ming Yin. 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, AIES '22*, page 295–306, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534201. URL <https://doi.org/10.1145/3514094.3534201>.
- [117] Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. Explainable active learning (XAL): Toward AI explanations as interfaces for machine teachers. *Proc. ACM Hum.-Comput. Interact.*, 4(3), January 2021. doi: 10.1145/3432934. URL <https://doi.org/10.1145/3432934>.
- [118] Mahmoud Ghorbel. Researchers at stanford have developed an artificial intelligence (ai) model,' stockbot', which uses lstms to predict stock prices with gains higher than the most aggressive etfs, Jul 2022. URL <https://www.marktechpost.com/2022/07/26/researchers-at-stanford-have-developed-an-artificial-intelligence-ai-model-stockbot-which-uses-lstm>
- [119] Felix Gille, Anna Jobin, and Marcello Ienca. What we talk about when we talk about trust: Theory of trust for AI in healthcare. *Intelligence-Based Medicine*, 1-2:100001, 2020. ISSN 2666-5212. doi: <https://doi.org/10.1016/j.ibmed.2020.100001>. URL <https://www.sciencedirect.com/science/article/pii/S2666521220300016>.
- [120] Nicole Gillespie. *Measuring trust in working relationships: The behavioral trust inventory*. Melbourne Business School, Melbourne, Australia, 2003.
- [121] Nicole Gillespie. Measuring trust in organizational contexts: An overview of survey-based measures. In Fergus Lyon, Guido Möllering, and Mark Saunders, editors, *Handbook of Research Methods on Trust*, chapter 17, pages 175–188. Edward Elgar, Cheltenham, UK; Northampton, MA, USA, 01 2011.
- [122] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations:

An overview of interpretability of Machine Learning. *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89, 2018.

- [123] Alyssa Glass, Deborah L. McGuinness, and Michael Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces, IUI '08*, page 227–236, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781595939876. doi: 10.1145/1378773.1378804. URL <https://doi.org/10.1145/1378773.1378804>.
- [124] Ella Glikson and Anita Woolley. Human trust in artificial intelligence: Review of empirical research (in press). *The Academy of Management Annals*, 14(2), August 2020. doi: <https://doi.org/10.5465/annals.2018.0057>.
- [125] Uri Gneezy and Aldo Rustichini. Pay enough or don't pay at all. *The Quarterly Journal of Economics*, 115(3):791–810, 2000. ISSN 00335533, 15314650. URL <http://www.jstor.org/stable/2586896>.
- [126] Dietz Graham and Den Hartog Deanne N. Measuring trust inside organisations. *Personnel Review*, 35(5):557–588, Jan 2006. ISSN 0048-3486. doi: 10.1108/00483480610682299. URL <https://doi.org/10.1108/00483480610682299>.
- [127] Dara Gruber, Ashley Aune, and Wilma Koutstaal. Can semi-anthropomorphism influence trust and compliance? exploring image use in app interfaces. In *Proceedings of the Technology, Mind, and Society, TechMindSociety '18*, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450354202. doi: 10.1145/3183654.3183700. URL <https://doi.org/10.1145/3183654.3183700>.
- [128] Jonathan Grudin. AI and HCI: Two fields divided by a common focus. *AI Magazine*, 30(4):48–57, September 2009. doi: 10.1609/aimag.v30i4.2271. URL <https://www.aaai.org/ojs/index.php/aimagazine/article/view/2271>.
- [129] Akshit Gupta, Debadeep Basu, Ramya Ghantasala, Sihang Qiu, and Ujwal Gadiraju. 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, WWW '22*, page 3531–3540, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3512248. URL <https://doi.org/10.1145/3485447.3512248>.

- [130] Kunal Gupta, Ryo Hajika, Yun Suen Pai, Andreas Duenser, Martin Lochner, and Mark Billingham. In AI we trust: Investigating the relationship between biosignals, trust and cognitive load in VR. In *25th ACM Symposium on Virtual Reality Software and Technology, VRST '19*, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450370011. doi: 10.1145/3359996.3364276. URL <https://doi.org/10.1145/3359996.3364276>.
- [131] Shivam Gupta, Sachin Modgil, Samadrita Bhattacharyya, and Indranil Bose. Artificial intelligence for decision support systems in the field of operations research: review and future scope of research. *Annals of Operations Research*, 308(1):215–274, Jan 2022. ISSN 1572-9338. doi: 10.1007/s10479-020-03856-6. URL <https://doi.org/10.1007/s10479-020-03856-6>.
- [132] Andreas Gutscher. A trust model for an open, decentralized reputation system. In *Trust Management, IFIPTM 2007*, pages 285–300. Springer US, New Brunswick, Canada, 2007. doi: 10.1007/978-0-387-73655-6\_19. URL [https://doi.org/10.1007/978-0-387-73655-6\\_19](https://doi.org/10.1007/978-0-387-73655-6_19).
- [133] H. Güngör. Creating value with Artificial Intelligence: A multi-stakeholder perspective. *Journal of Creating Value*, 6(1):72–85, 2020. doi: 10.1177/2394964320921071. URL <https://doi.org/10.1177/2394964320921071>.
- [134] Özgür Güreker, Andrea Bönsch, Lucas Braun, Christian Grund, Christine Harbring, Thomas Kittsteiner, and Andreas Staffeldt. Experimental economics in virtual reality, 12 2014.
- [135] Richard D. Hackathorn and Peter G. W. Keen. Organizational strategies for personal computing in decision support systems. *MIS Quarterly*, 5(3):21–27, 1981. ISSN 02767783. URL <http://www.jstor.org/stable/249288>.
- [136] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. de Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in Human-Robot interaction. *Human Factors*, 53(5):517–527, 2011. doi: 10.1177/0018720811417254. URL <https://doi.org/10.1177/0018720811417254>.
- [137] Jason L. Harman, John O'Donovan, Tarek Abdelzaher, and Cleotilde Gonzalez. Dynamics of human trust in recommender systems. In *Proceedings of the 8th ACM Conference on Recommender Systems, RecSys '14*, page 305–308, New York, NY, USA, 2014.

- Association for Computing Machinery. ISBN 9781450326681. doi: 10.1145/2645710.2645761. URL <https://doi.org/10.1145/2645710.2645761>.
- [138] Changlin He and Yufen Li. A survey of intelligent decision support system. In *Proceedings of the 2017 7th International Conference on Applied Science, Engineering and Technology (ICASET 2017)*, pages 201–206. Atlantis Press, 2017. ISBN 978-94-6252-340-1. doi: <https://doi.org/10.2991/icaset-17.2017.38>. URL <https://doi.org/10.2991/icaset-17.2017.38>.
- [139] Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. Presenting system uncertainty in automotive uis for supporting trust calibration in autonomous driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI '13*, page 210–217, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450324786. doi: 10.1145/2516540.2516554. URL <https://doi.org/10.1145/2516540.2516554>.
- [140] Ralph Hertwig and Andreas Ortmann. *Economists' and Psychologists' Experimental Practices: How They Differ, Why They Differ, And How they Could Converge*, volume 1, chapter 13, pages 253–272. Oxford University Press, Oxford, United Kingdom, 02 2003. URL [https://books.google.fr/books?id=f0I31h\\_G6UkC&pg=PA260&lpg=PA260&dq=financial+incentives+and+trust+experiment&source=bl&ots=-CRrjQeHv\\_&sig=ACfU3U2ID0VJinKgmlUgpFsomoQMD02GnQ&hl=en&sa=X&ved=2ahUKEwiA403C27XqAhVNOBoKHWGJBakQ6AEwDnoECAsQAQ#v=onepage&q=financial%20incentives%20and%20trust%20experiment&f=false](https://books.google.fr/books?id=f0I31h_G6UkC&pg=PA260&lpg=PA260&dq=financial+incentives+and+trust+experiment&source=bl&ots=-CRrjQeHv_&sig=ACfU3U2ID0VJinKgmlUgpFsomoQMD02GnQ&hl=en&sa=X&ved=2ahUKEwiA403C27XqAhVNOBoKHWGJBakQ6AEwDnoECAsQAQ#v=onepage&q=financial%20incentives%20and%20trust%20experiment&f=false).
- [141] Miriam Höddinghaus, Dominik Sondern, and Guido Hertel. The automation of leadership functions: Would people trust decision algorithms? *Computers in Human Behavior*, 116:106635, 2021.
- [142] Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, 2015. doi: 10.1177/0018720814547570. URL <https://doi.org/10.1177/0018720814547570>.
- [143] Kai Holländer, Philipp Wintersberger, and Andreas Butz. Overtrust in external cues of automated vehicles: An experimental investigation. In *Proceedings of the 11th International*

- Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '19, page 211–221, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368841. doi: 10.1145/3342197.3344528. URL <https://doi.org/10.1145/3342197.3344528>.
- [144] John Holmes and John Rempel. Trust in close relationships. *Journal of Personality and Social Psychology*, 49, 07 1985. doi: 10.1037//0022-3514.49.1.95.
- [145] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. Human factors in model interpretability: Industry practices, challenges, and needs. *Proc. ACM Hum.-Comput. Interact.*, 4(1), May 2020. doi: 10.1145/3392878. URL <https://doi.org/10.1145/3392878>.
- [146] Horizon Cancer Center. IBM Watson for oncology demo, March 2015. URL <https://www.youtube.com/watch?v=338CIHLVi7A>.
- [147] Gernot Horstmann and Arvid Herwig. Surprise attracts the eyes and binds the gaze. *Psychonomic Bulletin and Review*, 22(3):743–749, 2015. ISSN 15315320. doi: 10.3758/s13423-014-0723-1.
- [148] Larue Tone Hosmer. Trust: The connecting link between organizational theory and philosophical ethics. *The Academy of Management Review*, 20(2):379–403, 1995. ISSN 03637425. URL <http://www.jstor.org/stable/258851>.
- [149] Hsiao-Ying Huang and Masooda Bashir. Personal influences on dynamic trust formation in Human-Agent interaction. In *Proceedings of the 5th International Conference on Human Agent Interaction*, HAI '17, page 233–243, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450351133. doi: 10.1145/3125739.3125749. URL <https://doi.org/10.1145/3125739.3125749>.
- [150] Lenard Huff and Lane Kelley. Levels of organizational trust in individualist versus collectivist societies: A seven-nation study. *Organization Science*, 14(1):81–90, 2003. ISSN 10477039, 15265455. URL <http://www.jstor.org/stable/3086035>.
- [151] Anna Lena Hunkenschroer and Alexander Kriebitz. Is AI recruiting (un)ethical? a human rights perspective on the use of AI for hiring. *AI and Ethics*, Jul 2022. ISSN 2730-5961. doi: 10.1007/s43681-022-00166-4. URL <https://doi.org/10.1007/s43681-022-00166-4>.
- [152] J. S. Hyde. The gender similarities hypothesis. *Am Psychol*, 60(6):581–592, Sep 2005.

- [153] Institute for Quality and Efficiency in Health Care. What are systematic reviews and meta-analyses?, Jun 2016. URL <https://www.ncbi.nlm.nih.gov/books/NBK390295/>.
- [154] Joi Ito. What the boston school bus schedule can teach us about ai, May 2018. URL <https://www.wired.com/story/joi-ito-ai-and-bus-routes/>.
- [155] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306, 2009. ISSN 00426989. doi: 10.1016/j.visres.2008.09.007. URL <http://dx.doi.org/10.1016/j.visres.2008.09.007>.
- [156] 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. 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*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445385. URL <https://doi.org/10.1145/3411764.3445385>.
- [157] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 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*, FAccT '21, page 624–635, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445923. URL <https://doi.org/10.1145/3442188.3445923>.
- [158] Maurice Jakesch, Zana Bućinca, Saleema Amershi, and Alexandra Olteanu. How different groups prioritize ethical values for responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 310–323, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533097. URL <https://doi.org/10.1145/3531146.3533097>.
- [159] M.Tawfik Jelassi, Karen Williams, and Christine S Fidler. The emerging role of dss: From passive to active. *Decision Support Systems*, 3(4):299–307, 1987. ISSN 0167-9236. doi: [https://doi.org/10.1016/0167-9236\(87\)90101-1](https://doi.org/10.1016/0167-9236(87)90101-1). URL <https://www.sciencedirect.com/science/article/pii/0167923687901011>.
- [160] Theodore Jensen, Yusuf Albayram, Mohammad Maifi Hasan Khan, Md Abdullah Al Fahim, Ross Buck, and Emil Coman.

- The apple does fall far from the tree: User separation of a system from its developers in Human-Automation trust repair. In *Proceedings of the 2019 on Designing Interactive Systems Conference, DIS '19*, page 1071–1082, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450358507. doi: 10.1145/3322276.3322349. URL <https://doi.org/10.1145/3322276.3322349>.
- [161] Theodore Jensen, Mohammad Maifi Hasan Khan, Md Abdullah Al Fahim, and Yusuf Albayram. Trust and anthropomorphism in tandem: The interrelated nature of automated agent appearance and reliability in trustworthiness perceptions. In *Designing Interactive Systems Conference 2021, DIS '21*, page 1470–1480, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384766. doi: 10.1145/3461778.3462102. URL <https://doi.org/10.1145/3461778.3462102>.
- [162] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71, 2000. URL [https://doi.org/10.1207/S15327566IJCE0401\\_04](https://doi.org/10.1207/S15327566IJCE0401_04).
- [163] Weiwei Jiang, Zhanna Sarsenbayeva, Niels van Berkel, Chaofan Wang, Difeng Yu, Jing Wei, Jorge Goncalves, and Vassilis Kostakos. *User Trust in Assisted Decision-Making Using Miniaturized Near-Infrared Spectroscopy*. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450380966. URL <https://doi.org/10.1145/3411764.3445710>.
- [164] Zhuochen Jin, Shuyuan Cui, Shunan Guo, David Gotz, Jimeng Sun, and Nan Cao. Carepre: An intelligent clinical decision assistance system. *ACM Trans. Comput. Healthcare*, 1(1), 2020. doi: 10.1145/3344258. URL <https://doi.org/10.1145/3344258>.
- [165] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, Sep 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0088-2. URL <https://doi.org/10.1038/s42256-019-0088-2>.
- [166] Noel D. Johnson and Alexandra A. Mislin. Trust games: A meta-analysis. *Journal of Economic Psychology*, 32(5):865–889, June 2011. doi: 10.1016/j.joep.2011.05.00. URL <https://ideas.repec.org/a/eee/joepsy/v32y2011i5p865-889.html>.
- [167] Angie M. Johnston, Candice M. Mills, and Asheley R. Landrum. How do children weigh competence and benevolence when deciding whom to trust? *Cognition*, 144:76 – 90, 2015. ISSN

- 0010-0277. doi: <https://doi.org/10.1016/j.cognition.2015.07.015>. URL <http://www.sciencedirect.com/science/article/pii/S001002771530041X>.
- [168] K. G. Jöreskog. A general approach to confirmatory maximum likelihood factor analysis. *ETS Research Bulletin Series*, 1967(2):183–202, 1967. doi: 10.1002/j.2333-8504.1967.tb00991.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.1967.tb00991.x>.
- [169] Daniel Kahneman. *Evaluation by Moments: Past and Future*, chapter 38, pages 693–708. Cambridge University Press & Russell Sage Foundation, New York, USA, 09 2000. doi: 10.1017/CBO9780511803475.039.
- [170] Daniel Kahneman, Paul Slovic, and Amos Tversky. Judgment under uncertainty: Heuristics and biases. *Science*, 185:1124–1131, 1982. ISSN 0036-8075.
- [171] Anil Kalhan. Immigration policing and federalism through the lens of technology, surveillance, and privacy. *Ohio St. LJ*, 74:1105, 2013.
- [172] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. “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*, CHI ’22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517533. URL <https://doi.org/10.1145/3491102.3517533>.
- [173] Maria Kasinidou, Styliani Kleanthous, Pinar Barlas, and Jahna Otterbacher. I agree with the decision, but they didn’t deserve this: Future developers’ perception of fairness in algorithmic decisions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, page 690–700, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445931. URL <https://doi.org/10.1145/3442188.3445931>.
- [174] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists’ use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI



- '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376219. URL <https://doi.org/10.1145/3313831.3376219>.
- [175] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghui Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. Improving Human-AI partnerships in child welfare: Understanding worker practices, challenges, and desires for algorithmic decision support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517439. URL <https://doi.org/10.1145/3491102.3517439>.
- [176] Peter H. Kim, Cecily D. Cooper, Kurt T. Dirks, and Donald L. Ferrin. Repairing trust with individuals vs. groups. *Organizational Behavior and Human Decision Processes*, 120(1):1–14, 2013. doi: 10.1016/j.obhdp.2012.08.0. URL <https://ideas.repec.org/a/eee/jobhdp/v120y2013i1p1-14.html>.
- [177] F. H. Knight. *Risk, Uncertainty, and Profit*. Houghton Mifflin, New York, USA, 1921. URL <https://fraser.stlouisfed.org/files/docs/publications/books/risk/riskuncertaintyprofit.pdf>.
- [178] Melissa A. Koenig and Vikram K. Jaswal. Characterizing children’s expectations about expertise and incompetence: Halo or pitchfork effects? *Child Development*, 82(5):1634–1647, 2011. ISSN 00093920, 14678624. URL <http://www.jstor.org/stable/41289869>.
- [179] Agnieszka Kolasinska, Ivano Lauriola, and Giacomo Quadrio. Do people believe in artificial intelligence? a cross-topic multicultural study. In *Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good*, GoodTechs '19, page 31–36, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362610. doi: 10.1145/3342428.3342667. URL <https://doi.org/10.1145/3342428.3342667>.
- [180] KPMG. Controlling AI: The imperative for transparency and explainability. Technical report, KPMG, June 2019. URL <https://advisory.kpmg.us/articles/2019/controlling-ai.html>.
- [181] Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. Effects of proactive dialogue strategies on Human-Computer trust. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, page 107–116, New York,

- NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368612. doi: 10.1145/3340631.3394840. URL <https://doi.org/10.1145/3340631.3394840>.
- [182] Sari Kujala, Virpi Roto, Kaisa Väänänen, Evangelos Karapanos, and Arto Sinnelä. Ux curve: A method for evaluating long-term user experience. *Interact. Comput.*, 23:473–483, 2011. doi: 10.1016/j.intcom.2011.06.005.
- [183] Philipp Kulms and Stefan Kopp. More human-likeness, more trust? the effect of anthropomorphism on self-reported and behavioral trust in continued and interdependent Human-Agent cooperation. In *Proceedings of Mensch Und Computer 2019*, MuC'19, page 31–42, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450371988. doi: 10.1145/3340764.3340793. URL <https://doi.org/10.1145/3340764.3340793>.
- [184] Olli Lagerspetz. *Trust: The tacit demand*. Library of Ethics and Applied Philosophy. Springer, Dordrecht, Netherlands, December 2010.
- [185] Francesca Lagioia and Giuseppe Contissa. The strange case of Dr Watson : liability implications of AI evidence-based decision support systems in health care. *Eur. J. Leg. Stud.*, 12(2):241–289, 2020.
- [186] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT\* '19, page 29–38, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287590. URL <https://doi.org/10.1145/3287560.3287590>.
- [187] Vivian Lai, Chacha. Chen, Q. Vera Liao, Alison. Smith-Renner, and Chenhao Tan. Towards a science of Human-AI decision making: A survey of empirical studies, 2021, preprint.
- [188] Asheley R. Landrum, Candice M. Mills, and Angie M. Johnston. When do children trust the expert? benevolence information influences children's trust more than expertise. *Developmental Science*, 16(4):622–638, 2013. doi: 10.1111/desc.12059. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/desc.12059>.
- [189] Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J. König, and Nina Grgić-Hlača. “look! it's a computer program!

it's an algorithm! it's ai!": Does terminology affect human perceptions and evaluations of algorithmic decision-making systems? In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517527. URL <https://doi.org/10.1145/3491102.3517527>.

- [190] Stefan Larsson and Fredrik Heintz. Transparency in artificial intelligence. *Internet Pol. Rev.*, 9(2), May 2020.
- [191] Alexander Lascaux. Trust and uncertainty: a critical reassessment. *International Review of Sociology*, 18:1–18, 03 2008. doi: 10.1080/03906700701823613.
- [192] John Lee and Neville Moray. Trust, control strategies and allocation of function in Human-Machine systems. *Ergonomics*, 35(10):1243–1270, 1992. doi: 10.1080/00140139208967392. URL <https://doi.org/10.1080/00140139208967392>.
- [193] John Lee and Katrina See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46:50–80, February 2004. doi: 10.1518/hfes.46.1.50.30392.
- [194] John D. Lee and Neville Moray. Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40(1):153 – 184, 1994. ISSN 1071-5819. doi: <https://doi.org/10.1006/ijhc.1994.1007>. URL <http://www.sciencedirect.com/science/article/pii/S107158198471007X>.
- [195] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. Co-design and evaluation of an intelligent decision support system for stroke rehabilitation assessment. *Proc. ACM Hum.-Comput. Interact.*, 4(2), October 2020. doi: 10.1145/3415227. URL <https://doi.org/10.1145/3415227>.
- [196] Min Kyung Lee and Katherine Rich. Who is included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445570. URL <https://doi.org/10.1145/3411764.3445570>.
- [197] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic me-

- diation. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019. doi: 10.1145/3359284. URL <https://doi.org/10.1145/3359284>.
- [198] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4):611–627, Dec 2018. ISSN 2210-5441. doi: 10.1007/s13347-017-0279-x. URL <https://doi.org/10.1007/s13347-017-0279-x>.
- [199] Roy Lewicki and Chad Brinsfield. Measuring trust beliefs and behaviours. In Fergus Lyon, Guido Möllering, and Mark Saunders, editors, *Handbook of Research Methods on Trust*, chapter 3, pages 29–39. Edward Elgar, Cheltenham, UK; Northampton, MA, USA, 01 2011. doi: 10.4337/9781781009246.00013.
- [200] Roy J. Lewicki, Daniel J. McAllister, and Robert J. Bies. Trust and distrust: New relationships and realities. *The Academy of Management Review*, 23(3):438–458, 1998. ISSN 03637425. URL <http://www.jstor.org/stable/259288>.
- [201] J. David Lewis and Andrew Weigert. Trust as a social reality. *Social Forces*, 63(4):967–985, 1985. ISSN 00377732, 15347605. URL <http://www.jstor.org/stable/2578601>.
- [202] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy AI: From principles to practices. *ACM Comput. Surv.*, aug 2022. ISSN 0360-0300. doi: 10.1145/3555803. URL <https://doi.org/10.1145/3555803>. Just Accepted.
- [203] Ian Li, Jodi Forlizzi, Anind Dey, and Sara Kiesler. My agent as myself or another: Effects on credibility and listening to advice. In *Proceedings of the 2007 Conference on Designing Pleasurable Products and Interfaces*, DPPI '07, page 194–208, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595939425. doi: 10.1145/1314161.1314179. URL <https://doi.org/10.1145/1314161.1314179>.
- [204] Ting-Peng Liang. Recommendation systems for decision support: An editorial introduction. *Decision Support Systems*, 45(3):385–386, 2008. ISSN 0167-9236. doi: <https://doi.org/10.1016/j.dss.2007.05.003>. URL <https://www.sciencedirect.com/science/article/pii/S0167923607000796>. Special Issue Clusters.

- [205] Q. Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–15, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376590. URL <https://doi.org/10.1145/3313831.3376590>.
- [206] Q.Vera Liao and S. Shyam Sundar. Designing for responsible trust in AI systems: A communication perspective. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1257–1268, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533182. URL <https://doi.org/10.1145/3531146.3533182>.
- [207] Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D. Phillips. Calibration of probabilities: The state of the art. In *Decision Making and Change in Human Affairs*, pages 275–324. Springer Netherlands, Netherlands, 1977. doi: 10.1007/978-94-010-1276-8\_19. URL [https://doi.org/10.1007/978-94-010-1276-8\\_19](https://doi.org/10.1007/978-94-010-1276-8_19).
- [208] Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. The conflict between explainable and accountable decision-making algorithms. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2103–2113, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3534628. URL <https://doi.org/10.1145/3531146.3534628>.
- [209] Steven Lockey, Nicole Gillespie, Daniel Holm, and Ida Asadi Someh. A review of trust in Artificial Intelligence: Challenges, vulnerabilities and future directions. 01 2021. doi: 10.24251/HICSS.2021.664.
- [210] Steve Lohr. What ever happened to IBM's Watson?, Jul 2021. URL <https://www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html>.
- [211] Alex London. Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *The Hastings Center Report*, 49:15–21, 02 2019. doi: 10.1002/hast.973.
- [212] James L. Loomis. Communication, the development of trust, and cooperative behavior. *Human Relations*, 12(4):305–315, 1959. doi: 10.1177/001872675901200402. URL <https://doi.org/10.1177/001872675901200402>.

- [213] Zhuoran Lu and Ming Yin. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445562. URL <https://doi.org/10.1145/3411764.3445562>.
- [214] Ewa Luger and Abigail Sellen. “like having a really bad pa”: The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 5286–5297, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858288. URL <https://doi.org/10.1145/2858036.2858288>.
- [215] Niklas Luhmann. *Trust and Power*. Wiley, Chichester, Toronto, 1 edition, 1979. ISBN 0471997587.
- [216] Nikolas Luhmann. Familiarity, confidence, trust: Problems and alternatives. In Diego Gambetta, editor, *Trust: Making and Breaking Cooperative Relations*, pages 94–107. Basil Blackwell, Oxford, United Kingdom, January 2000.
- [217] Fergus Lyon, Guido Möllering, and Mark Saunders. *Handbook of Research Methods on Trust: Second Edition*. Edward Elgar Publishing, Cheltenham, United Kingdom, 01 2015. doi: 10.4337/9781782547419.
- [218] Henrietta Lyons, Eduardo Velloso, and Tim Miller. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), apr 2021. doi: 10.1145/3449180. URL <https://doi.org/10.1145/3449180>.
- [219] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. What’s the appeal? perceptions of review processes for algorithmic decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517606. URL <https://doi.org/10.1145/3491102.3517606>.
- [220] Stefanie M. Faas, Johannes Kraus, Alexander Schoenhals, and Martin Baumann. Calibrating pedestrians’ trust in automated vehicles: Does an intent display in an external hmi support trust calibration and safe crossing behavior? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21,

- New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445738. URL <https://doi.org/10.1145/3411764.3445738>.
- [221] P. Madhavan and D. A. Wiegmann. Similarities and differences between Human-Human and Human-Automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, 8(4): 277–301, 2007. doi: 10.1080/14639220500337708. URL <https://doi.org/10.1080/14639220500337708>.
- [222] Maria A. Madsen and Shirley Gregor. Measuring Human-Computer trust. In *Proceedings of the 11 th Australasian Conference on Information Systems*, pages 6–8, Brisbane, Australia, 2000. Australasian Conference on Information Systems (ACIS).
- [223] Danielle Magaldi and Matthew Berler. *Semi-structured Interviews*, pages 4825–4830. Springer International Publishing, Cham, 2020. ISBN 978-3-319-24612-3. doi: 10.1007/978-3-319-24612-3\_857. URL [https://doi.org/10.1007/978-3-319-24612-3\\_857](https://doi.org/10.1007/978-3-319-24612-3_857).
- [224] Rebecca Maguire, Phil Maguire, and Mark T. Keane. Making Sense of Surprise: An Investigation of the Factors Influencing Surprise Judgments. *Journal of Experimental Psychology: Learning Memory and Cognition*, 37(1):176–186, 2011. ISSN 02787393. doi: 10.1037/a0021609.
- [225] Mora Maldonado, Ewan Dunbar, and Emmanuel Chemla. Mouse tracking as a window into decision making. *Behavior Research Methods*, 51(3):1085–1101, Jun 2019. ISSN 1554-3528. doi: 10.3758/s13428-018-01194-x. URL <https://doi.org/10.3758/s13428-018-01194-x>.
- [226] C. Mantzavinos. Hermeneutics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Stanford, CA, USA, spring 2020 edition, 2020.
- [227] Ivana Markova and Alex Gillespie, editors. *Trust and Distrust*. Advances in Cultural Psychology. Information Age Publishing, Greenwich, CT, March 2008.
- [228] Ronald Marshall. Building trust early: The influence of first and second order expectations on trust in international channels of distribution. *International Business Review*, 12:421–443, 08 2003. doi: 10.1016/S0969-5931(03)00037-4.
- [229] Juliette Mattioli, Gabriel Pedroza, Souhail Khalfaoui, and Bertrand Leroy. Combining data-driven and knowledge-based ai

- paradigms for engineering ai-based safety-critical systems, Feb 2022. URL <https://hal.archives-ouvertes.fr/hal-03622260/document>.
- [230] Steffen Maurer, Rainer Erbach, Issam Kraiem, Susanne Kuhnert, Petra Grimm, and Enrico Rukzio. 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*, page 341–350, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359467. doi: 10.1145/3239060.3239078. URL <https://doi.org/10.1145/3239060.3239078>.
- [231] James H. Mayer, Roger C.;Davis. The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Business and Industrial Personnel*, 84(1):123–136, 1999. doi: 10.1037/0021-9010.84.1.123.
- [232] Roger C. Mayer, James H. Davis, and F. David Schoorman. An integrative model of organizational trust. *The Academy of Management Review*, 20(3):709–734, 1995. ISSN 03637425. URL <http://www.jstor.org/stable/258792>.
- [233] Daniel J. McAllister. Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *The Academy of Management Journal*, 38(1):24–59, 1995. ISSN 00014273. URL <http://www.jstor.org/stable/256727>.
- [234] Daniel J. McAllister. Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *The Academy of Management Journal*, 38(1):24–59, 1995. ISSN 00014273. URL <http://www.jstor.org/stable/256727>.
- [235] Bill McEvily and Marco Tortoriello. Measuring trust in organizational research: Review and recommendations. *Journal of Trust Research*, 1(1):23–63, 2011. doi: 10.1080/21515581.2011.552424. URL <https://doi.org/10.1080/21515581.2011.552424>.
- [236] D. McKnight and Norman Chervany. Trust and distrust definitions: One bite at a time. In R. Falcone, M. Singh, and Y. H. Tan, editors, *Trust in Cyber-societies: Integrating the Human and Artificial Perspectives*, pages 27–54. Springer, Heidelberg, Germany, 01 2001. doi: 10.1007/3-540-45547-7\_3.
- [237] D. Harrison McKnight, Larry L. Cummings, and Norman L. Chervany. Initial trust formation in new organizational relationships. *Academy of Management Review*, 23(3):473–490, 1998. doi:



- 10.5465/amr.1998.926622. URL <https://doi.org/10.5465/amr.1998.926622>.
- [238] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, 13(3):334–359, 2002. doi: 10.1287/isre.13.3.334.81. URL <https://pubsonline.informs.org/doi/abs/10.1287/isre.13.3.334.81>.
- [239] Stephanie M. Merritt. Affective processes in Human–Automation interactions. *Human Factors*, 53(4): 356–370, 2011. doi: 10.1177/0018720811411912. URL <https://doi.org/10.1177/0018720811411912>.
- [240] Joachim Meyer and John D. Lee. Trust, reliance, and compliance. In John D. Lee and Alex Kirlik, editors, *The Oxford Handbook of Cognitive Engineering*, pages 1–29. Oxford University Press, Oxford, UK, 05 2013. ISBN 9780199757183. URL <https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199757183.001.0001/oxfordhb-9780199757183-e-6>.
- [241] Wulf Uwe Meyer, Michael Niepel, Udo Rudolph, and Achim Schautzwohl. An Experimental Analysis of Surprise. *Cognition and Emotion*, 5(4):295–311, 1991. ISSN 14640600. doi: 10.1080/02699939108411042.
- [242] Microsoft. Responsible bots: 10 guidelines for developers of conversational ai. Technical report, Microsoft, USA, November 2018. URL <https://www.microsoft.com/en-us/research/publication/responsible-bots/>.
- [243] Barbara Misztal. *Trust in Modern Societies: The Search for the Bases of Social Order*. Library of Ethics and Applied Philosophy. Polity Press, Cambridge, England, 04 1996.
- [244] Shaswat Mohanty, Anirudh Vijay, and Nandagopan Gopakumar. StockBot: Using LSTMs to predict stock prices, 2022. URL <https://arxiv.org/abs/2207.06605>.
- [245] Michael Moore. *Confirmatory factor analysis*, pages 361–379. The Guilford Press, NY, USA, 07 2012. ISBN 9781606230770.
- [246] Drew M. Morris, Jason M. Erno, and June J. Pilcher. Electrodermal response and automation trust during simulated self-driving car use. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 61(1):1759–1762, 2017. doi: 10.1177/1541931213601921. URL <https://doi.org/10.1177/1541931213601921>.

- [247] B.M. Muir. *Operators' Trust in and Use of Automatic Controllers in a Supervisory Process Control Task*. Canadian theses on microfiche. University of Toronto, Toronto, Canada, 1989. ISBN 9780315510142. URL <https://books.google.fr/books?id=T94NSwAACAAJ>.
- [248] Bonnie M. Muir. Trust in automation: I. theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37:1905–1922, 1994. ISSN 1540-4560. doi: 10.1080/00140139408964957. URL <https://doi.org/10.1080/00140139408964957>.
- [249] Lea S. Müller, Sarah M. Meeßen, Meinald T. Thielsch, Christoph Nohe, Dennis M. Riehle, and Guido Hertel. *Do Not Disturb! Trust in Decision Support Systems Improves Work Outcomes under Certain Conditions*, page 229–237. MUC '20. Association for Computing Machinery, New York, NY, USA, 2020. ISBN 9781450375405. URL <https://doi.org/10.1145/3404983.3405515>.
- [250] Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. This thing called fairness: Disciplinary confusion realizing a value in technology. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019. doi: 10.1145/3359221. URL <https://doi.org/10.1145/3359221>.
- [251] Robert Münscher and Torsten M. Kühlmann. Using critical incident technique in trust research. In Fergus Lyon, Guido Möllering, and Mark Saunders, editors, *Handbook of Research Methods on Trust*, chapter 14, pages 161–172. Edward Elgar, Cheltenham, UK; Northampton, MA, USA, 01 2011.
- [252] Michael Naef and Jürgen Schupp. Measuring Trust: Experiments and Surveys in Contrast and Combination. IZA Discussion Papers 4087, Institute of Labor Economics (IZA), March 2009. URL <https://ideas.repec.org/p/iza/izadps/dp4087.html>.
- [253] Manisha Natarajan and Matthew Gombolay. Effects of anthropomorphism and accountability on trust in Human-Robot interaction. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, HRI '20*, page 33–42, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367462. doi: 10.1145/3319502.3374839. URL <https://doi.org/10.1145/3319502.3374839>.
- [254] National Commission on Informatics and Liberty. Self-assessment guide for Artificial Intelligence (AI)

- systems, Sep 2022. URL <https://www.cnil.fr/en/self-assessment-guide-artificial-intelligence-ai-systems>.
- [255] Lilynbsp; Hay Newman. A new google+ blunder exposed data from 52.5 million users, Dec 2018. URL <https://www.wired.com/story/google-plus-bug-52-million-users-data-exposed/>.
- [256] Michael Niepel, Udo Rudolph, Achim Schiltzwohl, and Wulf Uwe Meyer. Temporal Characteristics of the Surprise Reaction Induced by Schema-discrepant Visual and Auditory Events. *Cognition and Emotion*, 8(5):433–452, 1994. ISSN 14640600. doi: 10.1080/02699939408408951.
- [257] Domen Novak. *Engineering Issues in Physiological Computing*, pages 17–38. 03 2014. ISBN 978-1-4471-6391-6. doi: 10.1007/978-1-4471-6392-3\_2.
- [258] Tobias O. Nyumba, Kerrie Wilson, Christina J. Derrick, and Nibedita Mukherjee. The use of focus group discussion methodology: Insights from two decades of application in conservation. *Methods in Ecology and Evolution*, 9(1):20–32, 2018. doi: 10.1111/2041-210X.12860. URL <https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12860>.
- [259] James A. O'Brien and George Marakas. Decision support systems. In Fergus Lyon, Guido Möllering, and Mark Saunders, editors, *Management Information Systems*, chapter 10, pages 123–139. MITSource, Boston, MA: McGraw-Hill, Inc., 2007.
- [260] James A. O'Brien and George Marakas. Decision support systems. In Fergus Lyon, Guido Möllering, and Mark Saunders, editors, *Management Information Systems*, chapter 11, pages 141–155. MITSource, Boston, MA: McGraw-Hill, Inc., 2007.
- [261] Kenya Freeman Oduor and Christopher S. Campbell. Deciding when to trust automation in a policy-based city management game: Policity. In *Proceedings of the 2007 Symposium on Computer Human Interaction for the Management of Information Technology*, CHIMIT '07, page 2–es, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595936356. doi: 10.1145/1234772.1234775. URL <https://doi.org/10.1145/1234772.1234775>.
- [262] Institute of Business Ethics. Business ethics and artificial intelligence. Technical report, Internet Society, London, UK, January 2018. URL <https://www.ibe.org.uk/resource/>

[ibe-briefing-58-business-ethics-and-artificial-intelligence-pdf.html](#).

- [263] Claus Offe. *How can we trust our fellow citizens?*, chapter 3, pages 42–87. Cambridge UP, Cambridge, United Kingdom, 01 1999. URL [https://www.researchgate.net/publication/246388496\\_How\\_can\\_we\\_trust\\_our\\_fellow\\_citizens](https://www.researchgate.net/publication/246388496_How_can_we_trust_our_fellow_citizens).
- [264] Roobina Ohanian. Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attractiveness. *Journal of Advertising*, 19(3):39–52, oct 1990. doi: 10.1080/00913367.1990.10673191. URL <https://doi.org/10.1080%2F00913367.1990.10673191>.
- [265] Chitu Okoli. A guide to conducting a standalone systematic literature review. *Commun. Assoc. Inf. Syst.*, 37, 2015.
- [266] Chinasa T. Okolo, Srujana Kamath, Nicola Dell, and Aditya Vashistha. “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*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445420. URL <https://doi.org/10.1145/3411764.3445420>.
- [267] Linda Onnasch. Crossing the boundaries of automation—function allocation and reliability. *International Journal of Human-Computer Studies*, 76:12–21, 2015. ISSN 1071-5819. doi: <https://doi.org/10.1016/j.ijhcs.2014.12.004>. URL <https://www.sciencedirect.com/science/article/pii/S1071581914001670>.
- [268] Jeroen Ooge, Shotallo Kato, and Katrien Verbert. Explaining recommendations in e-learning: Effects on adolescents' trust. In *27th International Conference on Intelligent User Interfaces*, IUI '22, page 93–105, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391443. doi: 10.1145/3490099.3511140. URL <https://doi.org/10.1145/3490099.3511140>.
- [269] Rachel O'Dwyer. Algorithms are making the same mistakes as humans assessing credit scores. *Retrieved April, 17:2019*, 2018.
- [270] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. Understanding the impact of explanations on advice-taking: A user study for AI-based clinical decision support systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573.

doi: 10.1145/3491102.3502104. URL <https://doi.org/10.1145/3491102.3502104>.

- [271] Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. It's complicated: The relationship between user trust, model accuracy and explanations in ai. *ACM Trans. Comput.-Hum. Interact.*, 29(4), mar 2022. ISSN 1073-0516. doi: 10.1145/3495013. URL <https://doi.org/10.1145/3495013>.
- [272] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. Human-AI interaction in human resource management: Understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21*, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445304. URL <https://doi.org/10.1145/3411764.3445304>.
- [273] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. A slow algorithm improves users' assessments of the algorithm's accuracy. In *Proceedings of the 2019 Conference on Computer Supported Cooperative Work, volume 3 of CSCW '19*, New York, NY, USA, 2019. Association for Computing Machinery. doi: 10.1145/3359204. URL <https://doi.org/10.1145/3359204>.
- [274] Dhaval Parmar, Stefán Ólafsson, Dina Utami, Prasanth Murali, and Timothy Bickmore. *Navigating the Combinatorics of Virtual Agent Design Space to Maximize Persuasion*, page 1010–1018. AAMAS '20. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2020. ISBN 9781450375184.
- [275] Samir Passi and Steven J. Jackson. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), November 2018. doi: 10.1145/3274405. URL <https://doi.org/10.1145/3274405>.
- [276] P. M. Patrzyk, D. Link, and J. N. Marewski. Human-like machines: Transparency and comprehensibility. *Behav Brain Sci*, 40: e276, 01 2017.
- [277] P. Ivan Pavlov. Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex. *Annals of neurosciences*, 17(3):136–141, Jul 2010. ISSN 0972-7531. doi: 10.5214/ans.0972-7531.1017309. URL <https://doi.org/10.5214/ans.0972-7531.1017309>.

- [278] Carl J. Pearson, Allaire K. Welk, William A. Boettcher, Roger C. Mayer, Sean Streck, Joseph M. Simons-Rudolph, and Christopher B. Mayhorn. Differences in trust between human and automated decision aids. In *Proceedings of the Symposium and Bootcamp on the Science of Security, HotSos '16*, page 95–98, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342773. doi: 10.1145/2898375.2898385. URL <https://doi.org/10.1145/2898375.2898385>.
- [279] Brandon S. Perelman, Arthur W. Evans III, and Kristin E. Schaefer. Where do you think you're going? characterizing spatial mental models from planned routes. *J. Hum.-Robot Interact.*, 9(4), May 2020. doi: 10.1145/3385008. URL <https://doi.org/10.1145/3385008>.
- [280] Patricia Perry. Concept analysis: Confidence/self-confidence. *Nursing Forum*, 46(4):218–230, 2011. doi: 10.1111/j.1744-6198.2011.00230.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1744-6198.2011.00230.x>.
- [281] J. Paul Peter. Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research*, 16(1):6–17, 1979. ISSN 00222437. URL <http://www.jstor.org/stable/3150868>.
- [282] Gjalt-Jorn Peters. The alpha and the omega of scale reliability and validity: Why and how to abandon conbach's alpha and the route towards more comprehensive assessment od scale quality. *Euro Health Psychologist*, 16:56–69, 01 2014.
- [283] Gloria Phillips-Wren. Intelligent decision support systems. In Michael Doumpos and Evangelos Grigoroudis, editors, *Multicriteria decision aid and artificial intelligence*, chapter 2, pages 25–44. John Wiley & Sons, Nashville, TN, 2013.
- [284] Jonathan A. Plucker. Exploratory and confirmatory factor analysis in gifted education: Examples with self-concept data. *Journal for the Education of the Gifted*, 27(1):20–35, 2003. doi: 10.1177/016235320302700103. URL <https://doi.org/10.1177/016235320302700103>.
- [285] J. Potter and D. Edwards. *Discourse Analysis*, pages 419–425. Macmillan Education UK, London, 1996. ISBN 978-1-349-24483-6. doi: 10.1007/978-1-349-24483-6\_63. URL [https://doi.org/10.1007/978-1-349-24483-6\\_63](https://doi.org/10.1007/978-1-349-24483-6_63).
- [286] Daniel Power. *Decision Support Systems: Concepts and Resources for Managers*. 03 2002. ISBN 156720497X.

- [287] Pearl Pu and Li Chen. Trust building with explanation interfaces. In *Proceedings of the 11th International Conference on Intelligent User Interfaces, IUI '06*, page 93–100, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595932879. doi: 10.1145/1111449.1111475. URL <https://doi.org/10.1145/1111449.1111475>.
- [288] David V. Pynadath, Ning Wang, Ericka Rovira, and Michael J. Barnes. Clustering behavior to recognize subjective beliefs in Human-Agent teams. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '18*, page 1495–1503, New York, NY, USA, 2018. International Foundation for Autonomous Agents and Multiagent Systems.
- [289] Bako Rajaonah, Françoise Anceaux, and Fabrice Vienne. Study of driver trust during cooperation with adaptive cruise control. *Le travail humain*, 69(2):99–127, 2006. URL <https://doi.org/10.3917/th.692.0099>.
- [290] Bako Rajaonah, Françoise Anceaux, Nicolas Tricot, and Marie-Pierre Pacaux-Lemoine. Trust, cognitive control, and control: The case of drivers using an auto-adaptive cruise control. In *Proceedings of the 13th European Conference on Cognitive Ergonomics: Trust and Control in Complex Socio-Technical Systems, ECCE '06*, page 17–24, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 9783906509235. doi: 10.1145/1274892.1274896. URL <https://doi.org/10.1145/1274892.1274896>.
- [291] Divya Ramesh, Vaishnav Kameswaran, Ding Wang, and Nithya Sambasivan. 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, FAccT '22*, page 1917–1928, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533237. URL <https://doi.org/10.1145/3531146.3533237>.
- [292] Amy Rechkemmer and Ming Yin. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22*, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3501967. URL <https://doi.org/10.1145/3491102.3501967>.
- [293] B Reeves and C I Nass. *The media equation: How people treat computers, television, and new media like real people and places*. Center

for the Study of Language and Information. Cambridge University Press, 1996.

- [294] Samantha Reig, Selena Norman, Cecilia G. Morales, Samadrita Das, Aaron Steinfeld, and Jodi Forlizzi. A field study of pedestrians and autonomous vehicles. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI '18*, page 198–209, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359467. doi: 10.1145/3239060.3239064. URL <https://doi.org/10.1145/3239060.3239064>.
- [295] Samantha Reig, Selena Norman, Cecilia G. Morales, Samadrita Das, Aaron Steinfeld, and Jodi Forlizzi. A field study of pedestrians and autonomous vehicles. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI '18*, page 198–209, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359467. doi: 10.1145/3239060.3239064. URL <https://doi.org/10.1145/3239060.3239064>.
- [296] Rainer Reisenzein. Exploring the strength of association between the components of emotion syndromes: The case of surprise. *Cognition and Emotion*, 14(1):1–38, 2000. ISSN 02699931. doi: 10.1080/026999300378978.
- [297] Rainer Reisenzein and Markus Studtmann. On the Expression and Experience of Surprise: No Evidence for Facial Feedback, but Evidence for a Reverse Self-Inference Effect. *Emotion*, 7(3): 612–627, 2007. ISSN 15283542. doi: 10.1037/1528-3542.7.3.612.
- [298] Robin M. Richter, Maria Jose Valladares, and Steven C. Sutherland. Effects of the source of advice and decision task on decisions to request expert advice. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19*, page 469–475, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362726. doi: 10.1145/3301275.3302279. URL <https://doi.org/10.1145/3301275.3302279>.
- [299] Vincent Robbmond, Oana Inel, and Ujwal Gadiraju. Understanding the role of explanation modality in AI-assisted decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '22*, page 223–233, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392075. doi: 10.1145/3503252.3531311. URL <https://doi.org/10.1145/3503252.3531311>.



- [300] Loo Robert. A caveat on using single-item versus multiple-item scales. *Journal of Managerial Psychology*, 17(1):68–75, Jan 2002. ISSN 0268-3946. doi: 10.1108/02683940210415933. URL <https://doi.org/10.1108/02683940210415933>.
- [301] Jr Robert B. Lount, Chen-Bo Zhong, Niro Sivanathan, and J. Keith Murnighan. Getting off on the wrong foot: The timing of a breach and the restoration of trust. *Personality and Social Psychology Bulletin*, 34(12):1601–1612, 2008. doi: 10.1177/0146167208324512. URL <https://doi.org/10.1177/0146167208324512>. PMID: 19050335.
- [302] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. Overtrust of robots in emergency evacuation scenarios. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction, HRI '16*, page 101–108, New York, NY, USA, 2016. IEEE Press. ISBN 9781467383707.
- [303] Mark A. Robinson. Using multi-item psychometric scales for research and practice in human resource management. *Human Resource Management*, 57(3):739–750, 2018. doi: 10.1002/hrm.21852. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/hrm.21852>.
- [304] P Robinson. Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics*, 22(1):27–57, 03 2001. ISSN 0142-6001. doi: 10.1093/applin/22.1.27. URL <https://doi.org/10.1093/applin/22.1.27>.
- [305] John T. Roscoe. *Fundamental research statistics for the behavioral sciences*. New York Holt, Rinehart and Winston, 1969. ISBN 978-0-03-079135-2. URL <http://openlibrary.org/books/OL5685768M>.
- [306] Elizabeth Rosenzweig. Usability testing. In Elizabeth Rosenzweig, editor, *Successful User Experience: Strategies and Roadmaps*, chapter 7, pages 131 – 154. Morgan Kaufmann, Boston, MA, USA, 2015. ISBN 978-0-12-800985-7. doi: <https://doi.org/10.1016/B978-0-12-800985-7.00007-7>. URL <http://www.sciencedirect.com/science/article/pii/B9780128009857000077>.
- [307] Casey Ross and Ike Swetlitz. IBM pitched its Watson supercomputer as a revolution in cancer care. It’s nowhere close, Sep 2017. URL <https://www.statnews.com/2017/09/05/watson-ibm-cancer/>.

- [308] Jennifer M. Ross. *Moderators of trust and reliance across multiple decision aids*. PhD thesis, Department of Psychology in the College of Sciences at the University of Central Florida, 2008.
- [309] Julian B. Rotter. Interpersonal trust, trustworthiness, and gullibility. *American Psychologist*, 35(1):1–7, 1980. doi: 10.1037/0003-066X.35.1.1. URL <https://doi.org/10.1037/0003-066X.35.1.1>.
- [310] Denise Rousseau, Sim Sitkin, Ronald Burt, and Colin Camerer. Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, 23, July 1998. doi: 10.5465/AMR.1998.926617.
- [311] Royal College of Physicians. Artificial intelligence (ai) in health. Technical report, Royal College of Physicians, London, UK, September 2018. URL <https://www.rcplondon.ac.uk/projects/outputs/artificial-intelligence-ai-health>.
- [312] Nicole Salomons, Michael van der Linden, Sarah Strohkorb Sebo, and Brian Scassellati. Humans conform to robots: Disambiguating trust, truth, and conformity. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI '18*, page 187–195, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450349536. doi: 10.1145/3171221.3171282. URL <https://doi.org/10.1145/3171221.3171282>.
- [313] S Samoili, Cobo M Lopez, E Gomez Gutierrez, G De Prato, F Martinez-Plumed, and B Delipetrev. AI watch. defining artificial intelligence. (KJ-NA-30117-EN-N (online)), 2020. ISSN 1831-9424 (online). doi: 10.2760/382730(online).
- [314] Willem E. Saris and Irmtraud N. Gallhofer. *Criteria for the Quality of Survey Measures*, pages 173–217. John Wiley & Sons, Ltd, Hoboken, New Jersey, USA, 2007. ISBN 9780470165195. doi: 10.1002/9780470165195.ch9. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470165195.ch9>.
- [315] Shoko Sasayama. Is a ‘complex’ task really complex? validating the assumption of cognitive task complexity. *The Modern Language Journal*, 100(1):231–254, 2016. doi: <https://doi.org/10.1111/modl.12313>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/modl.12313>.
- [316] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 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), oct 2021. doi: 10.1145/3476089. URL <https://doi.org/10.1145/3476089>.
- [317] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES '19*, page 99–106, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314248. URL <https://doi.org/10.1145/3306618.3314248>.
- [318] Kristin E. Schaefer. *The Perception And Measurement Of Human-Robot Trust*. PhD thesis, Department of Psychology in the College of Sciences at the University of Central Florida, 2013.
- [319] Kristin E. Schaefer, Deborah R. Billings, James L. Szalma, Jeffrey K. Adams, Tracy Sanders, Jessie Y. C. Chen, and Peter A. Hancock. A meta-analysis of factors influencing the development of trust in automation: Implications for Human-Robot interaction. Technical report, U.S. Army Research Laboratory, July 2014.
- [320] Kristin E. Schaefer, Jessie Y. C. Chen, James L. Szalma, and P. A. Hancock. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, 58(3):377–400, 2016. doi: 10.1177/0018720816634228. URL <https://doi.org/10.1177/0018720816634228>. PMID: 27005902.
- [321] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. I can do better than your AI: Expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19*, page 240–251, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362726. doi: 10.1145/3301275.3302308. URL <https://doi.org/10.1145/3301275.3302308>.
- [322] Matthew U Scherer. Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *SSRN Electron. J.*, 2015.
- [323] Hanna Schneider, Julia Wayrauther, Mariam Hassib, and Andreas Butz. Communicating uncertainty in fertility prognosis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, page 1–11, New York, NY, USA, 2019.

- Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300391. URL <https://doi.org/10.1145/3290605.3300391>.
- [324] Jakob Schoeffer, Yvette Machowski, and Niklas Kuehl. A study on fairness and trust perceptions in automated decision making. *arXiv preprint arXiv:2103.04757*, 2021.
- [325] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. “there is not enough information”: On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, page 1616–1628, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533218. URL <https://doi.org/10.1145/3531146.3533218>.
- [326] F. David Schoorman, Roger C. Mayer, and James H. Davis. An integrative model of organizational trust: Past, present, and future. *Academy of Management Review*, 32(2):344–354, 2007. doi: 10.5465/amr.2007.24348410. URL [10.5465/amr.2007.24348410](https://doi.org/10.5465/amr.2007.24348410).
- [327] Ian A Scott, Stacy M Carter, and Enrico Coiera. Exploring stakeholder attitudes towards AI in clinical practice. *BMJ Health & Care Informatics*, 28(1), 2021. doi: 10.1136/bmjhci-2021-100450. URL <https://informatics.bmj.com/content/28/1/e100450>.
- [328] Kristen M. Scott, Sonja Mei Wang, Milagros Miceli, Pieter Delobelle, Karolina Sztandar-Sztanderska, and Bettina Berendt. Algorithmic tools in public employment services: Towards a jobseeker-centric perspective. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, page 2138–2148, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3534631. URL <https://doi.org/10.1145/3531146.3534631>.
- [329] Haeseung Seo, Aiping Xiong, and Dongwon Lee. Trust it or not: Effects of machine-learning warnings in helping individuals mitigate misinformation. In *Proceedings of the 10th ACM Conference on Web Science, WebSci '19*, page 265–274, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362023. doi: 10.1145/3292522.3326012. URL <https://doi.org/10.1145/3292522.3326012>.
- [330] Younho Seong, Ann M Bisantz, and Ann M Bisantz. Judgment and trust in conjunction with automated decision aids: A theoretical model and empirical investigation. In *Proceedings of the*

*Human Factors and Ergonomics Society Annual Meeting*, volume 46, pages 423–427. SAGE Publications Sage CA: Los Angeles, CA, 2002.

- [331] Fred Shaffer and J. P. Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in public health*, 5:258–258, Sep 2017. ISSN 2296-2565. doi: 10.3389/fpubh.2017.00258. URL <https://doi.org/10.3389/fpubh.2017.00258>. 29034226[pmid].
- [332] Ameneh Shamekhi, Q. Vera Liao, Dakuo Wang, Rachel K. E. Bellamy, and Thomas Erickson. Face value? exploring the effects of embodiment for a group facilitation agent. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173965. URL <https://doi.org/10.1145/3173574.3173965>.
- [333] Klaas Sijtsma. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1):107, Dec 2008. ISSN 1860-0980. doi: 10.1007/s11336-008-9101-0. URL <https://doi.org/10.1007/s11336-008-9101-0>.
- [334] Jeffrey A. Simpson. Psychological foundations of trust. *Current Directions in Psychological Science*, 16(5):264–268, 2007. doi: 10.1111/j.1467-8721.2007.00517.x. URL <https://doi.org/10.1111/j.1467-8721.2007.00517.x>.
- [335] Sim B. Sitkin and Nancy L. Roth. Explaining the limited effectiveness of legalistic "remedies" for trust/ distrust. *Organization Science*, 4(3):367–392, 1993. ISSN 10477039, 15265455. URL <http://www.jstor.org/stable/2634950>.
- [336] Janet A. Sniezek and Lyn M. Van Swol. Trust, confidence, and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes*, 84(2):288–307, 2001. ISSN 0749-5978. doi: <https://doi.org/10.1006/obhd.2000.2926>. URL <https://www.sciencedirect.com/science/article/pii/S0749597800929261>.
- [337] Internet Society. Artificial intelligence and machine learning: policy paper. Technical report, Internet Society, Reston, Virginia, United States, April 2017. URL <https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/>.
- [338] Special Interest Group on Artificial Intelligence. Dutch artificial intelligence manifesto. Technical report, ICT Research Platform Nederland, The Netherlands, September 2019.

- URL <http://ii.tudelft.nl/bnvki/wp-content/uploads/2018/09/Dutch-AI-Manifesto.pdf>.
- [339] Donna Spencer and Todd Warfel. Card sorting: a definitive guide, April 2004. URL <https://boxesandarrows.com/card-sorting-a-definitive-guide/>.
- [340] Logan Stapleton, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldechova, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, page 1162–1177, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533177. URL <https://doi.org/10.1145/3531146.3533177>.
- [341] STATNews. How does Watson for oncology work?, Sep 2017. URL <https://www.youtube.com/watch?v=UpFHNGF4F8o>.
- [342] Edward Stohr and Sivakumar Viswanathan. Recommendation systems: Decision support for the information economy. In *Emerging Information Technologies: Improving Decisions, Cooperation, and Infrastructure*, pages 21–44. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States, May 1999.
- [343] Nicole Sultanum, Michael Brudno, Daniel Wigdor, and Fanny Chevalier. More text please! understanding and supporting the use of visualization for clinical text overview. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18*, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173996. URL <https://doi.org/10.1145/3173574.3173996>.
- [344] Haoye Sun, Willem J. M. I. Verbeke, Rumen Pozharliev, Richard P. Bagozzi, Fabio Babiloni, and Lei Wang. Framing a trust game as a power game greatly affects interbrain synchronicity between trustor and trustee. *Social Neuroscience*, 14(6): 635–648, December 2019. doi: 10.1080/17470919.2019.1566171.
- [345] Harini Suresh, Natalie Lao, and Ilaria Liccardi. Misplaced trust: Measuring the interference of machine learning in human decision-making. In *12th ACM Conference on Web Science, WebSci '20*, page 315–324, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379892. doi: 10.1145/3394231.3397922. URL <https://doi.org/10.1145/3394231.3397922>.

- [346] Steven C. Sutherland, Casper Hartevelde, and Michael E. Young. 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*, page 2535–2544, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450331456. doi: 10.1145/2702123.2702609. URL <https://doi.org/10.1145/2702123.2702609>.
- [347] Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.
- [348] Henri Tajfel and John Turner. *The Social Identity Theory of Intergroup Behavior: Key Readings*, pages 276–293. 01 2004. ISBN 9780203505984. doi: 10.4324/9780203505984-16.
- [349] Karl Halvor Teigen and Gideon Keren. Surprises: Low probabilities or high contrasts? *Cognition*, 87(2):55–71, 2003. ISSN 00100277. doi: 10.1016/S0010-0277(02)00201-9.
- [350] Hiroyuki Tokushige, Takuji Narumi, Sayaka Ono, Yoshitaka Fuwamoto, Tomohiro Tanikawa, and Michitaka Hirose. Trust lengthens decision time on unexpected recommendations in Human-Agent interaction. In *Proceedings of the 5th International Conference on Human Agent Interaction, HAI '17*, page 245–252, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450351133. doi: 10.1145/3125739.3125751. URL <https://doi.org/10.1145/3125739.3125751>.
- [351] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. Second chance for a first impression? trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '21*, page 77–87, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383660. doi: 10.1145/3450613.3456817. URL <https://doi.org/10.1145/3450613.3456817>.
- [352] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 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, CHI '22*, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517732. URL <https://doi.org/10.1145/3491102.3517732>.

- [353] Ilaria Torre, Jeremy Goslin, Laurence White, and Debora Zanatto. Trust in artificial voices: A “congruency effect” of first impressions and behavioural experience. In *Proceedings of the Technology, Mind, and Society, TechMindSociety '18*, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450354202. doi: 10.1145/3183654.3183691. URL <https://doi.org/10.1145/3183654.3183691>.
- [354] Ilaria Torre, Emma Carrigan, Rachel McDonnell, Katarina Domijan, Killian McCabe, and Naomi Harte. The effect of multimodal emotional expression and agent appearance on trust in Human-Agent interaction. In *Motion, Interaction and Games, MIG '19*, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369947. doi: 10.1145/3359566.3360065. URL <https://doi.org/10.1145/3359566.3360065>.
- [355] Italo Trizano-Hermosilla and Jesús M. Alvarado. Best alternatives to Cronbach’s alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology*, 7:769, 2016. ISSN 1664-1078. doi: 10.3389/fpsyg.2016.00769. URL <https://www.frontiersin.org/article/10.3389/fpsyg.2016.00769>.
- [356] Efraim Turban and Jay E Aronson. *Decision Support Systems and Intelligent Systems*. Prentice Hall, Philadelphia, PA, 5 edition, November 1997.
- [357] UNI Global Union. 10 principles for ethical ai. Technical report, UNI Global Union, Nyon, Switzerland, December 2017. URL <http://www.thefutureworldofwork.org/opinions/10-principles-for-ethical-ai/>.
- [358] Hanneke Hooft van Huysduynen, Jacques Terken, and Berry Eggen. Why disable the autopilot? In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI '18*, page 247–257, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359467. doi: 10.1145/3239060.3239063. URL <https://doi.org/10.1145/3239060.3239063>.
- [359] Peter-Paul van Maanen, Francien Wisse, Jurriaan van Diggelen, and Robbert-Jan Beun. 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*, page 280–287, New York, NY, USA, 2011. IEEE Computer Soci-



- ety. ISBN 9780769545134. doi: 10.1109/WI-IAT.2011.117. URL <https://doi.org/10.1109/WI-IAT.2011.117>.
- [360] Lyn M Van Swol and Janet A Sniezek. Factors affecting the acceptance of expert advice. *Br J Soc Psychol*, 44(Pt 3):443–461, September 2005.
- [361] Michael Veale, Max Van Kleek, and Reuben Binns. 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, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3174014. URL <https://doi.org/10.1145/3173574.3174014>.
- [362] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. How to evaluate trust in AI-assisted decision making? a survey of empirical methodologies. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), October 2021. doi: 10.1145/3476068. URL <https://doi.org/10.1145/3476068>.
- [363] Cédric Villani, Yann Bonnet, Bertrand Rondepierre, et al. *For a meaningful artificial intelligence: Towards a French and European strategy*. Conseil national du numérique, France, 2018.
- [364] Rajan Vohra and Nripendra Das. Intelligent decision support systems for admission management in higher education institutes. *International Journal of Artificial Intelligence and Applications*, 2:63–70, 10 2011. doi: 10.5121/ijai.2011.2406.
- [365] Rudolf von Sinner. Trust and convivência. *The Ecumenical Review*, 57(3):322–341, 2005. doi: 10.1111/j.1758-6623.2005.tb00554.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1758-6623.2005.tb00554.x>.
- [366] J.F. Voss and T.A. Post. On the solving of ill-structured problems. In Michelene T H Chi, Robert Glaser, and Marshall J Farr, editors, *The Nature of Expertise*, chapter 9, pages 261–286. Psychology Press, New York, 1988.
- [367] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–15, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300831. URL <https://doi.org/10.1145/3290605.3300831>.

- [368] Lin Wang, Pei-Luen Patrick Rau, Vanessa Evers, Benjamin Krisper Robinson, and Pamela Hinds. When in rome: The role of culture & context in adherence to robot recommendations. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction, HRI '10*, page 359–366, New York, NY, USA, 2010. IEEE Press. ISBN 9781424448937.
- [369] M. Wang, A. Hussein, R. F. Rojas, K. Shafi, and H. A. Abbass. EEG-based neural correlates of trust in Human-Autonomy interaction. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 350–357, Bangalore, India, 2018. IEEE.
- [370] Ning Wang, David V. Pynadath, and Susan G. Hill. Trust calibration within a Human-Robot team: Comparing automatically generated explanations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction, HRI '16*, page 109–116, New York, NY, USA, 2016. IEEE Press. ISBN 9781467383707.
- [371] Ning Wang, David V. Pynadath, and Susan G. Hill. The impact of pomdp-generated explanations on trust and performance in Human-Robot teams. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, AAMAS '16*, page 997–1005, New York, NY, USA, 2016. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450342391.
- [372] Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in AI-assisted decision-making. In *26th International Conference on Intelligent User Interfaces, IUI '21*, page 318–328, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380171. doi: 10.1145/3397481.3450650. URL <https://doi.org/10.1145/3397481.3450650>.
- [373] Xinru Wang and Ming Yin. Effects of explanations in AI-assisted decision making: Principles and comparisons. *ACM Trans. Interact. Intell. Syst.*, feb 2022. ISSN 2160-6455. doi: 10.1145/3519266. URL <https://doi.org/10.1145/3519266>. Just Accepted.
- [374] Xiuxin Wang and Xiufang Du. Why does advice discounting occur? the combined roles of confidence and trust. *Front Psychol*, 9:2381, November 2018.
- [375] Eva K. Wendt, Bengt Fridlund, and Evy Lidell. Trust and confirmation in a gynecologic examination situation: a critical incident technique analysis. *Acta obstetricia et gynecologica Scandinavica*, 83 12:1208–1215, 2004.

- [376] Lawrence R. Wheeless and Janis Grotz. The measurement of trust and its relationship to self-disclosure. *Human Communication Research*, 3(3):250–257, 1977. doi: 10.1111/j.1468-2958.1977.tb00523.x. URL <https://doi.org/10.1111/j.1468-2958.1977.tb00523.x>.
- [377] T. Whelan. Social presence in multi-user virtual environments : A review and measurement framework for organizational research. 2008.
- [378] of Science and Technology Policy White House Office. American AI initiative: Year one annual report. Technical report, White House Office of Science and Technology Policy, Brussels, Belgium, February 2020. URL <https://www.whitehouse.gov/ai/>.
- [379] Lisa S Whiting. Semi-structured interviews: guidance for novice researchers. *Nurs Stand*, 22(23):35–40, 2008.
- [380] Philipp Wintersberger, Tamara von Sawitzky, Anna-Katharina Frison, and Andreas Riener. Traffic augmentation as a means to increase trust in automated driving systems. In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter, CHIItaly '17*, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450352376. doi: 10.1145/3125571.3125600. URL <https://doi.org/10.1145/3125571.3125600>.
- [381] Jacob O. Wobbrock and Julie A. Kientz. Research contributions in human-computer interaction. *Interactions*, 23(3):38–44, apr 2016. ISSN 1072-5520. doi: 10.1145/2907069. URL <https://doi.org/10.1145/2907069>.
- [382] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18*, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3174230. URL <https://doi.org/10.1145/3173574.3174230>.
- [383] Jun Xiao, John Stasko, and Richard Catrambone. 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*, page 1293–1302, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595935939. doi: 10.1145/1240624.1240820. URL <https://doi.org/10.1145/1240624.1240820>.

- [384] Yaqi Xie, Indu P Bodala, Desmond C. Ong, David Hsu, and Harold Soh. Robot capability and intention in trust-based decisions across tasks. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction, HRI '19*, page 39–47, New York, NY, USA, 2019. IEEE Press. ISBN 9781538685556.
- [385] Doris Xin, Eva Yiwei Wu, Doris Jung-Lin Lee, Niloufar Salehi, and Aditya Parameswaran. Whither automl? understanding the role of automation in machine learning workflows. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21*, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445306. URL <https://doi.org/10.1145/3411764.3445306>.
- [386] Rodrigo Yañez Gallardo and Sandra Valenzuela-Suazo. Critical incidents of trust erosion in leadership of head nurses. *Revista Latino-Americana de Enfermagem*, 20:143 – 150, 02 2012. ISSN 0104-1169. URL [http://www.scielo.br/scielo.php?script=sci\\$.\\_\\$arttext&pid=S0104-11692012000100019&nrm=iso](http://www.scielo.br/scielo.php?script=sci$._$arttext&pid=S0104-11692012000100019&nrm=iso).
- [387] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI '20*, page 189–201, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371186. doi: 10.1145/3377325.3377480. URL <https://doi.org/10.1145/3377325.3377480>.
- [388] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F. Antaki. Investigating the heart pump implant decision process: Opportunities for decision support tools to help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16*, page 4477–4488, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858373. URL <https://doi.org/10.1145/2858036.2858373>.
- [389] X. Jessie Yang, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah. Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI '17*, page 408–416, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450343367. doi: 10.1145/2909824.3020230. URL <https://doi.org/10.1145/2909824.3020230>.

- [390] Ilan Yaniv and Eli Kleinberger. Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, 83(2):260–281, 2000. ISSN 0749-5978. doi: <https://doi.org/10.1006/obhd.2000.2909>. URL <https://www.sciencedirect.com/science/article/pii/S0749597800929091>.
- [391] J. Frank Yates. *Judgment and decision making*. Judgment and decision making. Prentice-Hall, Inc, Englewood Cliffs, NJ, US, 1990. ISBN 0-13-511726-7 (Hardcover).
- [392] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 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, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300509. URL <https://doi.org/10.1145/3290605.3300509>.
- [393] Louise C. Young and Gerald S. Albaum. *Developing a measure of trust in retail relationships : a direct selling application*. School of Marketing, University of Technology of Sydney, Sydney Broadway, N.S.W, Australia, 2002.
- [394] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. *Keeping Designers in the Loop: Communicating Inherent Algorithmic Trade-Offs Across Multiple Objectives*, page 1245–1257. DIS '20. Association for Computing Machinery, New York, NY, USA, 2020. ISBN 9781450369749. URL <https://doi.org/10.1145/3357236.3395528>.
- [395] Kun Yu, Shlomo Berkovsky, Dan Conway, Ronnie Taib, Jianlong Zhou, and Fang Chen. Trust and reliance based on system accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, UMAP '16, page 223–227, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450343688. doi: 10.1145/2930238.2930290. URL <https://doi.org/10.1145/2930238.2930290>.
- [396] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, page 307–317, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450343480. doi: 10.1145/3025171.3025219. URL <https://doi.org/10.1145/3025171.3025219>.

- [397] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. Do I trust my machine teammate? an investigation from perception to decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19*, page 460–468, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362726. doi: 10.1145/3301275.3302277. URL <https://doi.org/10.1145/3301275.3302277>.
- [398] Kun-Hsing Yu and Isaac S Kohane. Framing the challenges of artificial intelligence in medicine. *BMJ Quality & Safety*, 28(3): 238–241, 2019. ISSN 2044-5415. doi: 10.1136/bmjqs-2018-008551. URL <https://qualitysafety.bmj.com/content/28/3/238>.
- [399] L. Yu and Y. Li. Artificial Intelligence Decision-Making Transparency and Employees' Trust: The Parallel Multiple Mediating Effect of Effectiveness and Discomfort. *Behav Sci (Basel)*, 12(5), Apr 2022.
- [400] Beste F. Yuksel, Penny Collisson, and Mary Czerwinski. Brains or beauty: How to engender trust in user-agent interactions. *ACM Trans. Internet Technol.*, 17(1), 2017. doi: 10.1145/2998572. URL <https://doi.org/10.1145/2998572>.
- [401] Mireia Yurrita, Dave Murray-Rust, Agathe Balayn, and Alessandro Bozzon. Towards a multi-stakeholder value-based assessment framework for algorithmic systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, page 535–563, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533118. URL <https://doi.org/10.1145/3531146.3533118>.
- [402] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 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*, page 295–305, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372852. URL <https://doi.org/10.1145/3351095.3372852>.
- [403] Jurong Zheng, Jens K. Roehrich, and Michael A. Lewis. The dynamics of contractual and relational governance: Evidence from long-term public–private procurement arrangements. *Journal of Purchasing and Supply Management*, 14(1):43–54, 2008. ISSN 1478-4092. doi: <https://doi.org/10.1016/j.pursup.2008.01.004>. URL <https://www.sciencedirect.com/science/article/pii/S1478409208000058>. Practice Makes Perfect: Special Issue of Best Papers of the 16th Annual IPSERA Conference 2007.

- [404] Qingxiao Zheng, Yiliu Tang, Yiren Liu, Weizi Liu, and Yun Huang. UX research on conversational Human-AI interaction: A literature review of the ACM digital library. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3501855. URL <https://doi.org/10.1145/3491102.3501855>.