International Journal on Software Tools for Technology Transfer
https://doi.org/10.1007/s10009-023-00699-x

GENERAL

Special issue: Dagstuhl 23492

Dynamic Calibration of Trust and Trustworthiness in AI-Enabled Systems

Magnus Liebherr · Ellen Enkel · Effie L-C. Law · Mohammad Reza Mousavi · Matteo Sammartino · Philipp Sieberg

the date of receipt and acceptance should be inserted later

Abstract
Trust is a multi-faceted phenomenon traditionally studied in human relations and more recently in human-machine interactions. In the context of AI-enabled systems, trust is about the belief of the user that in a given scenario the system is going to be helpful and safe. The system-side counterpart to trust is trustworthiness. When trust and trustworthiness are aligned with each other, there is calibrated trust. Trust, trustworthiness, and calibrated trust are all dynamic phenomena, evolving throughout the history and evolution of user beliefs, systems, and their interaction.
In this paper, we review the basic concepts of trust, trustworthiness and calibrated trust and provide definitions for them. We discuss their various metrics used in the literature, and the causes that may affect their dynamics, particularly in the context of AI-enabled systems. We discuss the implications of the discussed concepts for various types of stakeholders and suggest some challenges for future research.

Keywords Trust, Trustworthiness, Calibrated Trust, AI-Enabled Systems

1 Introduction

Trust [MDS95, Gam88, Har06, MTP00] is a user-centred multi-faceted notion, which has been traditionally studied in social sciences in human-human interactions. With the advent of human-machine interactions and more recently autonomous and AI-enabled systems trust has gained renewed interest [GAB+21, HB15, LS04]. This has been particularly driven by the increasing deployment of AI in high-stakes domains such as healthcare, finance, and autonomous driving, where trust directly impacts adoption, interaction quality, and safety. As AI systems take on more autonomous and decision-critical roles, ensuring appropriate levels of trust has become a central concern. In this context, many recent studies have focussed on defining [LS04] and measuring [ALCL23, BNR19, KdVW+21, RF23] trust in human-machine interactions. Users' trust in machines is not always justified [DVPJ+20]: under-trust in a system that is not trustworthy can lead to reduced benefit for users, and over-trust can lead to harm. Calibrated trust [DVPJ+20] happens when trust of users and trustworthiness of systems [NW10, KURD22, Jon12, Car23] meet. All these notions are parametric to users (personas and groups) and scenarios in which the system is used.

In this paper, we provide an overview of existing results, open challenges, potential solutions regarding trust and trustworthiness in data-driven AI-enabled systems. A common feature of many such systems is that there is typically little specification available and hence, establishing trustworthiness and calibrating trust are significant challenges. Our target audience is researchers, particularly those starting a research career in Trust(worthiness) for AI, and practitioners using such systems. We present the definitions and results in such a

Ellen Enkel
University of Duisburg-Essen
E-mail: ellen.enkel@uni-due.de

Effie L. Law
Durham University
E-mail: lai-chong.law@durham.ac.uk

Magnus Liebherr
University of Duisburg-Essen
E-mail: magnus.liebherr@uni-due.de

Mohammad Reza Mousavi
King's College Lonon
E-mail: mohammad.mousavi@kcl.ac.uk

Matteo Sammartino
Royal Holloway, University of London
E-mail: matteo.sammartino@rhul.ac.uk
Philipp Sieberg
Schotte Automotive Gmbh & Co.KG
E-mail: philipp.sieberg@schotteautomotive.de

way that they are also accessible for practitioners and policy makers. For this audience, we present rigorous definitions and an overview of available metrics and causes; we also reflect on the possible applications and consequences, and open issues for future research.

The contributions of this paper can be summarized as follows:

- An overview of definitions, assessment models, and factors of trust, trustworthiness and their calibration.
- A framework, based on the provided definitions, accounting for the dynamic interaction among trust, trustworthiness and their interactions.
- A reflection on the implication of these issues. We derive recommendations for system developers, policy makers, users and companies on how to deal with the levels of trust/trustworthiness.

The remainder of this paper is organized as follows. In Section 2, we review the different facets and definitions of trust, trustworthiness and calibrated trust and produce our definitions. In Section 3, we present the state of the art in assessing these concepts and the open challenges involved in their assessment. In Section 4, we discuss the factors that may have a positive or negative causal influence on trust and trustworthiness (and hence, their calibration) and discuss the dynamics of trust and trustworthiness in this context. Section 5 discusses the ongoing work embodying the implication of this research, e.g., in terms of standards and guidelines. Section 6 concludes the paper and presents an overview of the future road-map. We provide a glossary of terms as an appendix to help our readers navigate through the wealth of terminology and concepts.

2 Definitions

2.1 Trust

Defining trust. Trust stands as a foundational element within interpersonal relationships and is generally defined as "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 other party" [MDS95]. Mayer et al.'s (1995) model expands on this definition by identifying three key dimensions that shape the trustor's perception of trustworthiness, namely ability, benevolence, and integrity. This perception of trustworthiness influences trust, which in turn impacts the trustor's willingness to take risks. The extent to which trust translates into risk-taking is moderated by perceived risk, meaning that trust only becomes relevant in situations where vulnerability exists. Additionally, trustor's propensity—a general tendency to trust others based on personality or past experiences—acts as an antecedent, shaping initial trust levels before direct interactions with the trustee occur [MDS95]. In addition to considering risk as a moderator, other perspectives on trust place significantly greater emphasis on perceived risk as a central component. [Gam88, Har06, MTP00]. The perception of risk has to do with the epistemic state of each user. Although risk is inherent in trust, it does not negate the value of trust. Rather, it is the willingness to trust despite potential risks that makes trust a powerful and transformative force in interpersonal relationships, group dynamics, civic engagement and society as a whole [Rob16]. However, trust is not limited to the interpersonal domain. It also describes the way people interact with technology. In particular, the field of trust in AI has garnered significant attention in recent times [GAB+21, YW22, KKOT23].

To our current understanding, it is evident that interpersonal trust and trust in AI-based systems entail reliance on an external entity; however, the underpinnings and dynamics of trust exhibit notable distinctions between human interactions and engagement with AI-based systems, particularly concerning the dynamics of trust calibration with respect to the trustworthiness of the AI-enabled system. Currently, multiple competing definitions of trust in AI-based systems exist, and consensus on a specific definition has not been reached. In accordance with previous work in the area of trust and new technologies/AI applications [GAB+21, HB15, LS04], we define trust in AI-based systems as follows.

> **Trust** is defined as the user's (u) epistemic state that the system (s) is going to behave (b) as expected in a scenario (r) characterized by a certain level of risk and uncertainty.

Much like in interpersonal relationships, trust is about the state of people's minds and it significantly influences people's readiness to depend on AI-based systems. Below we list some of the factors contributing to trust, with a focus on trust in AI-based systems.

Cognitive trust vs. Emotional trust. Trust can be grounded in both affect, involving emotional aspects, and rational thinking, encompassing cognitive evaluations [McA95]. [MDS95] hinted at the potential emergence of an emotional connection in a trusting relationship. However, their model did not fully incorporate emotion. Instead, the influential factors have been construed through a cognitive lens, wherein trust is logically assessed by the trusting party based on trusting beliefs [LMT15, MCTC11]. Cognitive trust is rooted in logical reasoning and evidence, involving assessments that individuals make about the reliability, competence, and predictability of another party [SMD07]. This evaluation extends beyond human interactions, recognizing trust as a key factor in human relations with non-humans [HB15, PR97]. Researchers studying cognitive trust in AI describe it by the users' willingness to act on factual information or

advice and their perception of the technology's helpfulness, competence, or usefulness [HBS+11, HB15, LS04]. It's important to acknowledge that cognitive trust might also be influenced by factors like emotions and mood, and that emotional trust may mediate users' decisions to adopt a specific technology [KB06]. Although we know that emotions are the primary determinant of trusting behavior, the concept of emotional trust is under-explored and deserve further research in the context of human interactions with technology [HB15, LS04]. Generally, emotional trust is described by the sense of security individuals experience within emotional connections with others [GAB+21, JG05, LW85]. In the utilization of AI-based systems, tangibility has proven to be of significant importance, as well as the empathy and compassion exhibited by the AI towards users [GW20, LYQ+22, YW22]. [GW20], through their review of literature on human trust in AI, identified three forms of AI representations (robotic, virtual, and embedded) and their corresponding paths of cognitive and emotional trust. Overall, due to the information asymmetry between an AI system and its users, building cognitive trust proves to be challenging without a thorough understanding of how the system operates. Consequently, users often rely on emotional trust [KK22].

Theoretical models for trust in AI. A particularly relevant extension of trust models to human-machine interactions was introduced by [LM92] and later refined by [LS04]. Their work translated Mayer et al.'s [MDS95] interpersonal trust dimensions—ability, benevolence, and integrity—into the automation context, framing trust as a user's perception of a system's performance (ability), intent (benevolence), and predictability (integrity/reliability). This adaptation laid the groundwork for understanding trust in AI, as it emphasizes how users develop expectations about system behavior based on cognitive evaluations of these dimensions. Building on these insights, trust has been recognized as a pivotal factor shaping users' behavioral intentions towards the utilization of AI-based systems, consequently impacting adoption and diffusion in diverse domains [GAB+21, YW22]. In a recent review, encompassing diverse industrial areas, trust is elucidated as a substantial and affirmative predictor of intention, willingness, and usage behavior related to AI [KKOT23]. The importance of trust becomes also evident in both classic and modified iterations of key theories on new technology adoption (also applicable to the domain of AI-based systems), by including trust indirectly (Theory of Reasoned Action; Theory of Planned Behavior; Technology Acceptance Model) or directly (modified versions of the Technology Acceptance Model) within the models [Ajz91, Dav85, FA77, GKS03]. These frameworks also underscore the dynamic nature of trust in technology adoption processes, a concept that has been explicitly explored in additional theories. Dynamic trust models have been reported in various contexts such as the Internet of things [WCZ+20] algorithmic advice [DO22, HB15] or B2C Cross-Border E-Commerce [DWC22]. The prevailing consensus consolidates trust as an emergent phenomenon rather than a static state [Woo09]. In Explainable AI (XAI) a notional view of how trust could morph has been proposed, showing that users start cautiously or skeptically but become more trusting when provided with an effective initial explanation, leading to a state of justified trust; however, continued interaction with the XAI system may lead to instances of automation surprise [HMKL21].

2.2 Trustworthiness

Defining trustworthiness Trustworthiness has been used in various fields of social sciences such as Philosophy [Jon12, KS23, Car23], Psychology [Rot80, CDv+10], Political Economy [Har02], and outside social sciences in Neuroscience [DPJ12] and Computer Science and Engineering [NW10, KURD22]. It is noteworthy that trustworthiness is an under-researched concept, particularly when compared with trust [Har02]. The semantics of trustworthiness varies across different fields and there are even debates about how to define it within a field of study. For example, Hardin [Har02] defines trustworthiness as the "capacity to judge one's interest in fulfilling the trust". In much of the social sciences literature, trustworthiness is defined in terms of perceived trust. An important aspect of many earlier definitions of trustworthiness is its contextualization in terms of the subject (the assessor of trustworthiness), the object (the trustee), and the task (the scenario in which the interaction is supposed to happen) [Jon12, Car23]. We would like to provide a definition of trustworthiness that can be formalized and can provide a basis for quantitative measurement on the system and its properties. This will further enable quantifying the dynamic calibration between trust (using the metrics surveyed before) and trustworthiness (through the forthcoming definition). It is worth noting that trustworthiness is dynamic and subject to change during the system life cycle due to the continuous changes (e.g., due to changes and repair and also due to ageing of sensors and hardware). As a consequence, we define trustworthiness to be a property of the object of trust, given a specification of requirements provided by the user, as well as the scenario in which trustworthiness is defined. It is therefore inspired by the earlier contextualization of trustworthiness in earlier definitions [Jon12, KS23].

Trustworthiness: We call a system (s) trustworthy in a scenario (r), when it satisfies the user's (u) requirements (req_s) for the scenario. A system is trustworthy when it is trustworthy in all scenarios within its operational design domain.

Principles of Trustworthiness. Several organizations have proposed principles for trustworthy AI, including major tech companies such as Google, OpenAI, Microsoft, and IBM. These principles coalesce around the key requirements set out by the EU Commission in 2019 [(AI19]:

1. Human agency and oversight: AI should support human agency, safeguard fundamental rights of users and include oversight mechanisms.
2. Technical robustness and safety: AI systems must exhibit resilience, safety and security. Additionally, they should prioritize accuracy, reliability, and reproducibility. Contingency plans should be in place.
3. Privacy and data governance: AI systems should guarantee privacy and data protection, ensure the quality and integrity of data, and support data access protocols.
4. Transparency: AI systems should be properly documented to allow for traceability, with technical and decision-making processes that are explainable. Users should be made aware of the systems' capabilities and limitations.
5. Diversity, non-discrimination and fairness: AI systems should avoid unfair bias, be accessible, and engage pertinent stakeholders at every stage of their life cycle.
6. Societal and environmental sustainability: AI systems should contribute to the well-being of individuals, prioritize sustainability and environmental friendliness, and undergo thorough assessments for their societal impact.
7. Accountability: AI systems should incorporate mechanisms to guarantee responsibility and accountability. The algorithms, data, and design processes should facilitate auditability, with established measures for redress.

The EU definition is not the final word on defining trustworthiness, but it gives a comprehensive template for assessing trustworthiness on an abstract level. The principles put forward by the EU can further be used as metrics to characterize the requirements for different scenarios in different application domains. Moreover, such requirements can be turned into formal contracts between AI system providers and users to enforce the trustworthiness of the system [JMMG21].

There are published critiques of this definition [SC25, Kus24, SARH+21], some of which note that the EU guidelines are primarily based on ethical rather than legal principles, making their implementation challenging in specific contexts such as corporate governance, where harmonization with relevant regulations and governance principles is necessary [HP21]. Although other frameworks exist, there is considerable consensus on the key principles (see [Inn23, pag.13] for a comparison table).

2.3 Calibrated trust

Defining calibration. Trust and trustworthiness are supposed to mirror each other: trust is the belief of the user about the system matching their expectations, while trustworthiness is the property of the system that satisfies the requirements aligned with user needs, which, in turn, capture users' expectations. Trust is a personal and user-specific notion, while trustworthiness, in our definition, is objective and generalizable to a group, community or society [Sif12, SH23]. Calibrated trust, defined below, is when for a given user, the levels of trust and trustworthiness match, i.e., trust is placed on a system that is trustworthy in the given scenario:

> **Calibrated trust:** Trust of a user (u) in a system (s) in a scenario (r) is called calibrated when it matches the actual level of trustworthiness of the system for this scenario. Trust in a system is calibrated when it is calibrated for all users and all scenarios.

Calibration of trust is a dynamic process; typically the user builds an instantaneous trust when confronted by an AI-based system. Instantaneous trust may be informed by prior knowledge, by the guarantor of the system and the guarantees provided, as well as other cognitive and emotional factors briefly discussed in Section 2.1 and elaborated in Section 4. Both instantaneous trust and the dynamic changes of trust can lead to overtrust and distrust in AI-based systems. Below we briefly discuss the issues of overtrust and distrust. See Figure 1 [DVPJ+20] for a depiction of how calibration is determined by actual and perceived trustworthiness.

Under-trust/distrust. Under-trust in AI technology can have multifaceted consequences, manifesting in various ways [DVPJ+20]. Firstly, the failure to harness the full potential of Trustworthy Autonomous Systems (TAS) is a significant drawback. Underestimating or hesitating to rely on the capabilities of AI systems leads to missed opportunities for enhanced efficiency and effectiveness. Additionally, under-trust can result in suboptimal solutions or performance [DVPJ+20]. Doubting the capabilities of AI technology may lead to conservative decision-making, preventing the realization of optimal outcomes that could be achieved through a more confident integration of AI tools. Communication breakdowns can also arise as a consequence of under-trust. When humans lack confidence in AI systems, effective communication channels may be hindered, impeding the seamless collaboration between human operators and autonomous technologies [LRFI23, OY20]. Moreover, under-trust contributes to an increased workload on both human operators and TAS [FLC+23]. Human operators may feel compelled to overcompensate for perceived deficiencies in AI, leading to heavier cognitive load. Simultaneously, AI systems may not be utilized to their full potential, resulting in a suboptimal distribution of tasks and responsibilities. Lastly, under-trust can lead to a state of disuse or micromanagement. Human operators, due to a lack of trust in AI, may either refrain from

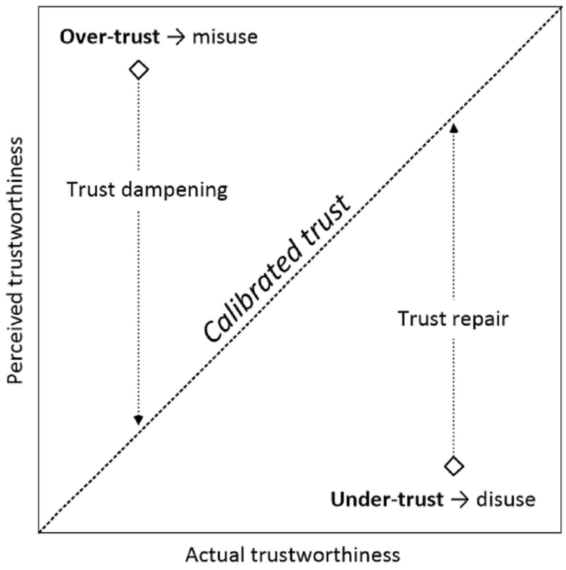

Fig. 1 Trust calibration [DVPJ+20] p. 462

using the technology altogether or micromanage its functions, undermining the purpose of employing autonomous systems for efficiency and autonomy [DVPJ+20, FLC+23].

Over-trust. Over-trust in AI presents serious challenges, often culminating in harmful outcomes. This unwarranted confidence in AI systems can be particularly perilous when the technology is not infallible or fully capable of performing a given task [ADBD+21]. The danger lies in relying excessively on AI without recognizing its limitations, potentially leading to critical errors or misinterpretations [DVPJ+20, GW20]. In instances where over-trust prevails, there is a concerning lack of guidance and control mechanisms. Users may neglect to provide adequate oversight or interventions, assuming that AI is entirely reliable [UBD21]. This absence of vigilant monitoring can exacerbate the risks associated with over-trust, as it fails to account for unexpected variables or evolving scenarios. Consequently, over-trust not only amplifies the potential for disastrous outcomes but also highlights the need for robust guidance and control measures to ensure the responsible and effective deployment of AI technologies in various contexts [GW20, UBD21].

Trust in AI-powered agents undergoes continuous calibration over time, particularly in prolonged collaborations between humans and agents [DVPJ+20]. User trust adapts dynamically based on the performance of the agent. Trust increases with positive interaction (matching the mental model of the user with the interaction with the AI based system) and trust can decrease when the expectation from the user is not matched with by the system (e.g. because the system has the wrong mental model of the user, overestimating the user's trust). In [WKM23a] comprehensive review of trust calibrations for automated systems, distinct patterns related to trust fluctuations and recovery, or resilience, were identified. Notably, some studies revealed that users were more responsive to declines in the agent's reliability than to improvements. The decline in trust was more pronounced than the increase, despite comparable changes in reliability in either direction [WRZ01]. Users were inclined to trust the agent even more after recovering from minor and brief reliability lapses, but not for significant and short ones [LS19]. Additionally, the nature of the error leading to reliability changes could significantly impact trust calibration; for example, false alarms were observed to diminish trust more than misses [CMH21]. However, these intriguing phenomena have primarily been analysed from cognitive rather than affective processes (i.e. cognitive rather than emotional trust), as [WKM23a] reported that only a few studies have explored the emotion-trust relationship in the context of automated systems [FKJ+21].

2.4 Summary

In this section, we provided three formal definitions for the three concepts of trust, trustworthiness, and calibrated trust. All three definitions are parameterized in terms of a user (u), a system (s), and a scenario (r). Trust is about the user's state (e.g., beliefs and emotions) concerning the system in a given scenario. Trustworthiness is about systems' properties with respect to users' requirements in a given scenario. Calibration is about the conformance of the earlier two notions. For each of the three concepts, we provided some aspects of the concept discussed in earlier research to motivate and contextualize our definitions. Table 1 summarizes the concepts, as well as the definitions, and aspects of each concept discussed in this section.

3 Assessment

3.1 Measures of trust

The varied conceptualizations and subsequent operationalizations of trust have given rise to a multitude of trust measures [ALCL23, BNR19, KdVW+21, RF23]. They can be categorized as either direct or indirect methods. Direct assessment involves self-report/subjective measures, while indirect evaluation utilizes behavioral/performance-based as well as physiological measures.

Self-report/subjective measures. These measures entail individuals providing information regarding their thoughts, feelings, attitudes, or behaviors. They rely on individuals' self-disclosure and are typically obtained through surveys, questionnaires, interviews, or other means where individuals report on their own experiences or perceptions.

Table 1 Summary of concepts, definitions, and discussed topics for: trust, trustworthiness, and calibrated trust

Concept	Definition	Discussed topics
Trust	We define trust in AI-based systems as the users (u) epistemic state that the system (s) is going to behave (b) as expected in a scenario (r) characterized by a certain level of risk and uncertainty.	Cognitive and emotional trust, Expectations, Theories of trust.
Trustworthiness	We call a system (s) trustworthy in a scenario (r), when it satisfies the user's (u) requirements for the scenario. A system is trustworthy when it is trustworthy in all scenarios within its operational design domain.	Agency and oversight, Privacy and data governance, Robustness and safety, Fairness, Sustainability, Accountability.
Calibrated trust	Trust of a user (u) in a system (s) in a scenario (r) is called calibrated when it matches the actual level of trustworthiness of the system for this scenario. Trust in a system is calibrated when it is calibrated for all users and all scenarios.	Calibrated trust, instantaneous trust, overtrust, distrust.

Self-report measures are most commonly used to assess human trust due to the ease of use and implementation in tasks or contexts [JBD00, Kör19, MG00, Sch16, WPTL+20]. We provide a summary of such measures in Table 2.

Utilizing a text analysis approach, [ALCL23] identified differences and similarities across the most used trust questionnaires, subsequently providing guidelines for its selection. In the context of AI, the following two specific questionnaires are most commonly used:

- Trust in Artificial Intelligent Agents Questionnaire [AGC21]; and
- TXAI Questionnaire [HMKL23].

In addition to employing specific questionnaires, general trust assessments or modified versions adapted from other contexts are also utilized to measure trust in AI. Furthermore, studies on trust in AI-based systems frequently incorporate custom measures, such as non-validated scales or individual items. In Table 2, we provide a concise overview of the self-report/subjective measures of trust.

Behavioral/performance-based measures. These measures assess trust based on observable actions, responses, or behaviors exhibited by individuals in a given context. These measures focus on tangible and external manifestations of trust, providing insights into how individuals interact with or rely on a particular entity, system, or situation. Examples of behavioral measures in the context of trust might include:

- Compliance and Agreement Rate [VDVM13]
- Decision Time [YCC17]
- Reliance [PM10]
- Response Time [KPB18]

The determination of the parameter used as a behavior-related measure for evaluating trust depends on the system or AI technologies considered. For instance, in the integration of AI-based technologies in driving, reaction time serves as a valuable metric for gaining insight into trust in the system [KPB18]; however, it offers limited insights or may not be relevant to conversational AI applications. Unlike self-report measures that rely on individuals' subjective perceptions, behavioral measures offer an objective and tangible way to assess trust by examining actual behaviors and actions in real-world or simulated scenarios. These measures are particularly valuable for understanding trust dynamics in situations where individuals may not accurately or completely express their trust through self-reporting.

Physiological measures. These measures capture biological responses within the human body as indicators of trust in a given situation. Certain physiological reactions such as changes in heart rate, skin conductance or brain activity, which reflect emotional and cognitive states, are used to draw conclusions about changes in trust. Typical measures are:

- Heart rate variability [WN14] - Electrodermal activity or galvanic skin response [KZCM15]
- Neural measures [GPC+16, HBB+14, JDL19]
- Eye gaze tracking [GKH+15, HLVK16]

Physiological measures enhance our understanding of underlying processes associated with trust (see Table 3). In order to draw conclusions, it is necessary to link physiological measures to context. Just as behavioral measures are selected, the choice of a metric is contingent upon the specific AI technology under consideration, taking into account factors such as the environment (e.g., natural environment), the situation (e.g., any kind of movement included), and individual aspects (e.g., clinical history).

To date, several systematic literature reviews on human trust in technologies provide further insights into related measures [ASL20, ALCL23, BKH+22, BNR19, KdVW+21, RF23]; integrating such reviews can chart a landscape of existing methods and tools for measuring trust in AI-based systems. In the future, it is crucial to consolidate the substantial variability in metrics employed for evaluating human trust in AI and to establish consensus on the adoption of standardized methods [GW20]. Furthermore, when employing trust measures in the realm of new technologies and

Table 2 Summary of self-reported / subjective measure of trust.

Questionnaire	Number of items	Subscales	Likert-scale	Example	Context	Psychometric characteristics
Trust Between People and Automation [JBD00]	12	-	1-7	"The system is deceptive."	Automated system	Reliability ↑ [JBD00, RF23] Reliability of scale structure ↓ [WPTL+20]
Human-computer trust questionnaire [MG00]	25	Perceived Reliability, Perceived Technical Competence, Perceived Understandability, Faith, Personal Attachment	1-7	"The system performs reliably."	subjective measure of "cognition based" and "affective-based" trust	Reliability, Validity ↑ [MG00], Reliability ↑, Construct validity ↓ [DKKT17]
TiA scale [Kör19]	19	Reliability/Competence, Understanding/Predictability, Familiarity, Intention of Developers, Propensity to Trust, Trust in Automation	1-5	"The system works reliably."		Reliability, validity ↑ [Kör19, RF23]
Trust Perception Scale for Human-Robot Interactions [Sch16]	40 (and 14 item sub-scale)	-	0%-100%	"What percent of the time will this robot be responsible?"	Human-Robot-Interaction	Reliability ↑, validity ↓ [RF23]
Propensity to Trust Scale [Mer11]	6	-	1-5	"I usually trust machines until there is a reason not to."	Propensity to trust, not specific technology	Reliability ↑ [Mer11]
HCTM [GSL19]	12	-	1-5	"I believe that there could be negative consequences when using (—)."		Reliability ↑ [GSL19]
[MCK02]	16	Trusting beliefs , Trusting intention	1-7	"I can always rely on LegalAdvice.com in a tough legal situation."		Reliability ↑ [MCK02]
TAIA (Trust in Artificial Intelligent Agents; [AGC21])	12	Predictability, Consistency, Utility, Faith, Dependability, Understanding	1-7	Not included		Reliability, Validity and internal consistency ↑ [AGC21]
TXAI [HMKL23]	8	-	1-5	"I am confident in the [tool]. I feel that it works well."	XAI	Reliability ↑ [HMKL23]

AI systems, it is essential to consider the various phases of usage. Subjective measures are most effectively utilized in the pre- and post-technology use phases, as integrating them into the actual usage phase is difficult and can potentially disrupt individuals. In contrast, behavioral measures can be solely applied during the actual usage phase, as they offer data derived from interactions with the system or technology. Physiological measures are also commonly employed during the usage phase. However, to continuously track the progression of trust from pre-use to post-use and gain insights into specific situations, such as initial encounters with the technology, physiological measures can provide a more detailed understanding (Figure 2).

3.2 Trustworthiness

It follows from our definition (see Section 2.2) that trustworthiness can be put on a formal ground, for a given set of scenarios, if the expected behaviors of the system in those scenarios are formalized. The level of confidence and the precision of the measured trustworthiness are then proportional to the details made available in the requirement specification for the given set of scenarios.

A number of assessment frameworks are available. The FAIE-H toolkit designed by the Open Roboethics Institute [Ope20] aims to perform ethics assessment of healthcare AI projects, mainly focusing on the development and deployment

Table 3 Summary of physiological measures of trust.

Parameter	Explanations
Heart rate variability	Describes the fluctuations in the time interval between successive heartbeats and can be indicative of the autonomic nervous system's activity, providing insights into emotional responses and potential trust-related states [WN14]
Electrodermal activity	This is measured using electrodes in contact with the skin, detecting changes in ionic activity influenced by sweat [BL00, Cri02]. Changes can signify emotional arousal, potentially reflecting trust or anxiety [KZCM15]
Neural measures	Neural measures involve the examination of brain activity patterns to gain insights into cognitive processes related to trust. Techniques such as functional magnetic resonance imaging (fMRI), functional near-infrared spectroscopy (fNIRS), or electroencephalography (EEG) are employed to capture neural responses. These measures provide direct information about how the brain responds to trust-related stimuli, offering valuable data on the cognitive aspects of trust perception. Neural measures enhance our understanding of the underlying neurological processes associated with trust in various contexts, including human interactions with technology or AI [GPC+16, HBB+14, JDL19].
Eye gaze tracking	Involves monitoring the direction and duration of an individual's eye movements. By analyzing where a person directs their gaze, researchers can gain insights into visual attention patterns and cognitive processing related to trust. For instance, prolonged gaze at certain stimuli may indicate heightened interest or trust, while quick shifts in gaze may suggest uncertainty or scepticism [GKH+15, HLVK16].

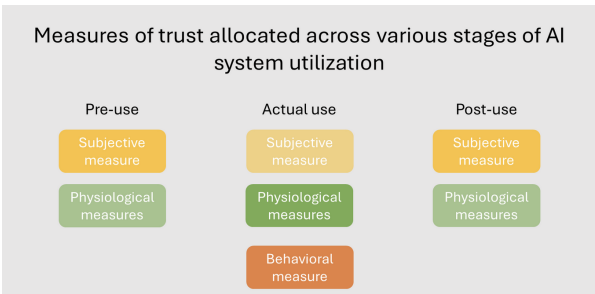

Fig. 2 Allocation of measurement methods according to their primary temporal use in the process of interaction with AI systems.

stages. The Assessment List for Trustworthy AI (ALTAI) [(AI20] is an online tool that prompts users with a set of questions related to the dimensions of trustworthiness. The outcomes are subsequently visualized in a spider diagram. A recent example is the Z-Inspection toolkit [ZBB+21], a holistic approach which aims to orchestrate assessments from a team of multi-disciplinary experts, and can be applied at any stage of the AI system lifetime. Due to this multidisciplinary and multifaceted nature of trustworthiness, in the remainder of this section, we provide an overview of some of the key aspects of trustworthiness, as exemplified in the EU Guideline for Trustworthy AI [(AI19], and provide respective metrics to measure them.

Human Oversight and Agency. The degree of human agency and oversight in relation to an AI system is in direct correlation with the perceived level of risk associated with the scenario. This relationship has been categorized in [(AI19] into three distinct paradigms: human-in-the-loop, human-on-the-loop, and human-in-command. Other works [AF22] emphasize the need for a more nuanced understanding of these different levels of human interaction with AI systems. We may quantify the appropriate level of agency/oversight by carrying out a risk assessment in different scenarios, using one of the frameworks recently proposed (e.g., [oST23]), and comparing the result with the expected level of agency/oversight in each scenario. It is widely acknowledged that higher-risk AI necessitates more rigorous oversight to ensure responsible development, deployment, and operation.

Technical Robustness / Safety. Assessing technical robustness and safety can be achieved by evaluating compliance with established guidelines, such as those outlined by the UK NCSC [Nat23], and adherence to standards such as ISO 10218 [fSI11], and the IEEE P7009 standard for fail-safe design of autonomous and semi-autonomous systems [oEI24]. More rigorous approaches within formal methods offer greater ability to quantify properties. Exact (complete) approaches for deep neural networks include:

- Encoding natural networks and safety properties as constraints satisfaction problems. This approach ensures that the system operates within certain safety bounds [Ehl17, KHI+19]; and
- Assessing resilience against deliberate attempts to manipulate its behavior, known as adversarial attacks [HPG+17, NKR+18]. There are approaches to find "nearest" (with respect to specific metrics) adversarial examples [BIL+16, TXT17].

To mitigate complexity, incomplete approaches are available, based on abstract interpretation [GMDC+18, SGM+18], simulation [XTJ18], and approximation [WZC+18].

Table 4 Summary of measures of privacy as a part of trustworthiness metrics: differential privacy and informational leakage.

Measure	Explanations
Differential privacy	Two parameters ε and δ represent the privacy guarantee, i.e., the probability an attacker might guess whether an individual's data was in the training set, and the probability this guarantee may not hold, respectively [DKM+06]. Recent work by Jayaraman and Evans [JE19] investigates the practical challenges of using differentially private machine learning. Estimation methods for privacy guarantees are presented in [ZBWT+23].
Information leakage	The amount of information about an individual leaked by an AI system. Work by Hannun et al. [HGvdM22] explores using a specific measure (Fisher information) for quantifying leakage. Information leakage can also be measured by evaluating how effective certain attacks are. For instance, membership inference attacks allow attackers to potentially learn if a data point was used to train the model (see, e.g., [HSS+22])

Although less studied, approaches for other models exist, such as support vector machines [RZ19], and for identifying nearest adversarial examples in decision tree ensembles [KTJ16, CZS+19].

Privacy and Data Governance. Data ownership, accountability, and management are crucial aspects of data governance [KB10]. Assessing them involves evaluating the clarity and comprehensiveness of roles and responsibilities, and a rigorous evaluation of implemented measures, ensuring they meet the compliance and quality standards established by regulations such as GDPR. For instance, Privacy Impact Assessments (PIA) and Data Protection Impact Assessments (DPIA) should be carried out in certain cases. Other standards include ISO 31700 Privacy by Design Standard, KMPG's Privacy by Design Assessment Framework, ISO 27001.

The literature on assessing the privacy of AI systems provides some concrete measures listed in Table 4.

Diversity and Fairness. Given two groups of agents that differ only in a number of protected characteristics, an AI-enabled system is fair when the outcome of the system does not change significantly for similar inputs only differing on legally protected or diversity-related characteristics. Fairness is typically categorized into two broad categories: group and individual fairness. Group fairness refers to guaranteeing a statistical notion of similarity of outcomes across different groups of the society characterized by their sensitive attributes. Individual fairness requires the outcomes of an AI-enabled system to be identical (or similar with respect to a distance measure) for individuals that only differ with respect to their sensitive attribute (or are similar, again with respect to a distance measure on their other features). There are different metrics for measuring the change of outcomes and similar inputs. We refer to recent textbooks [BHN23, KR19] and surveys [MMS+21, PS22] on the definitions of fairness for machine learning and AI systems. In the remainder of this section, we provide only some examples of such metrics and how they can be used to measure fairness. We conclude by reviewing some of the shortcomings of using these metrics as well as mentioning some other diversity-related metrics that can be measured on AI-enabled systems.

Some commonly used metrics for measuring fairness include: statistical parity difference (SPD), equal opportunity difference (EOD), average odds difference (AOD) and disparate impact (DI). Their definitions are provided in Table 5.

Despite the rich literature on quantitative metrics of fairness and diversity, researchers in this area find that optimizing for specific metrics can result in more unfair systems [CGN+23]. Also such metrics can be gamed and the powerful actors in an ecosystem may use the metric that works best for them or optimize for the commonly-used metrics without providing genuine support for fairness or diversity [Nar22].

Societal and environmental sustainability. To assess the societal impact of AI systems, *Algorithm Impact Assessment* tools have been proposed, such as the one from the AI Now Institute [RSCW18], which targets public agencies. With the advent of generative AI, impact assessment has become more challenging, as contextual factors, human interactions, and the combination of modalities (e.g., text, images, videos) have become increasingly relevant [WRM+23]. It is crucial to measure impact across several dimensions, some of which apply differently to various modalities [STA+23].

Measuring the precise environmental impact can be difficult, as most AI systems are deployed in the cloud, where the majority of carbon emissions stem from construction, infrastructure, and hardware manufacturing [GKL+22].

When considering AI algorithms in isolation, most energy consumption occurs during the training phase, which can last for months. Tools for measuring the carbon footprint are available, such as the "Machine Learning Emissions Calculator" [LLSD19] and the"Carbon Tracker" [AKS20].

Accountability. The main challenge regarding the accountability of AI systems is the lack of comprehensive "internal auditing", i.e., auditing processes that should be applied before deployment by companies designing such systems, rather than after deployment, when issues have already arisen [RSW+20].

Table 5 Summary of four common metrics for fairness as a part of trustworthiness: statistical parity difference (SPD), equal opportunity difference (EOD), Average Odds Difference (AOD), and disparate impact (DI).

Fairness Metric	Intuition	Formalization	Explanation
Statistical Parity Difference (SPD)	Equal rate of favorable outcomes between majority and minority	$SPD = P[Y_{pred} = 1 \mid A = 1] - P[Y_{pred} = 1 \mid A = 0]$	The difference of conditional probability for predicting favorable outcomes between the majority and minority conditions. Non-zero values will classify the amount of bias.
Disparate Impact (DI)	Similar to SPD, only the difference is replaced with the ratio	DI = P[Y=1 | A=1] / P[Y=1 | A = 0]	The ratio of conditional probability for predicting favorable outcomes between the majority and minority conditions. Non-zero values will classify the amount of bias.
Equal Opportunity (EOD)	Equal rate of true (justified) favorable outcomes between majority and minority	$EOD = P[Y_{pred} = 1 \mid A = 1, Y = 1] - P[Y_{pred} = 1 \mid A = 0, Y = 1]$	The difference of conditional probability for predicting true positive (favorable) outcomes between the majority and minority conditions. Non-zero values will classify the amount of bias.
Average Odd Difference (AOD)	Average of differences for false (unjustified) and true favorable outcomes between the majority and minority	$AOD = avg_{iin\{0,1\}}(.5 * (P[Y_{pred} = 1 \mid A = 1, Y = i] - P[Y_{pred} = 1 \mid A = 0, Y = i]))$	The difference of conditional probability for predicting true positive (favorable) outcomes between the majority and minority conditions. Non-zero values will classify the amount of bias in both cases.

To quantify accountability, organizations need to formalize ethical principles guiding development and measure compliance at each stage. Outcomes should be documented, and proactive risk analysis should be conducted [RSW+20]. Key Performance Indicators can also be used to better measure accountability [PDSdG21].

3.3 Calibrated trust

Trust calibration is typically assessed through one of the following three methods (we refer to the survey by Wischnewski et al. [WKM23b] for a more detailed overview):

1. Relative measures: comparing perceived trustworthiness among different groups and systems, e.g., systems with high and low trustworthiness; the relative differences can be used to establish calibration of trust among user groups;
2. Correlative measure: estimating the correlation between the dynamics of trust and trustworthiness, for instance by measuring how trust changes as trustworthiness evolves ("trust sensitivity"); and
3. Behavioral measures: measuring the deviation from "ideal" user behavior (e.g., response time) under different circumstances (e.g., in scenarios featuring different trustworthiness levels).

Given its ease of integration, the first method is most frequently employed in studies evaluating trust calibration. In the context of XAI, Naiseh et al. [NAJA23] used the cognitive-based trust scale introduced by Madsen and Gregor [MG00], which measures how effectively the XAI interface aids users in comprehending, relying on and perceiving the technical proficiency of the AI. Furthermore, the authors examined behavioral markers of trust calibration, leveraging an objective metric proposed by Wang et al. [WYAL19], which assesses the accuracy of participants' decisions. In addition to the advantages of this method, Wischnewski and colleagues [WKM23b] mention that the method doesn't indicate whether a system with lower capabilities might be seen as overly trustworthy or untrustworthy, leading to overtrust or under-trust. Similarly, a high-capability system might cause either excessive or insufficient perceived trustworthiness, resulting in overtrust or undertrust. Many interventions are possible to calibrate trust, including 1) providing prior information about the system's trustworthiness factors, 2) providing run-time information about the systems decisions and their underlying causes, and 3) receiving and providing feedback after interactions [WKM23b].

4 Causes and Correlations

4.1 Trust Factors

When considering antecedents of trust in AI, again we need to apply them to the broad variety of AI-based systems (e.g., algorithms, automated vehicles, chatbots, and robots), resulting in a wide range of factors that can impact trust. Previous works propose a categorization of trust antecedents in AI-based systems into three overarching groups relating to the human trustor, the technology trustee, and shared contextual factors [KKBH23]. In the spectrum of AI-based systems, factors within the human trustor category underscore the

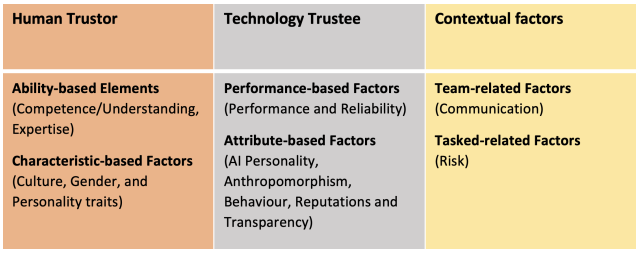

Fig. 3 Antecedents of trust in AI-enabled systems.

significance of ability-based elements (Competency/Understanding and Expertise) and along with characteristics-based factors (Culture, Gender, and Personality traits). In the realm of the technology trustee category, performance-based factors (Performance and Reliability) along with attribute-based factors (e.g., AI Personality, Anthropomorphism, Behavior, Reputation, and Transparency) are identified to be relevant. Additionally, within the third category team-related (Communication) and task-related factors (Risk) have been reported to be relevant [HBS+11, HKK+21, HB15, KKBH23, YW22].

Regarding trustee-related factors, we would like to emphasize the role of anthorpomorphism and more generally, embodiment. In interpersonal relationships trust is notably contingent upon the physical attributes of the trusted individual [CH09, DSY12]. In AI-based systems there is a great variety of embodiment forms, which means that the degree of physical appearance varies greatly: as a physical robot, as a virtual agent or bot, or in forms that are invisible to the user, embedded in a computer or another tool. Current empirical data indicate that the extent and nature of embodiment, coupled with the degree of machine intelligence integrated into the technology, have considerable implications for the trust exhibited by users[GW20].

Figure 3 describes the relationship of AI-relevant measure of trust (see Section 3.1, the color coding describes the category to which those factors belong), assessed before, during and after interacting with the AI system as well as the influencing factors on building trust, derived from the three categories human trustee, technology trustee and contextual factors. In order to influence trust, a feedback loop from the measures to the influencing factors indicates the possibility of strengthening certain factors to enhance trust in the AI-based system.

4.2 Trustworthiness Factors

The principles of trustworthiness, covered in Section 2.2, can be incorporated into the requirement specification and system design, and can be guaranteed by construction. Examples of such trustworthy-by-construction techniques include privacy-preserving architectures for AI [ACG+16] and AI-enabled systems [HRC20]; certified (adversarial) robustness for AI [CRK19, LCWC19]; and inherently explainable AI models [NATJA21].

However, not all components of an AI-enabled system are designed based on the trustworthiness principles and some are obtained as third-party off-the-shelf components. In such cases, in order to ensure trustworthiness, a rigorous specification of trustworthiness requirements [ABC+23] and structured validation and verification techniques [MCF+23] are needed. The techniques involved in validation and verification include: formal verification (including theorem proving and model checking), various forms of testing (including model-based testing and falsification), and various forms of user studies [AMV23].

When shortcomings in trustworthiness principles are detected, those can be remedied by deploying fixes, re-synthesizing and re-training parts of the system, and / or developing and deploying complementary components and wrappers to ensure the trustworthiness requirements [APN+20].

4.3 Calibration and Interplay

Trustworthiness of the AI-based system and human trust will change during the course of interaction between the human and the system. Based on the mental model of the human (how the system will behave in certain situations or given certain commands, trust is gained when the system behaves accordingly, trust is lost if the system reacts unexpectedly. How easy trust is lost or gained depends on the previous "trust level" caused by the influencing factors described in Section 4.1. If human trust is low, measured in an initial assessment before interacting with the system, the system needs to communicate or display its trustworthiness more openly and clearly than if human trust is initially high.

However, during the interaction between the human and system, human trust will be influenced. If the human gains more trust (adapting their mental model), the system can reduce its explanation and the human is willing to reduce the control over the system (e.g. from human-in-control to human-in-the-loop). In order to seamlessly adapt the configuration of the system towards the level of human trust/mental model assessed, the system needs to integrate certain trust measurements to adapt automatically its explanations or, if not possible, to describe which level of trust is needed to successfully use the system (often generally indicated as "beginner" or "expert level" user).

5 Implications for Stakeholders

The description of trust, trustworthiness, and their interactions cannot be uniform for all stakeholders. It is evident that the general public does not have the same level of understanding of AI as a programmer or certifier. Therefore, the

public may need additional explanations that are apparent to developers. The following distinguishes between various stakeholders, including the wider public, certification bodies, investigators, expert witnesses, and lawyers.

5.1 Implications for the General Public

Artificial intelligence is transforming society in numerous ways. It is crucial that individuals possess an understanding of how these systems operate when working with or utilizing them in public. To fully harness the potential of the technology, trust and trustworthiness in AI need to be aligned to create calibrated trust. One way to achieve this could be the education of the general public about the opportunities and risks of using AI transparently, and how over- or under- trust can affect both the environment and the society. To achieve this goal, discussion panels should facilitate open dialogue, supported by easily understandable media that are accessible to the general public through mass media such as TV, newspapers and social media. However, this approach may be too simplistic. It could be argued that the burden of understanding these complex systems should not fall solely on end users [Rya20]. An alternative option would be to design systems and contexts in which trustworthy technologies are the norm, and to develop strategies to ensure that only reliable systems are deployed. Furthermore, the development of AI systems could be informed by the understanding that these systems adapt to the trust level of users and provide clear signals when they are over- or under-trusted. This would shift the responsibility for ensuring the trustworthiness of AI systems from the general public to the developers and deployers of these technologies.

5.2 Implications for Certification Authorities and Assessors

In the context of certifying AI systems, independent certification bodies should have a calibrated trust in the technology. Over- or under -trust would be fatal, especially in the case of system safety acceptance. Depending on how safety-critical a system is, it must meet different safety requirements. The safety requirements for an AI system for automated driving are significantly higher than for an AI system for creative image creation. An error in automated driving can have fatal consequences, whereas an error in the second example only produces undesirable results. Certification bodies must be able to verify that AI developers have thoroughly tested and checked their systems to achieve a trustworthy AI system. Depending on the safety requirements, the developers will need to provide the technical description of the system, the source code, and information about how and with which data the AI has been trained.

5.3 Implications for Investigators, Expert Witnesses and Lawyers

To ensure a fair judgement or appropriate compensation in the event of a dispute or accident, investigators, lawyers or other experts require access to reliable information. This information should be provided by those responsible for the safety and inspection of the AI. It is also important that lawyers, investigators and experts can comprehend the causes and consequences of the situation. Over- or under-trust is not acceptable in this role. To achieve reliable trust in AI, lawyers, investigators and experts must also consider the attributes of the developers. This includes fundamental aspects such as the implementation of quality assurance and a data security strategy. Certification of these aspects, for example, in accordance with [fSI15] or [fSI01] and TISAX, enhances trustworthiness.

6 Conclusions

This paper explores trust and trustworthiness in AI systems, and proposes a formalization that relies on the system, the scenario, and the user, highlighting the highly contextual nature of these concepts. Alignment of trust and trustworthiness is essential to achieve calibrated trust, i.e., a state in which the user's trust reflects the actual level of trustworthiness for a specific scenario. Calibration is a dynamic process, as the user acquires experience and their expectations adapt. Under-trust or over-trust may occur when the alignment is not achieved, and both can have detrimental effects on the effectiveness and safety of AI systems.

As formalization requires quantifying trust and trustworthiness, we review a range of approaches to measure trust, from subjective self-reports to objective behavioral assessments, and approaches to measure trustworthiness, detailing methods for assessing a variety of dimensions, as well as methods to measure trust calibration.

We analyze factors that can impact these concepts. For trust, the impacts stem from a combination of human factors, technology attributes, and contextual influences. For trustworthiness, we describe ways in which its principles can be integrated into AI system design, highlighting the need for validation and verification. For calibrated trust, effective communication about the AI system's behavior and capabilities can help the user adjust their trust levels appropriately. Finally, we discuss the incarnation of (calibrated) trust and trustworthiness in a range of contexts, emphasizing the need for tailored explanations about the system to different stakeholders. However, several significant challenges remain. One major challenge is the dynamic calibration of trust over time. Trust levels are not static; they fluctuate based on system performance and user experiences, and current systems lack

the adaptability to manage these fluctuations effectively. Furthermore, balancing cognitive and emotional trust presents a complex issue. While AI systems can support cognitive trust by offering transparency and logical reasoning, fostering emotional trust—rooted in feelings of security and connection—requires further attention.

Assessing trustworthiness also remains problematic, especially across diverse scenarios and use cases. Many AI systems still lack the transparency and robustness necessary for users to gauge their reliability confidently. This is compounded by the dual issue of overtrust and distrust, where users either rely too heavily on systems or fail to utilize them fully, both of which can lead to suboptimal or even dangerous outcomes.

Additionally, while there is growing research on trust in AI, much of it focuses on short-term interactions. Little is understood about the dynamics of trust in long-term engagements, which are critical for many real-world applications. Ethical concerns and fairness in AI systems also pose ongoing challenges. Ensuring that AI operates in a way that is both transparent and fair while maintaining performance requires more sophisticated solutions than currently available.

Trust calibration becomes particularly difficult in high-risk domains, such as healthcare or autonomous driving, where system failures can have severe consequences. In these areas, achieving calibrated trust is essential, yet it remains one of the most pressing challenges. Finally, addressing these issues will require stronger cross-disciplinary collaboration, integrating insights from fields such as cognitive psychology, computer science, and ethics to develop more comprehensive frameworks for trust and trustworthiness in AI.

Future research must focus on overcoming these challenges by refining the methods used to assess trust, designing systems that can adapt to trust dynamics in real time, and ensuring that AI systems adhere to ethical and fairness standards. Only through this approach can we build AI systems that are not only effective but also trustworthy, ensuring their safe and beneficial integration into society.

7 Acknowledgments

Mohammad Reza Mousavi has been partially supported by the UKRI Trustworthy Autonomous Systems Node in Verifiability, Grant Award Reference EP/V026801/2, EPSRC project on Verified Simulation for Large Quantum Systems (VSL-Q), grant reference EP/Y005244/1 and the EPSRC project on Robust and Reliable Quantum Computing (RoaRQ), Investigation 009 Model-based monitoring and calibration of quantum computations (ModeMCQ), grant reference EP/W032635/1.

References

ABC+23. Dhaminda B. Abeywickrama, Amel Bennaceur, Greg Chance, Yiannis Demiris, Anastasia Kordoni, Mark Levine, Luke Moffat, Luc Moreau, Mohammad Reza Mousavi, Bashar Nuseibeh, Subramanian Ramamoorthy, Jan Oliver Ringert, James Wilson, Shane Windsor, and Kerstin Eder. On specifying for trustworthiness. *Commun. ACM*, 67(1):98–109, dec 2023.

ACG+16. Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 308–318, New York, NY, USA, 2016. Association for Computing Machinery.

ADBD+21. Alexander M Aroyo, Jan De Bruyne, Orian Dheu, Eduard Fosch-Villaronga, Aleksei Gudkov, Holly Hoch, Steve Jones, Christoph Lutz, Henrik Sætra, Mads Solberg, et al. Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. *Paladyn, Journal of Behavioral Robotics*, 12(1):423–436, 2021.

AF22. Marc Anderson and Karën Fort. Human where? a new scale defining human involvement in technology communities from an ethical standpoint. *The International Review of Information Ethics*, 31(1), Nov. 2022.

AGC21. Anton Angelgardt, Elena Gorbunova, and Maria Chumakova. An assessment of trust in artificial intelligent agents: Tool development. *Higher School of Economics Research Paper No. WP BRP*, 128, 2021.

(AI19. High-Level Expert Group on Artificial Intelligence (AI HLEG). Ethics guidelines for trustworthy AI, 2019. Last accessed: 08 April 2024.

(AI20. High-Level Expert Group on Artificial Intelligence (AI HLEG). Assessment list for trustworthy AI (ALTAI), 2020. Last accessed: 08 April 2024.

Ajz91. Icek Ajzen. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2):179–211, 1991. Theories of Cognitive Self-Regulation.

AKS20. Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems, July 2020. arXiv:2007.03051.

ALCL23. Areen Alsaid, Mengyao Li, Erin K Chiou, and John D Lee. Measuring trust: A text analysis approach to compare, contrast, and select trust questionnaires. *Frontiers in Psychology*, 14, 2023.

AMV23. Hugo Araujo, Mohammad Reza Mousavi, and Mahsa Varshosaz. Testing, validation, and verification of robotic and autonomous systems: A systematic review. *ACM Trans. Softw. Eng. Methodol.*, 32(2), mar 2023.

APN+20. Raja Ben Abdessalem, Annibale Panichella, Shiva Nejati, Lionel C. Briand, and Thomas Stifter. Automated repair of feature interaction failures in automated driving systems. In *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis*, ISSTA 2020, page 88–100, New York, NY, USA, 2020. Association for Computing Machinery.

ASL20. Ighoyota Ben Ajenaghughrure, Sonia Da Costa Sousa, and David Lamas. Measuring trust with psychophysiological signals: a systematic mapping study of approaches used. *Multimodal Technologies and Interaction*, 4(3):63, 2020.

BHN23. Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

BIL+16. Osbert Bastani, Yani Ioannou, Leonidas Lampropoulos, Dimitrios Vytiniotis, Aditya Nori, and Antonio Criminisi. Measuring neural net robustness with constraints. *Advances in neural information processing systems*, 29, 2016.

BKH+22. Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão, and Sonia Sousa. A systematic literature review of user trust in AI-enabled systems: An hci perspective. *International Journal of Human–Computer Interaction*, pages 1–16, 2022.

BL00. Margaret M Bradley and Peter J Lang. Measuring emotion: Behavior, feeling, and physiology. In *Cognitive neuroscience of emotion*, pages 242–276. Oxford University Press, 2000.

BNR19. Matthew Brzowski and Dan Nathan-Roberts. Trust measurement in human–automation interaction: A systematic review. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 1595–1599. SAGE Publications Sage CA: Los Angeles, CA, 2019.

Car23. J Adam Carter. Simion and kelp on trustworthy AI. *Asian Journal of Philosophy*, 2(1):18, 2023.

CDv+10. Luke J. Chang, Bradley B. Doll, Mascha van 't Wout, Michael J. Frank, and Alan G. Sanfey. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology*, 61(2):87–105, 2010.

CGN+23. Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *J. Mach. Learn. Res.*, 24:312:1–312:117, 2023.

CH09. Jinsook E Cho and Haiyan Hu. The effect of service quality on trust and commitment varying across generations. *International journal of consumer studies*, 33(4):468–476, 2009.

CMH21. Jing Chen, Scott Mishler, and Bin Hu. Automation error type and methods of communicating automation reliability affect trust and performance: An empirical study in the cyber domain. *IEEE Transactions on Human-Machine Systems*, 51(5):463–473, 2021.

Cri02. Hugo D Critchley. Electrodermal responses: what happens in the brain. *The Neuroscientist*, 8(2):132–142, 2002.

CRK19. Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, 09–15 Jun 2019.

CZS+19. Hongge Chen, Huan Zhang, Si Si, Yang Li, Duane Boning, and Cho-Jui Hsieh. Robustness verification of tree-based models. *Advances in Neural Information Processing Systems*, 32, 2019.

Dav85. Fred D Davis. *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. PhD thesis, Massachusetts Institute of Technology, 1985.

DKKT17. Igor Dolgov, Elizabeth K Kaltenbach, Ahmed S Khalaf, and Zachary O Toups. Measuring human performance in the field. In *Human Factors in Practice*, pages 37–54. CRC Press, 2017.

DKM+06. Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25*, pages 486–503. Springer, 2006.

DO22. Stefan Daschner and Robert Obermaier. Algorithm aversion? on the influence of advice accuracy on trust in algorithmic advice. *Journal of Decision Systems*, 31(sup1):77–97, 2022.

DPJ12. Milena Dzhelyova, David I. Perrett, and Ines Jentzsch. Temporal dynamics of trustworthiness perception. *Brain Research*, 1435:81–90, 2012.

DSY12. Jefferson Duarte, Stephan Siegel, and Lance Young. Trust and credit: The role of appearance in peer-to-peer lending. *The Review of Financial Studies*, 25(8):2455–2484, 2012.

DVPJ+20. Ewart J De Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics*, 12(2):459–478, 2020.

DWC22. Xinyi Ding, Tianlan Wei, and Cong Cao. Towards a four-dimensional dynamic trust model in b2c cross-border e-commerce. In *Proceedings of the 2022 13th International Conference on E-Education, E-Business, E-Management, and E-Learning*, pages 414–418, 2022.

Ehl17. Rüdiger Ehlers. Formal verification of piece-wise linear feed-forward neural networks. In Deepak D'Souza and K. Narayan Kumar, editors, *Automated Technology for Verification and Analysis - 15th International Symposium, ATVA 2017, Pune, India, October 3-6, 2017, Proceedings*, volume 10482 of *Lecture Notes in Computer Science*, pages 269–286. Springer, 2017.

FA77. Martin Fishbein and Icek Ajzen. *Belief, attitude, intention, and behavior: An introduction to theory and research*. Longman Higher Education, 1977.

FKJ+21. 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*, pages 1492–1503, 2021.

FLC+23. Vincent Fer, Daniel Lafond, Gilles Coppin, Mathias Bollaert, Olivier Grisvard, and Pierre De Loor. Trust in automation: Analysis and model of operator trust in decision aid AI over time. In *Conference on Artificial Intelligence for Defense*, 2023.

fSI01. International Organization for Standardization (ISO). ISO/IEC 27001:2022 information security, cybersecurity and privacy protection — information security management systems — requirements, 201. Last accessed: 01 July 2024.

fSI11. International Organization for Standardization (ISO). ISO 10218-1:2011 robots and robotic devices safety requirements for industrial robots, 2011. Last accessed: 08 April 2024.

fSI15. International Organization for Standardization (ISO). ISO 9001:2015 quality management systems — requirements, 2015. Last accessed: 01 July 2024.

GAB+21. Omri Gillath, Ting Ai, Michael S Branicky, Shawn Keshmiri, Robert B Davison, and Ryan Spaulding. Attachment and trust in artificial intelligence. *Computers in Human Behavior*, 115:106607, 2021.

Gam88. Diego Gambetta. *Trust: Making and breaking cooperative relations*. Wiley-Basil Blackwell, 1988.

GKH+15. Christian Gold, Moritz Körber, Christoph Hohenberger, David Lechner, and Klaus Bengler. Trust in automation–before and after the experience of take-over scenarios in a highly automated vehicle. *Procedia Manufacturing*, 3:3025–3032, 2015.

GKL+22. Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S. Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. Chasing carbon: The elusive environmental footprint of computing. *IEEE Micro*, 42(4):37–47, 2022.

GKS03. David Gefen, Elena Karahanna, and Detmar W Straub. Trust and tam in online shopping: An integrated model. *MIS quarterly*, pages 51–90, 2003.

GMDC+18. Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2018.

GPC+16. Kimberly Goodyear, Raja Parasuraman, Sergey Chernyak, Poornima Madhavan, Gopikrishna Deshpande, and Frank Krueger. Advice taking from humans and machines: An fmri and effective connectivity study. *Frontiers in Human Neuroscience*, 10:542, 2016.

GSL19. Siddharth Gulati, Sonia Sousa, and David Lamas. Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, 38(10):1004–1015, 2019.

GW20. Ella Glikson and Anita Williams Woolley. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660, 2020.

Har02. R. Hardin. *Trust and Trustworthiness*. Russell Sage Foundation Series on Trust. Russell Sage Foundation, 2002.

Har06. Russell Hardin. *Trust*, volume 10. Polity, 2006.

HB15. Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3):407–434, 2015.

HBB+14. Leanne M Hirshfield, Philip Bobko, Alex Barelka, Stuart H Hirshfield, Mathew T Farrington, Spencer Gulbronson, and Diane Paverman. Using noninvasive brain measurement to explore the psychological effects of computer malfunctions on users during human-computer interactions. *Advances in Human-Computer Interaction*, 2014:2–2, 2014.

HBS+11. Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC 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.

HGvdM22. Awni Y. Hannun, Chuan Guo, and Laurens van der Maaten. Measuring data leakage in machine-learning models with fisher information (extended abstract). In Luc De Raedt, editor, *IJCAI*, pages 5284–5288, 2022.

HKK+21. Peter A Hancock, Theresa T Kessler, Alexandra D Kaplan, John C Brill, and James L Szalma. Evolving trust in robots: specification through sequential and comparative meta-analyses. *Human factors*, 63(7):1196–1229, 2021.

HLVK16. Sebastian Hergeth, Lutz Lorenz, Roman Vilimek, and Josef F Krems. Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human factors*, 58(3):509–519, 2016.

HMKL21. Robert Hoffman, Shane Mueller, Gary Klein, and Jordan Litman. Measuring trust in the XAI context. In *Tech reports for the DARPA XAI program: Task Area 2*. PsyArXiv, 2021.

HMKL23. Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, 5:1096257, 2023.

HP21. Eleanore Hickman and Martin Petrin. Trustworthy AI and corporate governance: the eu's ethics guidelines for trustworthy artificial intelligence from a company law perspective. *European Business Organization Law Review*, 22:593–625, 2021.

HPG+17. Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

HRC20. Muneeb Ul Hassan, Mubashir Husain Rehmani, and Jinjun Chen. Differential privacy techniques for cyber physical systems: A survey. *IEEE Communications Surveys & Tutorials*, 22(1):746–789, 2020.

HSS+22. Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.

Inn23. Innovate UK. Report on the core principles and opportunities for responsible and trustworthy ai, 2023.

JBD00. 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.

JDL19. Eun-Soo Jung, Suh-Yeon Dong, and Soo-Young Lee. Neural correlates of variations in human trust in human-like machines during non-reciprocal interactions. *Scientific reports*, 9(1):9975, 2019.

JE19. Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1895–1912, 2019.

JG05. Devon Johnson and Kent Grayson. Cognitive and affective trust in service relationships. *Journal of Business research*, 58(4):500–507, 2005.

JMMG21. 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*, pages 624–635, 2021.

Jon12. Karen Jones. Trustworthiness. *Ethics*, 123(1):61–85, 2012.

KB06. Sherrie YX Komiak and Izak Benbasat. The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS quarterly*, pages 941–960, 2006.

KB10. Vijay Khatri and Carol V Brown. Designing data governance. *Communications of the ACM*, 53(1):148–152, 2010.

KdVW+21. Spencer C Kohn, Ewart J de Visser, Eva Wiese, Yi-Ching Lee, and Tyler H Shaw. Measurement of trust in automation: A narrative review and reference guide. *Frontiers in psychology*, 12:604977, 2021.

KHI+19. Guy Katz, Derek A. Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljić, David L. Dill, Mykel J. Kochenderfer, and Clark Barrett. The marabou framework for verification and analysis of deep neural networks. In Isil Dillig and Serdar Tasiran, editors, *Computer Aided Verification*, pages 443–452, Cham, 2019. Springer International Publishing.

KK22. Nakyung Kyung and Hyeokkoo Eric Kwon. Rationally trust, but emotionally? the roles of cognitive and affective trust in laypeople's acceptance of AI for preventive care operations. *Production and Operations Management*, 2022.

KKBH23. Alexandra D Kaplan, Theresa T Kessler, J Christopher Brill, and PA Hancock. Trust in artificial intelligence: Meta-analytic findings. *Human factors*, 65(2):337–359, 2023.

KKOT23. Sage Kelly, Sherrie-Anne Kaye, and Oscar Oviedo-Trespalacios. What factors contribute to the acceptance of artificial intelligence? a systematic review. *Telematics and Informatics*, 77:101925, 2023.

Kör19. Moritz Körber. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20*, pages 13–30. Springer, 2019.

KPB18. Moritz Körber, Lorenz Prasch, and Klaus Bengler. Why do i have to drive now? post hoc explanations of takeover requests. *Human factors*, 60(3):305–323, 2018.

KR19. Michael Kearns and Aaron Roth. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, 2019.

KS23. Christoph Kelp and Mona Simion. What is trustworthiness? *Noûs*, 57(3):667–683, 2023.

KTJ16. Alex Kantchelian, J Doug Tygar, and Anthony Joseph. Evasion and hardening of tree ensemble classifiers. In *International conference on machine learning*, pages 2387–2396. PMLR, 2016.

KURD22. Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. Trustworthy artificial intelligence: A review. *ACM Comput. Surv.*, 55(2), jan 2022.

Kus24. Isabel Kusche. Possible harms of artificial intelligence and the eu ai act: fundamental rights and risk. *Journal of Risk Research*, 0(0):1–14, 2024.

KZCM15. Ahmad Khawaji, Jianlong Zhou, Fang Chen, and Nadine Marcus. Using galvanic skin response (gsr) to measure trust and cognitive load in the text-chat environment. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1989–1994, 2015.

LCWC19. Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

LLSD19. Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *CoRR*, abs/1910.09700, 2019.

LM92. John Lee and Neville Moray. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10):1243–1270, 1992.

LMT15. Nancy K Lankton, D Harrison McKnight, and John Tripp. Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, 16(10):1, 2015.

LRFI23. Minha Lee, Peter Ruijten, Lily Frank, and Wijnand IJsselsteijn. Here's looking at you, robot: The transparency conundrum in hri. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 2120–2127. IEEE, 2023.

LS04. John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

LS19. Yidu Lu and Nadine Sarter. Feedback on system or operator performance: Which is more useful for the timely detection of changes in reliability, trust calibration and appropriate automation usage? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 312–316. SAGE Publications Sage CA: Los Angeles, CA, 2019.

LW85. J David Lewis and Andrew Weigert. Trust as a social reality. *Social forces*, 63(4):967–985, 1985.

LYQ+22. Xingyang Lv, Yufan Yang, Dazhi Qin, Xingping Cao, and Hong Xu. Artificial intelligence service recovery: The role of empathic response in hospitality customers' continuous usage intention. *Computers in Human Behavior*, 126:106993, 2022.

McA95. Daniel J McAllister. Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management journal*, 38(1):24–59, 1995.

MCF+23. Mohammad Reza Mousavi, Ana Cavalcanti, Michael Fisher, Louise Dennis, Rob Hierons, Bilal Kaddouh, Effie Lai-Chong Law, Rob Richardson, Jan Oliver Ringer, Ivan Tyukin, and Jim Woodcock. Trustworthy autonomous systems through verifiability. *Computer*, 56(2):40–47, 2023.

MCK02. 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.

MCTC11. D Harrison Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F Clay. Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on management information systems (TMIS)*, 2(2):1–25, 2011.

MDS95. 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.

Mer11. Stephanie M Merritt. Affective processes in human–automation interactions. *Human Factors*, 53(4):356–370, 2011.

MG00. Maria Madsen and Shirley Gregor. Measuring human-computer trust. In *11th australasian conference on information systems*, volume 53, pages 6–8. Citeseer, 2000.

MMS+21. Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), July 2021.

MTP00. Linda D Molm, Nobuyuki Takahashi, and Gretchen Peterson. Risk and trust in social exchange: An experimental test of a classical proposition. *American Journal of Sociology*, 105(5):1396–1427, 2000.

NAJA23. Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *Int. J. Hum. Comput. Stud.*, 169:102941, 2023.

Nar22. Arvind Narayanan. The limits of the quantitative approach to discrimination, 2022. Available online from https://www.cs.princeton.edu/~arvindn/talks/baldwin-discrimination/baldwin-discrimination-transcript.pdf.

Nat23. National Cyber Security Centre (NCSC). Guidelines for secure ai system development, 2023. Last accessed: 08 April 2024.

NATJA21. Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. Explainable recommendation: when design meets trust calibration. *World Wide Web*, 24(5):1857–1884, 2021.

NKR+18. Nina Narodytska, Shiva Kasiviswanathan, Leonid Ryzhyk, Mooly Sagiv, and Toby Walsh. Verifying properties of binarized deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

NW10. PG Neumann and Robert NM Watson. Capabilities revisited: A holistic approach to bottom-to-top assurance of trustworthy systems. In *Fourth Layered Assurance Workshop, Austin, Texas, December*, 2010.

oEI24. Institute of Electrical and Electronics Engineers IEEE. IEEE P7009: IEEE Draft Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems, 2024. Last accessed: 08 April 2024.

Ope20. Open Roboethics Institute. Foresight into AI ethics in healthcare (faie-h): A toolkit for creating an ethics roadmap for your healthcare AI project, 2020. Last accessed: 08 April 2024.

oST23. National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). Technical report, US Department of Commerce, 2023.

OY20. Kazuo Okamura and Seiji Yamada. Adaptive trust calibration for human-AI collaboration. *Plos one*, 15(2):e0229132, 2020.

PDSdG21. Chris Percy, Simo Dragicevic, Sanjoy Sarkar, and Artur S. d'Avila Garcez. Accountability in AI: from principles to industry-specific accreditation. *AI Commun.*, 34(3):181–196, 2021.

PM10. Raja Parasuraman and Dietrich H Manzey. Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52(3):381–410, 2010.

PR97. Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

PS22. Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3), February 2022.

RF23. Yosef S Razin and Karen M Feigh. Converging measures and an emergent model: A meta-analysis of human-automation trust questionnaires. *arXiv preprint arXiv:2303.13799*, 2023.

Rob16. Blaine G Robbins. What is trust? a multidisciplinary review, critique, and synthesis. *Sociology compass*, 10(10):972–986, 2016.

Rot80. Julian B Rotter. Interpersonal trust, trustworthiness, and gullibility. *American psychologist*, 35(1):1, 1980.

RSCW18. Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. Algorithmic impact assessments: a practical framework for public agency. *AI Now*, 9, 2018.

RSW+20. Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *FAT* '20*, pages 33–44. ACM, 2020.

Rya20. Mark Ryan. In ai we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, 26(5):2749–2767, 2020.

RZ19. Francesco Ranzato and Marco Zanella. Robustness verification of support vector machines. In *Static Analysis: 26th International Symposium, SAS 2019, Porto, Portugal, October 8–11, 2019, Proceedings 26*, pages 271–295. Springer, 2019.

SARH+21. Nathalie A. Smuha, Emma Ahmed-Rengers, Adam Harkens, Wenlong Li, James MacLaren, Riccardo Piselli, and Karen Yeung. How the EU can achieve legally trustworthy AI: A response to the European Commission's proposal for an artificial intelligence act. *SSRN*, 2021.

SC25. Eugenia Stamboliev and Tim Christiaens. How empty is trustworthy ai? a discourse analysis of the ethics guidelines of trustworthy ai. *Critical Policy Studies*, 19(1):39–56, 2025.

Sch16. Kristin E Schaefer. Measuring trust in human robot interactions: Development of the "trust perception scale-hri". In *Robust intelligence and trust in autonomous systems*, pages 191–218. Springer, 2016.

SGM+18. Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. Fast and effective robustness certification. *Advances in neural information processing systems*, 31, 2018.

SH23. Joseph Sifakis and David Harel. Trustworthy autonomous system development. *ACM Transactions on Embedded Computing Systems*, 22(3):1–24, 2023.

Sif12. Joseph Sifakis. Trustworthy computing systems. In *SENSORNETS*, page 5, 2012.

SMD07. F David Schoorman, Roger C Mayer, and James H Davis. An integrative model of organizational trust: Past, present, and future, 2007.

STA+23. Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Canyu Chen, Hal Daumé III, Jesse Dodge, Isabella Duan, et al. Evaluating the social impact of generative ai systems in systems and society. *arXiv preprint arXiv:2306.05949*, 2023.

TXT17. Vincent Tjeng, Kai Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. *arXiv preprint arXiv:1711.07356*, 2017.

UBD21. Daniel Ullrich, Andreas Butz, and Sarah Diefenbach. The development of overtrust: An empirical simulation and psychological analysis in the context of human–robot interaction. *Frontiers in Robotics and AI*, 8:554578, 2021.

VDVM13. Kees Van Dongen and Peter-Paul Van Maanen. A framework for explaining reliance on decision aids. *International Journal of Human-Computer Studies*, 71(4):410–424, 2013.

WCZ+20. Eric Ke Wang, Chien-Ming Chen, Dongning Zhao, Wai Hung Ip, and Kai Leung Yung. A dynamic trust model in internet of things. *Soft Computing*, 24:5773–5782, 2020.

WKM23a. Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023.

WKM23b. Magdalena Wischnewski, Nicole C. Krämer, and Emmanuel Müller. Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In Albrecht Schmidt, Kaisa Väänänen, Tesh Goyal, Per Ola Kristensson, Anicia Peters, Stefanie Mueller, Julie R. Williamson, and Max L. Wilson, editors, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 755:1–755:16. ACM, 2023.

WN14. Adam Waytz and Michael I Norton. Botsourcing and outsourcing: Robot, british, chinese, and german workers are for thinking—not feeling—jobs. *Emotion*, 14(2):434, 2014.

Woo09. DD Woods. Trust emerges from the dynamics of reciprocity, responsibility and resilience in networked systems. presentation at the working meeting on trust in cyberdomains. *Institute for human and machine Cognition, Pensacola, FL. Supported by the human Effectiveness Directorate, Air Force Research Laboratory, Wright-Patterson AFB, OH*, 2009.

WPTL+20. Heather M Wojton, Daniel Porter, Stephanie T. Lane, Chad Bieber, and Poornima Madhavan. Initial validation of the trust of automated systems test (toast). *The Journal of social psychology*, 160(6):735–750, 2020.

WRM+23. Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*, 2023.

WRZ01. Douglas A Wiegmann, Aaron Rich, and Hui Zhang. Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science*, 2(4):352–367, 2001.

WYAL19. Danding Wang, Qian Yang, Ashraf M. Abdul, and Brian Y. Lim. Designing theory-driven user-centric explainable AI. In Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos, editors, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 601. ACM, 2019.

WZC+18. Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. Towards fast computation of certified robustness for relu networks. In *International Conference on Machine Learning*, pages 5276–5285. PMLR, 2018.

XTJ18. Weiming Xiang, Hoang-Dung Tran, and Taylor T Johnson. Output reachable set estimation and verification for multilayer neural networks. *IEEE transactions on neural networks and learning systems*, 29(11):5777–5783, 2018.

YCC17. Beste F Yuksel, Penny Collisson, and Mary Czerwinski. Brains or beauty: How to engender trust in user-agent interactions. *ACM Transactions on Internet Technology (TOIT)*, 17(1):1–20, 2017.

YW22. Rongbin Yang and Santoso Wibowo. User trust in artificial intelligence: A comprehensive conceptual framework. *Electronic Markets*, 32(4):2053–2077, 2022.

ZBB+21. Roberto V. Zicari, John Brodersen, James Brusseau, Boris Düdder, Timo Eichhorn, Todor Ivanov, Georgios Kararigas, Pedro Kringen, Melissa McCullough, Florian Möslein, Naveed Mushtaq, Gemma Roig, Norman Stürtz, Karsten Tolle, Jesmin Jahan Tithi, Irmhild van Halem, and Magnus Westerlund. Z-Inspection (R): A process to assess trustworthy ai. *IEEE Transactions on Technology and Society*, 2(2):83–97, 2021.

ZBWT+23. Santiago Zanella-Beguelin, Lukas Wutschitz, Shruti Tople, Ahmed Salem, Victor Rühle, Andrew Paverd, Mohammad Naseri, Boris Köpf, and Daniel Jones. Bayesian estimation of differential privacy. In *International Conference on Machine Learning*, pages 40624–40636. PMLR, 2023.

A Glossary of Terms

- **Trust**: A user-centered concept, referring to the epistemic state that the system is going to behave as expected in a given situation, characterized by a certain level of risk and uncertainty.
- **Trustworthiness**: A property of a system where it satisfies the user's requirements in a specific scenario. A system is considered trustworthy if it consistently meets these expectations across various scenarios.
- **Calibrated Trust**: Trust of a user in a system that matches the actual level of trustworthiness for a given scenario. Calibration is considered dynamic, where trust adapts based on ongoing interaction with the system.
- **Cognitive trust**: Trust of a user in a system that matches the actual level of trustworthiness for a given scenario. Calibration is considered dynamic, where trust adapts based on ongoing interaction with the system.
- **Emotional trust**: Trust that emerges from emotional connections, often involving the feeling of security users experience during interactions with the system.
- **Overtrust**: A situation where a user places too much trust in a system, exceeding the system's actual trustworthiness, potentially leading to harmful outcomes.
- **Distrust**: A scenario where a user's level of trust in a system is lower than the system's actual trustworthiness, which can lead to underuse of the system's capabilities.
- **Human-in-the-loop**: A control paradigm where human operators are actively involved in the decision-making processes of the system.
- **Human-on-the-loop**: A control paradigm where human operators monitor and oversee the system's operations but intervene only when necessary.
- **Human-in-command**: A control paradigm where human operators have overarching control over the system, including both decision-making and oversight capabilities.
- **Adversarial Attacks**: Deliberate manipulations aimed at exploiting the weaknesses of AI systems, often to produce erroneous outputs.
- **Differential Privacy**: A technique used to protect individuals' privacy by introducing randomness into the data analysis process, ensuring the data output is indistinguishable from queries with or without a particular individual's data.
- **Statistical Parity**: A fairness metric ensuring that the rate of favorable outcomes is the same across majority and minority groups.
- **Explainability (XAI)**: The degree to which a human user can understand the reasons behind an AI system's decisions or actions.