Computers in Human Behavior 170 (2025) 108671

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM)

Nadine Schlicker [a,*], Kevin Baum [b], Alarith Uhde [c], Sarah Sterz [d], Martin C. Hirsch [a], Markus Langer [e,**]

[a] Philipps University of Marburg, Germany
[b] Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany
[c] Ritsumeikan University, Japan
[d] Saarland University, Germany
[e] Albert-Ludwig University of Freiburg, Germany

ARTICLE INFO

Keywords:
Trust in automation
Trustworthiness
Trustworthy AI
Human-centered design
Calibrated trust

ABSTRACT

Designing trustworthy AI-based systems and enabling external parties to accurately *assess* the trustworthiness of these systems are crucial objectives. Only if trustors assess system trustworthiness accurately, they can base their trust on adequate expectations about the system and reasonably rely on or reject its outputs. However, the process by which trustors assess a system's actual trustworthiness to arrive at their perceived trustworthiness remains underexplored. In this paper, we conceptually distinguish between actual and perceived trustworthiness, trust propensity, trust, and trusting behavior. Drawing on psychological models of how humans assess other people's characteristics, we present the two-level Trustworthiness Assessment Model (TrAM). At the micro level, we propose that trustors assess system trustworthiness based on cues associated with the system. The accuracy of this assessment depends on cue relevance and availability on the system's side, and on cue detection and utilization on the human's side. At the macro level, we propose that individual micro-level trustworthiness assessments propagate across different trustors – one stakeholder's trustworthiness assessment of a system affects other stakeholders' trustworthiness assessments of the same system. The TrAM advances existing models of trust and sheds light on factors influencing the (accuracy of) trustworthiness assessments. It contributes to theoretical clarity in trust research, has implications for the measurement of trust-related variables, and practical implications for system design, stakeholder training, AI alignment, and AI regulation related to trustworthiness assessments.

1. Introduction

Trust is a much-discussed topic in algorithmic decision-making, especially in the area of artificial intelligence ("AI")[1] (Glikson & Woolley, 2020; Jacovi et al., 2021; Kaplan et al., 2023). In the development of trust, the process by which a human assesses the trustworthiness of a system, leading to their *perception* of trustworthiness, is crucial. We call this process the trustworthiness assessment process. Only if someone's trustworthiness assessment is accurate, they can base their trust on adequate expectations about the system's capabilities and limitations (Lee & See, 2004), and make informed decisions about their trusting behavior. In contrast, overestimating a system's trustworthiness can reduce vigilance and oversight (Hardré, 2016) and underestimating a system's trustworthiness may cause users to disregard valid system outputs (Hoff & Bashir, 2015). These issues are closely linked to research on achieving "calibrated trust," which aims to optimize the joint

This article is part of a special issue entitled: The social bridge: Trust published in Computers in Human Behavior.

* Corresponding author. Philipps-Universität Marburg, Institute for AI in Medicine, Baldingerstraße, Marburg, 35043, Germany.
** Corresponding author. Albert-Ludwigs-Universität Freiburg, Department of Psychology, Section Work and Organizational Psychology, Engelbergerstraße 41, Freiburg, 79085, Germany.
E-mail addresses: nadine.schlicker@uni-marburg.de (N. Schlicker), markus.langer@psychologie.uni-freiburg.de (M. Langer).

[1] In the remainder of this paper, we use the term (AI) system to refer to any kind of semi-automated and automated decision systems. This includes systems that produce software-based decisions such as embedded and cyber-physical systems (e.g., robots), algorithmic systems, and data-based systems (e.g., machine learning or deep neural networks).

https://doi.org/10.1016/j.chb.2025.108671
Received 19 June 2023; Received in revised form 20 September 2024; Accepted 14 April 2025
Available online 14 April 2025

performance of humans and AI systems (Muir, 1987; Cancro et al., 2022; Lee & See, 2004). However, instead of focusing on the heterogeneously defined and measured concept of "trust calibration" (Carter et al., 2023; Lee & See, 2004; Wischnewski et al., 2023) we tackle the conceptually upstream challenge of an accurate assessment of trustworthiness (Baer et al., 2018; Mayer et al., 1995; van der Werff et al., 2021) that we then connect to the concept of calibrated trust in section 5.2.

Despite its importance, to date there has been little theoretical effort to describe the trustworthiness assessment process. Thus, it remains unclear how people arrive at their perceptions of the trustworthiness of specific systems. Empirical research has shown that characteristics associated with the actual trustworthiness of a system (e.g., accuracy, fairness; High-Level Expert Group on Artificial Intelligence, 2019) influence people's perceptions of trustworthiness (Schelble et al., 2022). However, the fact that a system is actually trustworthy does not seem to be sufficient to develop high levels of perceived trustworthiness (Dietvorst et al., 2015; Dzindolet et al., 2003; Longoni et al., 2019). Conversely, systems that actually lack trustworthiness may still be perceived as trustworthy (Kocielnik et al., 2019; Merritt, Lee, et al., 2015; Papenmeier et al., 2022). Relatedly, different people may arrive at different perceptions of trustworthiness depending on the characteristics of the human-system interaction (Dietvorst et al., 2015), the person assessing the system (Merritt, Unnerstall, et al., 2015), and the situation (Longoni et al., 2019). For example, when two people assess system trustworthiness, one might utilize their first-hand experience with the system and the other might utilize certification labels by certification institutions (Ehsan et al., 2021; Jacovi et al., 2021; Knowles & Richards, 2021). Although the characteristics of the system remain the same, each person may end up with a different perception of the system's trustworthiness (Kay et al., 2015; Papenmeier et al., 2022).

In this conceptual paper, we aim to specify the trustworthiness assessment process, i.e. the process through which people assess a system's *actual trustworthiness* to form their *perceived trustworthiness* of the system. To this end, we introduce the Trustworthiness Assessment Model (TrAM) with two levels: the micro-level trustworthiness assessment process and the macro-level trustworthiness propagation process. At the micro level, we build on models from psychology that specify how humans assess characteristics of other humans that are not directly accessible (e.g., personality) (i.e. Funder's Realistic Accuracy Model (Funder, 1995) and Brunswik's Lens Model (Kirlik, 2006; Brunswik, 1956; Hammond & Stewart, 2001; Kuncel, 2018)) and translate them to the case of humans assessing artificial systems. At the macro level, we describe the trustworthiness propagation process that proposes that different stakeholders (e.g., system designers, certification institutions, end users) assess system trustworthiness to form their perception of trustworthiness. Following their assessment, they produce cues that can influence other stakeholders in their (micro-level) trustworthiness assessment of the system.

We see the following contributions of this paper. First, we aim for more conceptual clarity in trust research. This addresses growing criticism about heterogeneity in how trust-related terminology is used (Saβmannshausen et al., 2023; Vereschak et al., 2021), variety in the measurement of trust-related variables, and general criticism on trust as a research target (Bolton, 2022). We distinguish between actual and perceived trustworthiness, trust propensity, trust, and trusting behavior, which we believe needs to be made explicit to advance research on trust in AI systems (see also Vereschak et al., 2021). We discuss our proposed model in relation to these concepts and in relation to research on calibrated trust. Second, we extend established trust models (Lee & See, 2004; Mayer et al., 1995), by explicating the transition from a system's actual trustworthiness to a trustor's perceived trustworthiness. We define and refine key concepts of the trustworthiness assessment (actual trustworthiness, perceived trustworthiness, system characteristics, and individual standards) and highlight factors at the micro and the macro level in the trustworthiness assessment process that may influence the accuracy of trustworthiness assessments in interactions with systems and based on the social and societal contexts in which those systems are used. Third, we derive practical implications of the TrAM for system designers, for human trustors, and for legislation, especially in light of current developments toward regulation such as the European AI Act, standardization, and certification of AI systems (DIN, 2020; UNESCO, 2023; European Parliament, 2024; High-Level Expert Group on Artificial Intelligence, 2019; Hauschke et al., 2022). In conclusion, the TrAM offers a theoretical framework to systematically analyze the factors that may influence the accuracy of trustworthiness assessments at the micro and macro level.

2. On the relation of trust propensity, trustworthiness, trust, and trusting behavior

This paper examines the trustworthiness assessment process. We do not attempt to provide a comprehensive overview of all concepts related to trust. Instead, we explain how the TrAM adds to existing trust models, describe how the key concepts in the trust development process are related, and distinguish the concepts in the TrAM from concepts that may have been used interchangeably in research. Our propositions in this section mainly build on the propositions by Mayer et al.'s (1995) Organizational Trust Model.

Trust requires at least two parties: The trustor, the party that trusts (or does not trust), and the trustee, the party that is trusted (or is not trusted). In the dynamic process of trust development (as stated in Mayer et al., 1995) for organizational trust and adopted by Lee & See, 2004) for trust in automation), four concepts are particularly important: *propensity to trust, trustworthiness*, *trust,* and *trusting behavior* (see the right hand part within the grey dashed frame in Fig. 1).[2]

Propensity to trust in human interactions has been characterized as a "general willingness to trust others" [106, p. 715] and a "generalized expectancy that others can be relied upon" [5, p. 166]. Propensity to trust develops across the lifespan and may be particularly important in unfamiliar situations with unknown trustees (Baer et al., 2018). While the disposition to trust *humans* cannot be directly extrapolated to the disposition to trust *machines*, the literature on trust in automation also recognizes the notion of propensity to trust as a stable disposition (Hoff & Bashir, 2015; Jessup et al., 2019; Lee & See, 2004). For example, it has been defined as "an individual's overall tendency to trust automation, independent of context or a specific system" [64, p. 413]. Propensity to trust plays a major role in differences in the trust development process between trustors who experience the same system. Research suggests that propensity to trust is a relatively stable human disposition (Baer et al., 2018), which limits the opportunity to externally influence the trust development process through propensity to trust.

Trustworthiness has been used to refer to two sides of one coin (e.g., de Visser et al., 2014; Lee & See, 2004; Mayer et al., 1995; Schlicker & Langer, 2021; Liao & Sundar, 2022). First, trustworthiness has been referred to as an "objective attribute of the trustee" (Zerilli et al., 2022; see also Green, 2022; Jacovi et al., 2021; Kelp & Simion, 2022; McLeod and Zalta, 2021) for a similar conceptualization of trustworthiness as a property of the trustee). Second, trustworthiness has been referred to as a trustor's subjective perception of a trustee's attributes (Mayer et al., 1995). In this understanding, trustworthiness perceptions reflect a trustor's assessment of the trustee's abilities, principles, and intentions, as well as a trustor's beliefs about whether the trustee will help the trustor to achieve their goals (Baer et al., 2018; Colquitt et al., 2007; Lee & Moray, 1992; Lee & See, 2004; Mayer et al., 1995). Perceptions of trustworthiness are situational and subjective (Baer et al., 2018; Hoff & Bashir, 2015; Mayer et al., 1995). Situational means that a trustor

[2] For the purposes of this paper, we do not provide a comprehensive review of other factors that influence trust. We refer readers to papers that highlight these factors such as (Baer et al., 2018; Mayer et al., 1995; Rousseau, Sitkin, Burt, & Camerer, 1998).

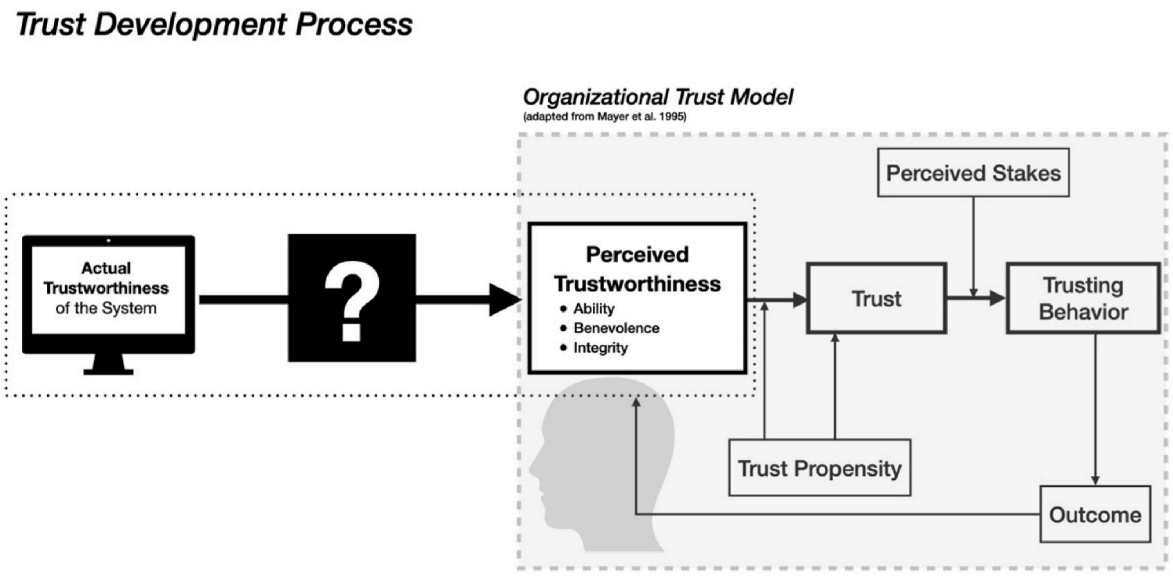

Fig. 1. The trust development process. We propose that the trust development process consists of (1) the process from a trustor's perceived trustworthiness to trusting behavior as described by the Organizational Trust Model by Mayer et al. (1995) (right side of this figure) and (2) the trustworthiness assessment process through which trustors reach their perceived trustworthiness given the actual trustworthiness of a system (left side of this figure). The latter is the theoretical gap the current paper aims to fill. Note that we have slightly adapted Mayer et al.'s figure and renamed what they called "risk taking in a relationship" to trusting behavior, and renamed what they called "perceived risk" to perceived stakes in order to emphasize that it is about the weighting of risks *and* benefits.

assesses a trustee's trustworthiness with respect to a specific task, in a specific context, and at a specific time (Mayer et al., 1995). Subjective means that whether a trustee is considered trustworthy depends on the trustor's cognitive and affective assessment in light of the trustor's individual goals, values, and abilities in a situation in which the trustor considers relying on the trustee (Muir, 1987; Chiou & Lee, 2021; McAllister, 1995).

Regardless of whether trustworthiness is considered an attribute of the trustee or a subjective perception of attributes of a trustee, research suggests that trustworthiness consists of multiple facets (Baer et al., 2018; Dietz & Den, 2006; High-Level Expert Group on Artificial Intelligence, 2019; Jacovi et al., 2021; Lee & See, 2004; Mayer et al., 1995). For instance, Mayer et al. (1995) propose that ability (a "group of skills, competencies, and characteristics that enable a party to have influence within some specific domain", p. 717), benevolence ("the extent to which a trustee is believed to want to do good to the trustor, aside from an egocentric profit motive", p. 718), and integrity ("the trustor's perception that the trustee adheres to a set of principles that the trustor finds acceptable", p. 719) subsume all facets of (perceived) trustworthiness in interpersonal trust. Lee & See (2004) translated these facets to trust in automation referring to performance ("the competency or expertise as demonstrated by its ability to achieve the operator's goals", p. 59), purpose ("corresponds to faith and benevolence and reflects the perception that the trustee has a positive orientation towards the trustor", p. 59) and process ("the degree to which the automation's algorithms are appropriate for the situation and able to achieve the operator's goals", p. 59). Instead of referring to "perceived trustworthiness", Lee and See refer to these facets as the "bases of trust" (p. 59). In order to clearly distinguish between the two uses of the term trustworthiness, we propose that on the side of the trustee, there is an *actual trustworthiness* (AT); on the side of the trustor, there is a *perceived trustworthiness* (PT) of the system.

Trust is influenced by PT (Baer et al., 2018; Colquitt et al., 2007; Mayer et al., 1995; Möllering, 2006). While we acknowledge the diversity of trust definitions (Muir, 1987; Lee & See, 2004; Rempel et al., 1985; Rousseau et al., 1998; Shin, 2019; Thielmann & Hilbig, 2015), we follow Mayer et al. (106, p. 712), who define trust 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".

Trusting behavior is the behavioral manifestation of trust (e.g., reliance, compliance (de Visser et al., 2020; Vereschak et al., 2021), risk-taking (Mayer et al., 1995)). Although trust and trusting behavior can be positively related (Körber, 2019), the stakes associated with the actual decision to engage in trusting behavior toward a trustee prevent trust from translating directly into trusting behavior. More precisely, any decision to actually engage in trusting behavior is associated with an expected positive outcome (benefits), but also with a possible negative outcome (risks), i.e. the *perceived stakes*. For example, even if the degree of trust is high, if the perceived risks are also high, trustors may decide to perform a task themselves (Buçinca et al., 2020; Yin et al., 2019; Zhang et al., 2020). Furthermore, sometimes a trustor's behavior may appear to reflect trusting behavior, but may actually reflect the consequences of situational factors, such as time pressure (Rieger & Manzey, 2022) or social conformity (Asch, 1955). Nevertheless, we expect perceptions of trustworthiness, trust, and trusting behavior to be positively related (Vereschak et al., 2021).

Fig. 2 shows the conceptual relationship between AT and PT, as well as trust and trusting behavior. The above summary of the main concepts in trust research reflects our understanding of the research on trust (in automation and AI) and is the basis for the argumentation in this paper. We propose that whether human's trust and trusting behavior are well-grounded (McLeod and Zalta, 2021) (i.e. lead to an optimization of expected outcomes) depends heavily on the characteristics of the trustee and their relation to the goals of the trustor. Therefore, it is crucial to understand the conceptually upstream trustworthiness assessment process (i.e. the transition between AT and PT) and the factors that influence the accuracy of the trustworthiness assessment process. This is crucial because research indicates that the trustworthiness assessment process is prone to inaccuracies, given that even the assessment of seemingly objective system characteristics such as system accuracy varies between observers (Dzindolet et al., 2002; Madhavan & Wiegmann, 2007; Papenmeier et al., 2022; Rieger et al., 2022). Inaccurate trustworthiness assessments also become evident in research on the automation bias phenomenon (i.e. that people initially tend to overestimate the trustworthiness of systems for certain tasks (Merritt, Unnerstall, et al., 2015; Parasuraman & Manzey, 2010) and in research

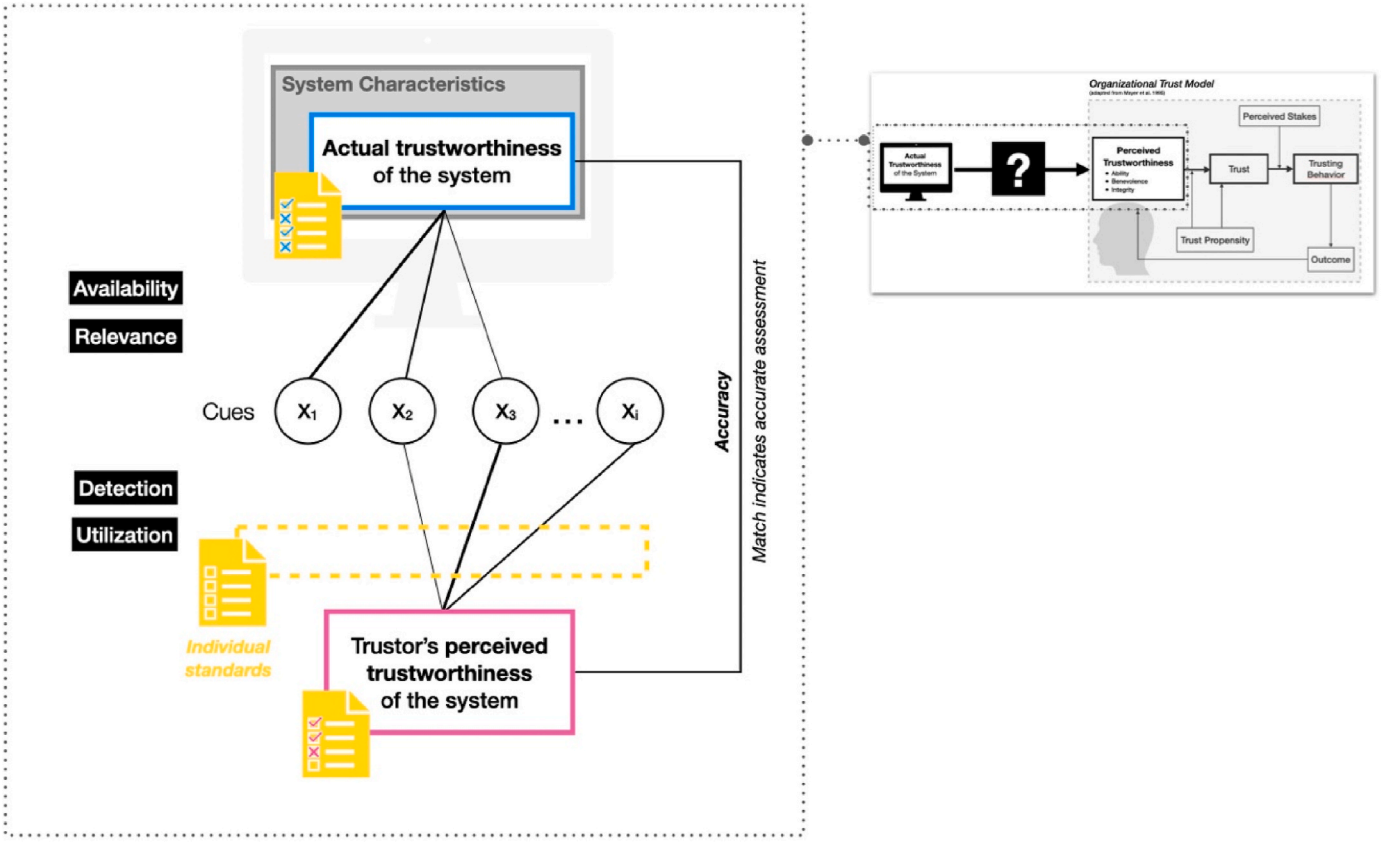

Fig. 2. The micro level of the TrAM showing the relation between actual trustworthiness (AT) and perceived trustworthiness (PT). The location of the trustworthiness assessment process in the model by (Mayer et al., 1995) is indicated by the grey dotted frame (showing where Fig. 2 zooms into Fig. 1). The relation between AT and PT via cues is based on the ideas of the Brunswik Lens Model (Kirlik, 2006; Hammond & Stewart, 2001) and the factors influencing an accurate trustworthiness assessment (black boxes) are based on Funder's Realistic Accuracy Model (Funder, 1995). AT manifests in cues and trustors use cues to form their PT of a system (see also Fig. 3). The line thickness indicates on the side of the system how relevant a cue is for the system's AT, and on the side of the trustor how heavily a cue is weighted by the trustor to form their PT. The yellow list icons reflect our requirement list metaphor (introduced in section 3.2), which indicates that individual standards (yellow dashed frame, empty list), actual trustworthiness (list filled out in blue), and perceived trustworthiness (list filled out in purple) are related. The individual standards act as a requirement list for a trustworthy system from the trustor's perspective. They influence which cues are searched for and how they are utilized. For more detail on the relation between individual standards, PT and AT see Fig. 4. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

on algorithm aversion (i.e. that people seem to underestimate the trustworthiness of systems after seeing the system fail (Dietvorst et al., 2015).

Previous research has sometimes distinguished between AT and PT, and has also emphasized the importance of accurately assessing AT (Muir, 1987; de Visser et al., 2020; Lee & See, 2004). However, to the best of our knowledge, there is no research that explicitly describes the trustworthiness assessment process and the factors that influence the accuracy of trustworthiness assessments. In the remainder of this paper, we aim to fill this gap: We shed light on the process that links AT and PT and the factors that influence the accuracy of the trustworthiness assessment at the micro and macro level.

3. The micro level – the trustworthiness assessment process

3.1. Foundations and overview of the micro-level trustworthiness assessment process

The micro level of the TrAM builds on models from psychology that describe human judgment processes (Brunswik, 1956), that outline how people assess other people's traits (Kirlik, 2006; Hammond, 1996; Kuncel, 2018), and that propose what factors influence the accuracy of these assessments (e.g., Funder, 1995). It describes the process that lies outside of existing trust models, but reflects a central basis for these models, that is, the formation of a trustor's PT through their assessment of a trustee's AT (see Fig. 2). Characteristics of other people that are to be assessed are often latent constructs. For instance, this applies to cognitive abilities or personality. Both can neither be directly observed or measured to obtain their "true value" (i.e. their ground truth) (Borsboom et al., 2003). Nevertheless, we, as a society, define them (e.g., early concepts for cognitive abilities; Spearman, 1961), refine these definitions (e.g., later concepts of cognitive ability; Carroll, 1993), operationalize them (e.g., in written tests or questionnaires of cognitive ability; Kaufman et al., 2006; Laurent et al., 1992; Prewett, 1995) and thus assess these latent constructs more or less reliably.

We also assess latent constructs in interactions with people (Funder, 1995). For example, we observe that a person with a colorful wardrobe is talkative and assess that this person appears to be extraverted. We interpret available cues (e.g., a person's behavior and their clothing) to assess a person's actual characteristics (e.g., actual extraversion) resulting in our perceived characteristics (e.g., perceived extraversion) of that person. This assessment varies in its accuracy depending on (1) the characteristics of the target, (2) the cues available for assessing the target's characteristics, and (3) the observer assessing the target's characteristics (Funder, 1995). Regardless of its accuracy, the assessment and the corresponding expectations can influence the observer's subsequent assessments and their behavior toward the target (Goffman, 2006; Funder, 1995; Human & Biesanz, 2013; Rosenthal & Jacobson,

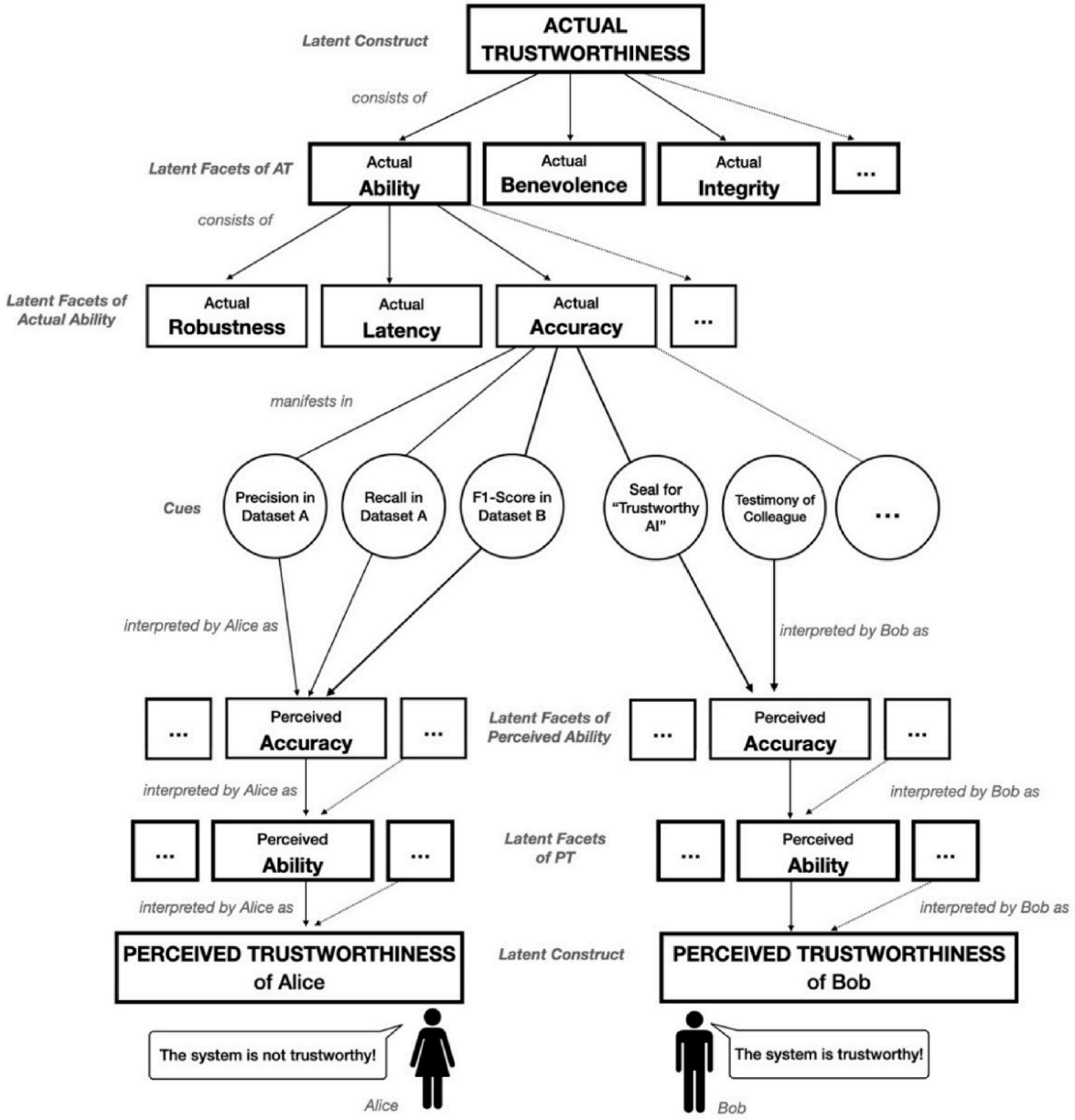

Fig. 3. An example of the assumed relations between the actual trustworthiness (AT), the latent facets that contribute to it, and the observable cues that form the interface for different trustors (in this example Alice and Bob) to infer their perceived trustworthiness (PT) of the system. The facets that make up AT and (PT) are similar, but the cues used to infer PT may differ between trustors (even if they have similar standards as assumed in this example). The cues are of different relevance to assess the (facets of) system's AT, which is indicated by the different line thickness. In this example, the cue "testimony of colleague" is not related to the actual accuracy of the system (indicated by the absent line), but Bob uses it to assess the system's AT. Note, that this should not imply that testimonies of others are always irrelevant cues.

1968). For example, underestimating someone's cognitive abilities might lead to suboptimal task assignment in a work context.

We apply this to a trustor's assessment of a system's trustworthiness. Specifically, in the micro-level trustworthiness assessment process (Fig. 2), we propose that trustors cannot directly assess a system's AT. Instead, they assess a system's AT on the basis of available cues. These cues are, to some extent, relevant to the assessment of a system's AT. Human trustors then need to detect and utilize certain (and possibly many) cues to arrive at their PT. Consequently, the degree to which a trustor's PT matches a system's AT depends on the *availability* and *relevance* of cues on the system's side and on the *detection* and *utilization* of cues on the human's side. We propose that this relationship between AT and PT sets the stage for everything that follows in the trust development process (e.g., in terms of well-grounded trust and trusting behavior).

3.2. Main concepts of the trustworthiness assessment model (TrAM)

3.2.1. Actual trustworthiness (AT)

We define a system's AT as a latent, i.e. not directly observable, construct that indicates the true value of a system's trustworthiness (in the sense of the Realistic Accuracy Model; Funder, 1995). AT consists of several facets. For example, benevolence, integrity, and ability are treated as facets of AT (High-Level Expert Group on Artificial Intelligence, 2019; Lee & See, 2004; Mayer et al., 1995). However, there is typically no way to perfectly measure these individual facets (e.g., because only a subset of all available data can be accessed to assess system performance), nor to accurately assess how they combine to AT (Hoff & Bashir, 2015; Lee & See, 2004). Most or even all of these facets of AT will consist of subfacets (Hoff & Bashir, 2015; Lee & See, 2004). For instance, system accuracy will be a facet of a system's ability. Because

Fig. 4. The requirement list metaphor describes the relationship between individual standards (empty list on the left), actual trustworthiness (AT, list filled out in blue on the top right), and perceived trustworthiness (PT, list filled out in purple on the bottom right). In this case, since the list is filled out differently for the PT than for the AT, the accuracy of the trustworthiness assessment is suboptimal.[31] Ticked boxes indicate that a standard is met, crossed out boxes indicate that a standard is not met, and empty boxes indicate uncertainty regarding a given standard. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

system accuracy itself is also not directly measurable, people use manifest variables, i.e. observable cues, to assess system accuracy. Cues for system accuracy could be accuracy metrics presented in the system's manual (e.g., F1-scores, precision, recall) (the upper part of Fig. 3 illustrates the relations between the latent construct AT, its facets, and the observable cues; Lee et al., 2022).

To further specify AT, it is necessary to elaborate on two concepts in our model that affect AT: *individual standards* for system trustworthiness and *system characteristics*. What constitutes a trustworthy system differs between trustors and depends on their *individual standards* for system trustworthiness (Muir, 1987; Gabriel, 2020; Knowles & Richards, 2021). Individual standards answer the question: "*What makes a system trustworthy for me*?" As a metaphor, we can think of individual standards as a requirement list that contains all the facets that make up a perfectly trustworthy system for the trustor, as shown in Figs. 2 and 4. We use the requirement list metaphor throughout the manuscript to illustrate the main concepts of the model.

Individual standards are influenced by trustors' goals and interests, the culture in which trustors grew up (e.g., collectivistic vs. individualistic cultures; Awad et al., 2018), ethically and socially motivated values (e.g., cooperation vs. competition) (Textor et al., 2022), as well as the normative and regulatory framework in which trustors operate (e.g., under the influence of the European General Data Protection Regulation). Individual standards are not arbitrary: they refer to a common understanding of the concept of trustworthiness of systems, which is also influenced by public discussions about trustworthy AI (Drobotowicz et al., 2021; High-Level Expert Group on Artificial Intelligence, 2019; High-Level Expert Group on Artificial Intelligence, 2020). Thus, individual standards could be clustered by the facets that are assumed to constitute trustworthiness (e.g., ability, benevolence, integrity; Mayer et al., 1995). However, the specification of these facets of trustworthiness is likely to differ between different (groups of) individuals. For instance, the specification of a system's ability depends on the users' goals (e.g., screening vs. decision making), interests (e.g., saving money vs. preserving data privacy), and their personal alternatives to the system (e.g., one's own ability to perform the task at hand).

Differences in the individual standards may also result from the granularity (from general to detailed) of the trustor's standards. For example, when considering fairness, a general individual standard, held by Bob might be that "a system should be fair", while a more detailed individual standard held by Alice might be that "a system should treat everyone equally, regardless of their individual attributes" (Deutsch, 1975; Wachter et al., 2021). While Alice has an explicit standard of fairness (i.e. equality), Bob might implicitly define "fair" in terms of equity, where those who contribute more receive more (Deutsch, 1975). As a result, Alice and Bob might assume that they agree that systems should be fair, without realizing that they would disagree on the details (Ross et al., 1977).

Individual standards have a top-down influence on what cues a trustor searches for and how these cues are utilized. For instance, if Alice cares about data privacy, making it an important individual standard for her, she might actually read the terms and conditions, searching for cues that serve as evidence for trustworthiness with respect to data privacy standards. In contrast, Bob may care less about data privacy and thus simply accept the terms of service.

We refer to *system characteristics* as (often only theoretically available) context-free facts that subsume everything that could theoretically be known about a system. System characteristics answer the question: "*What are the characteristics of this system?*" on a descriptive level. For example, system characteristics may contain information such as the

functionality, robustness, and reliability of the system (in a specific data set), but also the user-interface design, the company behind it, and the data used for training and testing. It is important to distinguish that system characteristics in this (only theoretically existing) context-free space are also *objective* in the sense that they do not depend on the trustor. In contrast, AT is always subjective in the sense that it is relative to the trustor's individual standards of trustworthiness.

To summarize, a system's AT depends on the system characteristics and on the individual standards of trustworthiness. Specifically, AT reflects the degree to which the system characteristics match a trustor's individual standards for trustworthiness with respect to a particular task and point in time. Thus, AT answers the question: "*How trustworthy is the system actually with respect to my individual standards*?" In the requirement list metaphor, AT is a function of the respective ticked and crossed out boxes on the "individual standards requirement list" if a perfect assessment was possible (see Fig. 4).

To further clarify the distinction, *system characteristics* answer questions such as "What is the accuracy of an AI system in a particular data set?"; *AT* answers questions such as "Is the AI system's accuracy high enough?" What is high enough may be different for different trustors. For example, in medical diagnostics, evaluating system accuracy may require a comparison with how well human experts perform on the same task. While a novice user may perform significantly better *with* the system, an expert user may perform better *without* the system (Wei et al., 2022). Consequently, when assessing the trustworthiness of a system by reference to the individual standard: "Will the system produce better results than I would?", the novice user might judge 85 % accuracy as trustworthy, while the expert might judge the system as trustworthy only if it achieves 95 % accuracy. System characteristics might also be different in reference to the trustors' characteristics, even if their individual standards are similar (Birhane, 2022; Benčević et al., 2024; Keller et al., 2022). For example, two trustors (one with light skin and one with dark skin) with the same standards (e.g., the AI system should classify a melanoma with 90 % accuracy), might encounter different AT's of the system, due to the system's different capabilities of classifying melanoma on light (95 % accuracy) vs. dark skin (80 % accuracy).

The AT of a system can be influenced by both, changes in the system characteristics and by changes in the individual standards. Changes in the system characteristics can result from improvements to the system, for instance, by optimizing its technical functionalities (e.g., algorithmic fairness (Bellamy et al., 2019; Li et al., 2021; Serna et al., 2022), robustness (Cisse et al., 2017; Xie & Wu, 2019), explainability (Setzu et al., 2021; Ribeiro et al., 2016), and safety (Toreini et al., 2020)). Changes in the system characteristics may also be due to changes in the environment. Consider a machine learning-based decision support system for detecting different types of viruses. As the virus mutates, it reduces the accuracy of the algorithm and thus changes its system characteristics (Jalaian et al., 2019; Hoens et al., 2012).

Changes in the individual standards may be due to new experiences or knowledge gained by the trustor (Flathmann et al., 2021; Hoff & Bashir, 2015; Textor et al., 2022). For instance, a trustor's standards of system accuracy may increase, because they have improved their own skill at a task. Individual standards may become more nuanced because trustors' understanding of different accuracy metrics may change, and as a result, their individual standard may evolve from "the system should be accurate" to "the system should have high specificity". Trustors could also lower their individual standards, which would increase the system's AT (Toreini et al., 2020). Lowering individual standards might be useful, if trustors have unrealistic expectations about system capabilities (e.g., expectations that systems perform near perfection; Madhavan & Wiegmann, 2007).

3.2.2. Perceived trustworthiness (PT)

PT reflects the result of a trustor's assessment of the trustee's AT (Baer et al., 2018; Mayer et al., 1995). Thus, PT is reflected by the question "*How trustworthy do I think the system is with respect to my individual standards?*" Speaking in our requirement list metaphor, PT concerns how the trustor fills out their requirement list reflecting their individual standards. Specifically, there are three possible states of the boxes on the requirement list depending on whether a system meets an individual standard (i.e. the perceived *presence* of trustworthiness reflected by the ticked boxes), does not meet an individual standard (i.e. the perceived *absence* of trustworthiness reflected by the crossed-out boxes), and whether the trustor is uncertain about the status of the box (i.e. the perceived uncertainty reflected by the empty boxes) (see Fig. 4). In most cases, after the trustworthiness assessment, there is still uncertainty about the trustee's AT. This means that trustors may not be able to mark (i.e. tick or cross out) some of the boxes after the assessment, or they may be unsure about whether they should mark certain boxes.

PT is another latent construct that can only be measured indirectly, as is common in research assessing people's perceptions of system trustworthiness (e.g., by asking people to report on their perceived trustworthiness of a system (Alam & Mueller, 2021; Jian et al., 2000; Madsen & Gregor, 2000; Körber, 2019; Rieger et al., 2022), or by observing people's interactions with systems (Vereschak et al., 2021; Zhang et al., 2020)). The facets that contribute to overall PT are similar to those that contribute to overall AT (see Fig. 3). PT may consist of the same general facets (e.g., ability, integrity, benevolence) and same subfacets (e.g., accuracy for ability) as AT.

3.2.3. Cues

Cues are the interface between AT and PT. According to de Visser et al., 2014 , a cue is an "information element that can be used to make a trust[worthiness] assessment about an agent" (term in square brackets added by the authors). Cues are pieces of evidence that presumably provide insight into the AT of a system (de Visser et al., 2014; Cancro et al., 2022; Liao & Sundar, 2022). Single cues may provide only narrow or even misleading insights into the AT of a system, and each cue is related to the AT to some degree. Thus, trustors are constantly (consciously or unconsciously) searching for, encountering, and interpreting cues to assess the AT of a system. Trustor's PT is determined in part by how they actively and passively select, interpret, and weigh cues. In line with this, research has investigated how different cues are weighted and how they are associated with trust in decision aids (Conway et al., 2016).

A variety of cues can be used to assess system trustworthiness. For example, cues can be the aesthetics of a user interface, information included in the user manual, the time it takes to produce an output, a blue screen, information about the inputs a system uses, specific outputs of a system, the reported accuracy of a classifier, information about the uncertainty associated with a classification output, an explanation for the system's recommendation, a label indicating trustworthy AI, or the logo of a company (Scharowski et al., 2023; de Visser et al., 2014; Cancro et al., 2022; Liao & Sundar, 2022; van der Werff et al., 2021). Cues can also come from other people (e.g., employee testimonials, see Section 4). Providing a comprehensive list of possible cues is beyond the scope of this paper, but research has already proposed design recommendations and taxonomies to classify cues. For instance, de Visser et al., (2014) provide a "trust cue design taxonomy". Cues mentioned by the authors are, for example, confidence levels that a system provides for its outputs, or information about goals and purposes of the system. Among other criteria, the authors organize these cues according to different "bases of trust" (e.g., intent, performance, process). We agree that what these authors discuss are manifest cues for assessing a system and its characteristics, but rather than calling the cues "trust cues", we suggest to call them *trustworthiness* cues, since they are cues for assessing facets of system trustworthiness. Similarly, Cancro et al. (2022) provide a taxonomy of system information (i.e. cues) that according to the authors can be used for "trust calibration". Examples of cues in their work are explanations that provide information about a system's uncertainty, the rationale for a system's actions, and its past performance under

similar conditions. Again, we agree that these are all cues that help to assess a system, but instead of enabling "trust calibration", we say that such cues aim to enable an accurate *trustworthiness assessment*.

3.3. Relations between the model components and influencing factors

Fig. 2 shows the relations between the micro-level components of the TrAM. Trustworthiness assessment is accurate when the trustor's PT matches the system's AT. According to Funder Funder, (1995), this accuracy depends on the *relevance* and *availability* of cues on the side of the system and on the *detection* and *utilization* of cues on the side of the trustor.

On the system side, cue relevance and availability determine how accurately AT can be assessed from the cues. *Cue relevance* defines how indicative a cue is for the AT of a system. Relevant cues correlate strongly with the (facets of) AT of a system. For instance, a relevant cue for AT may be information about a system's performance on a task. In contrast, less relevant cues correlate weakly with AT. A less relevant cue may be the popularity of a brand (Jacovi et al., 2021; Roy et al., 2001). In order to make an accurate and informed assessment, relevant cues must be available. In some cases, cues may also be uncorrelated with AT, although trustors may use them to form their PT. For example, Hoff and Bashir (2015) mention that increasing the anthropomorphism of automation can promote greater trust, which has been shown to be related to trustworthiness assessments and trust in systems (Glikson & Woolley, 2020; Schaefer et al., 2016). However, any system could be designed to be anthropomorphic, regardless of its actual capabilities. Thus, anthropomorphism may be a cue that is not related to the AT of a system, but it may be a cue that is often utilized by trustors and strongly influences the PT of the system.

Cue availability refers to the fact that cues can only be detected if they are accessible to the trustor. For example, the quality of a training data set could be a relevant cue. However, users may not have access to information about the training data. There are ideas in the literature that are consistent with the need to consider ways to make relevant trustworthiness cues available to human stakeholders. For instance, model cards (Mitchell et al., 2019) and fact sheets can highlight the purpose, algorithm, training and test data sets, potential biases, and model development of the system (Arnold et al., 2019; Baracaldo et al., 2022).

On the trustor's side, *cue detection* and *cue utilization* determine how strongly cues relate to trustor's PT and thus affect the accuracy of the assessment of AT. *Cue detection* means that relevant and available cues must be detected by the trustor. Possible factors influencing cue detection are trustor's mood and emotions (Bechara et al., 1997; Bower, 1981; Forgas, 1995; Merritt, 2011), attentional capacity (Hawkins et al., 1990), situation awareness (Endsley, 2017), time pressure (Rieger & Manzey, 2022), or their experience with a system (Thompson et al., 2008). In addition, a trustor's individual standards could top-down guide their attention to and detection of trustworthiness cues (Geyer & Müller, 2009; Jensen et al., 2011). Beyond that, user interface properties such as low contrast could make cue detection more difficult. For example, information about the training data set could be stored behind a hyperlink with low visibility. In this example, relevant information is available, but it may not be detected.

Cue utilization means that trustors must correctly interpret a relevant, available, and detected cue. In other words, even when relevant information is available and detected, trustors must weigh this information appropriately. Research has provided evidence for inaccuracies of trustor's PT due to suboptimal cue utilization. For instance, perceptions regarding system accuracy vary with the perceived difficulty of the system's task, implying that contextual information may influence cue utilization (Papenmeier et al., 2022). Furthermore, lay people's PT of a system appears to be more strongly influenced by a small sample of personal experiences with a system than by information from large-scale testing (Dietvorst & Bharti, 2020; Rechkemmer & Yin, 2022; Yin et al., 2019). There may also be sequence effects that influence cue utilization. For instance, the first-error effect (Bahner et al., 2008) suggests that system errors (as a cue for the absence of system trustworthiness) are weighted more heavily when they occur at the beginning of a human-system interaction than when they occur later in the interaction.

Cue utilization may be influenced by cognitive biases or implicit attitudes toward automation or AI (Langer et al., 2022; Merritt et al., 2013, 2015a). For instance, such effects have been shown in research on the automation bias as a phenomenon associated with overtrust in systems (Parasuraman & Manzey, 2010) and in research on algorithm aversion as a phenomenon associated with undertrust after seeing systems fail (Dietvorst et al., 2015). Finally, the domain, task, and system knowledge of the human trustor might influence cue utilization (Zhang et al., 2020). For example, correctly interpreting whether a system is failing seems to be influenced by the trustor's ability to perform a task without assistance (Merritt, Lee, et al., 2015). In contrast, little domain-, task-, and system knowledge (which could also be reflected in rather unspecific individual standards) may lead to inadequate utilization of cues, which may lead to inaccurate assessments of system's AT (e.g., assuming that a high-quality user interface or the logo of a popular company indicates high system accuracy; Bansal et al., 2019; Jacovi et al., 2021; Koh & Sundar, 2010).

Funder (1995) emphasizes that relevance, availability, detection, and utilization all determine accurate assessments of target characteristics. If only irrelevant cues are available, this will prevent an accurate assessment of AT. If no cues are available or detected, it is difficult to accurately assess system trustworthiness. Finally, improper utilization of relevant, available, and detected cues will result in a low accuracy of the trustworthiness assessment.

4. The macro level – the trustworthiness propagation process

The assessment of system trustworthiness does not occur in isolation. It is embedded in a societal and a social context, and many stakeholders are involved in the trustworthiness assessment before, for example, an end user assesses a system's AT. In line with this, previous work has proposed a broader view when addressing issues of trust in AI systems (Chiou & Lee, 2021; Knowles & Richards, 2021; Papagni et al., 2022; Liao & Sundar, 2022). For instance, research has suggested that certification bodies and expert communities are important mediators in the development of public trust in AI systems – and thus in the development of individual trust in AI-based systems (Knowles & Richards, 2021). Other research has highlighted the importance of social cues and heuristics in assessing system trustworthiness (Ehsan et al., 2021; Liao & Sundar, 2022). For example, seeing others interact with or talk about a system, can strongly influence trustworthiness assessments.

To understand how the (accuracy of a) stakeholder's trustworthiness assessment of a system is influenced by the trustworthiness assessment of other stakeholders, we propose that the macro-level trustworthiness assessment of a system can be viewed as a net of micro-level trustworthiness assessments of different trustors, each of which provides new trustworthiness cues to other trustors (see Fig. 5). We call the process of linked trustworthiness assessments and cue provision and reception in this graph the *trustworthiness propagation process*.

4.1. Overview of the macro level of the trustworthiness assessment model

Throughout the trustworthiness propagation process, different stakeholders assess a system's AT to arrive at their PT. For instance, stakeholders along the system life cycle may be system designers, certification bodies, auditors, supervisors, and end users. These stakeholders utilize cues, arrive at their PT based on these cues, and as a result may produce new cues (e.g., certificates, system manuals, testimonials) for other stakeholders that assess system trustworthiness (see Fig. 5). This implies that trustworthiness cues arise not only from the system itself, but also from the stakeholders who assess the AT of a system with respect to their individual standards. Given that AT likely differs

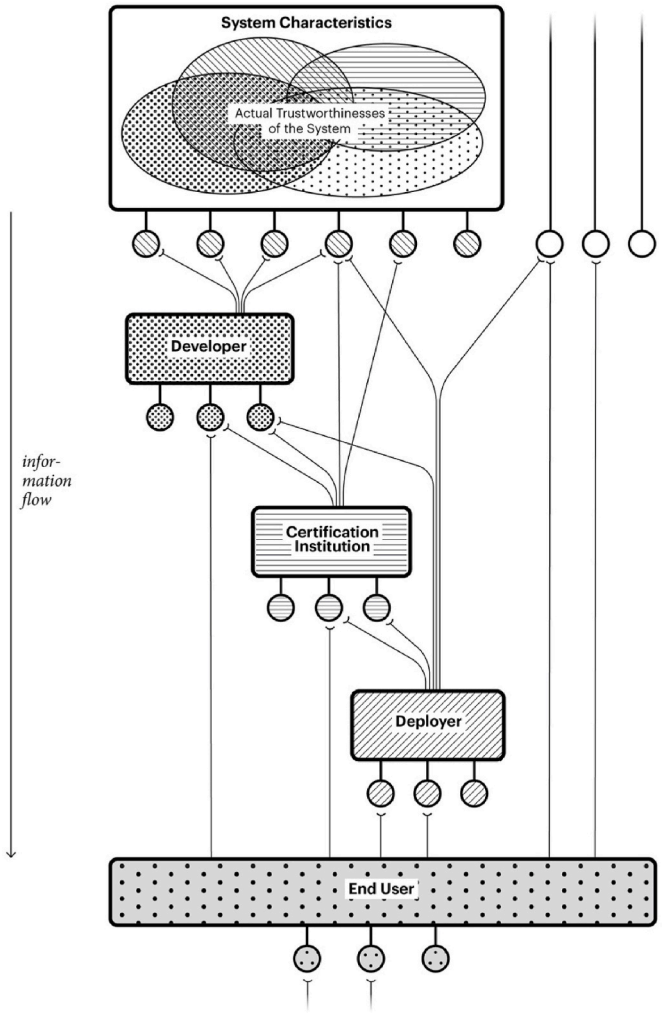

Fig. 5. The macro-level trustworthiness propagation process. Various stakeholders use cues (circles) from different origins (different shadings of the cues) to conduct micro level trustworthiness assessment processes. The cues they use can stem from the system (primary cues) or from other stakeholders (secondary cues). After their micro-level assessments, the stakeholders produce secondary cues for other stakeholders to use in their trustworthiness assessment. This figure shows a sample trustworthiness propagation process that could continue (which is indicated by the blurred line below "End User"). The differently shaded ovals in the system characteristics box indicate that there are as many actual trustworthinesses as there are trustors along the trustworthiness propagation process.

between stakeholders assessing the same system, there are as many ATs as there are trustors assessing the system and those ATs may overlap to varying degrees (see Fig. 5).

An example of a trustworthiness propagation process may start with a system designer assessing the AT of a system they have developed. Their assessment may produce new cues (e.g., performance metrics, a summary of system functionalities). Next, a certification institution may assess this system to allow it to enter the market. For its assessment, the certification institution might use the cues provided by the system, as well as cues provided by the system designer. As a result of its assessment, the certification institution may produce new cues, for example a certification label (Scharowski et al., 2023). Those cues can then be used by managers and product adopters to form their PT of the system. This trustworthiness propagation process may reach an end user who uses all available cues to form their PT of the system. Yet, the trustworthiness propagation process can continue. For example, end users may provide cues to other end users or to management, who then update their assessment of system trustworthiness. Any stakeholder in the process can provide cues to system designers who may, for example, take a bug report as a cue to lower their confidence in the system's ability and thus their PT. As a consequence, the designers may improve the system, and thus change the system characteristics.

The trustworthiness propagation process is thus a sequence of different trustors conducting the micro-level trustworthiness assessment process and producing secondary cues for other stakeholders to use in their trustworthiness assessment process. It is not a pipeline, nor does it involve a strict hierarchy; rather, it takes place in a complex social network of stakeholders. In the next section, we describe the trustworthiness propagation process and the possible trustors involved.

4.2. *The trustworthiness propagation process*

The two main ideas of the trustworthiness propagation process are (1) that there exists a net of stakeholders who form their own PT of the system through micro-level trustworthiness assessment processes, and (2) that these stakeholders produce, what we call, *secondary cues* that can then be used by other stakeholders to form their PT of the system. In contrast to *primary cues* that stem directly from the system, *secondary cues* result from another stakeholder's assessment of the system's trustworthiness. Secondary cues convey information about the system that is "colored" by the individual standards of the stakeholders who produced those cues.

Different stakeholders may have different approaches to accessing primary cues. For instance, a system designer might test a system to obtain performance information about the system such as accuracy, precision, and recall within a given data set – cues that would be primary cues. Similarly, a certification institution may conduct its own testing of the system as a part of its audit process and may receive similar primary cues from the system. However, the certification institution may also receive information from the system designer about the accuracy of the system – a secondary cue. In this case, the information may have been limited to accuracy only because the designer decided that information about precision and recall did not need to be available. In our parlance: potentially relevant cues may not be available to the certification institution.

Secondary cues are important in increasing the efficiency of the micro-level trustworthiness assessment process. Conducting a thorough, attentive, and intentional micro-level trustworthiness assessment process can be challenging and time-consuming. Secondary cues can streamline the assessment by indicating that another trustor has already completed the trustworthiness assessment. This aligns with research arguing that most people may lack the expertise to evaluate a system's documentation and instead rely on expert judgements (Knowles & Richards, 2021), known as "trust by proxy" (Jacovi et al., 2021; Laux et al., 2023). This concept is also reflected in the notion of "social trust" or "trust in trust" in research on digital information on the internet, which suggests that people rely on reviews and references that reflect others' assessments of the original information (Kelton et al., 2008). A prototypical example of a secondary cue is the provision of labels (e.g., labels used for certification or standardization; DIN, 2020; Zicari et al., 2021).

Beyond efficiency, secondary cues can also strongly influence the accuracy of trustworthiness assessments (Guffey & Loewy, 2012; Ross et al., 1977). For example, secondary cues can be intentionally inaccurate (e.g., the provision of intentionally deceptive cues, such as fraudulent data in the Volkswagen emissions scandal; Baum, 2016, pp. 633–647; Jung & Sharon, 2019). Secondary cues can also be unintentionally inaccurate due to the secondary cue provider's inaccurate assessment of system trustworthiness (e.g., a colleague's uninformed report about a system). Furthermore, secondary cues may be intentionally incomplete because stakeholders are trying to communicate efficiently (e.g., stakeholders may only communicate a single system performance measure) or because they are not allowed to communicate certain information (e.g., due to intellectual property or privacy

regulations).

Stakeholders' individual standards shape their assessment of system trustworthiness and, consequently, their provision of secondary cues. In other words, Alice's standards for a trustworthy system may not match Bob's standards. Yet, Bob may believe that Alice's trustworthiness assessment process includes similar standards. If Bob heavily utilizes secondary cues provided by Alice, and if Alice's individual standards do not match Bob's, he might mistakenly believe that the system is trustworthy for him. Bob and Alice are not aligned in terms of their individual standards. As such at least one of them is also not aligned with the system.

Given the importance of secondary cues in the trustworthiness assessment process, and given the potential misalignment of individual standards: (Gabriel, 2020) it is important to try to understand each stakeholder involved in the trustworthiness propagation process, their individual standards, and the secondary cues they may produce. We will now consider examples of stakeholders and their possible individual standards and the secondary cues they provide. The following is not intended to be comprehensive; it is intended to provide an intuition about the relevance of the trustworthiness propagation process.

The first stakeholders in the trustworthiness propagation process are often the **system designers**. System designers initiate the development process of a system. They are in a continuous feedback loop with the system and have the opportunity to change the system if it does not meet their individual standards.

Individual standards: System designers may have a specific use case in mind when developing the system. For example, this could be classifying heart sounds in medicine (Melms et al., 2023; Thompson et al., 2019) or picking the best applicant in hiring (Langer et al., 2021). Designers' individual standards for a trustworthy system might include stable performance (Friedler et al., 2021; Hauer et al., 2021; Parikh et al., 2019), combined with specific requirements in the labeling process, a higher performance than the current state-of-the-art AI system on a benchmark data set (Oliveira et al., 2021; Russakovsky et al., 2015), and robustness.

Secondary cues: Following their assessment and with reference to regulatory requirements, system designers will provide secondary cues such as a documentation (Knowles & Richards, 2021), a summary of their training, validation and test data, accuracy metrics, limitations and capabilities of the system, information presented in the user interface, and model cards (Mitchell et al., 2019).

At least for the foreseeable future, market approval for AI systems, especially of those used in high-risk contexts, is likely to be granted only if the system meets a list of requirements consistent with standards agreed upon within a particular institutionalized community of values. Consequently, the next stakeholder in the trustworthiness propagation process might be a **certification institution**.

Individual standards: The standards of certification institutions are part of an ongoing debate about the market authorization of AI. Standards are discussed and agreed upon at (inter)national levels. For instance, the European High-Level Expert Group on Artificial Intelligence (2019) attempted to define trustworthiness from a European perspective, highlighting various facets of what might constitute trustworthy AI (e.g., transparency, robustness, diversity, non-discrimination, and fairness). In addition, various ethical guidelines attempted to define what constitutes AT through a societal and scientific debate (Bærøe et al., 2020; Drobotowicz et al., 2021; Floridi, 2019; Morik et al., 2021; Zicari et al., 2021). The outcome of these debates sets the criteria that must be met for a trustworthy system. Note that this means that the same system could be considered as not trustworthy enough by a certification institution in the EU and trustworthy in, say, the US. The system applying for market approval will then be validated against these criteria by people representing the respective certification or approval body.

Secondary cues: Secondary cues provided by the certification institution may include reports, certification labels (e.g., indicating the "trustworthiness" of the system), and system disclaimers (e.g., indicating possible limitations and risks associated with using a particular system).

After market authorization, the next stakeholder in the trustworthiness propagation process might be **deployers.** Deployers are, for instance, vendors of AI systems or the senior management of an organization that decides whether to use a system in the organization.

Individual standards: Their individual standards for trustworthiness will be tied to their business goal. For example, a system would be considered trustworthy if it promised to increase profits and employee satisfaction, or if it promised to reduce risks and costs.

Secondary cues: Cues provided by the deployers might include a speech by the CEO supporting the use of a system or a company-wide advertising campaign for a new system. Another cue may be the reluctance of a senior manager to implement a system (Rousseau et al., 1998).

Although there may be more stakeholders involved in the trustworthiness propagation process (e.g., various certification institutions or deployers), we end with the **end user**. End users can be employees, casual home users, or researchers. End users play a special role in the trustworthiness propagation process, because they (1) repeatedly assess a system's AT by interacting with the system, and (2) have to decide whether a system is trustworthy in general, but also whether they want to rely on a single output (i.e. end users have to decide whether a single system recommendation belongs to the 5 % incorrect cases reported in the documentation of the AI system; Cancro et al., 2022). Presumably, continuous daily interactions allow end users to quickly adjust their weighting and utilization of cues. Users may get better (e.g., due to more knowledge and experience) or worse (e.g., due to habituation and less awareness) at detecting and utilizing cues.

Individual standards: The goals and domain knowledge of end users as well as their technical expertise and knowledge of AI systems can be heterogeneous. End user's individual standards may often be implicit, but may become more explicit and specific as they interact with the systems.

Secondary cues: End users may provide secondary cues to other end users. For instance, imagine that team members are using an AI system and seem happy with it. This can be used by other end users to assess system trustworthiness. End users may also be the ones who test a particular AI system and then try to convince their employers to deploy it in the organization. These secondary cues in the form of testimonials can be shared through various communication channels (e.g., through customer reviews in app stores, in interpersonal communication with colleagues, on social media).

5. Discussion

In seminal models of trust in automation (i.e. Lee & See, 2004) and organizational trust (i.e. Mayer et al., 1995) there formerly existed a gap because these models only included the PT of the trustor without describing how trustors arrive at their PT given the trustees' AT. In this paper, we introduce the TrAM that complements existing trust models (e.g., Lee & See, 2004; Mayer et al., 1995) by adding the trustee (and its characteristics), specifying the concepts relevant to the trustworthiness assessment (system characteristics, individual standards, AT, and PT), describing their relations, and emphasizing the role of PT in the overall trust development process. Bringing more attention to the trustworthiness assessment as a part of the trust development process may help to guide future research in understanding the transition from PT to trust and trusting behavior. Fig. 6 displays how the trust development process depicted in Fig. 1 is complemented by the TrAM.

5.1. Conceptual implications: how the TrAM relates to the trust development process

What we outline in this section is not intended to be a complete theory of the relations between PT, trust, trust propensity, and trusting

behavior. Instead, we hope that this section will stimulate discussion about the relations and distinctions between these concepts.

The propositions of the TrAM have implications regarding the specifications of trust and trusting behavior in the trust development process. Starting with the outcome of what we suggested in the micro-level of the TrAM, the PT consists of three parts: (1) the perceived presence of trustworthiness, (2) the perceived absence of trustworthiness, and (3) the perceived uncertainty regarding the trustee (see Fig. 7). The trustor's perceptions concerning the presence or absence of trustworthiness can be accurate to some degree, as indicated by the green (accurate) and red (inaccurate) areas on the outside of the circle. The proportions of accurate and inaccurate assessments reflect how accurately a trustor's PT reflects the system's AT. The perceived uncertainty reflects the remaining attributed uncertainty with respect to the trustee. Specifically, we assume that prior to all assessments and for a person with no information at all about a system, the circle is completely filled with uncertainty (blue empty boxes in the requirement list metaphor). When a trustor assesses the trustworthiness of a system, uncertainty reduces and is in parts replaced by (1) the area indicating the presence of trustworthiness (light grey; checked boxes), which starts to fill the circle clockwise at 12 o'clock, followed by (2) the area indicating the absence of trustworthiness (dark grey; crossed out boxes).

Questions following from our propositions are related to how the trust development process continues after the trustworthiness assessment process and how the other main concepts in the trust development process (i.e. trust, trust propensity, perceived stakes, and trusting behavior) relate to the outcomes of the trustworthiness assessment process. This question is also reflected in a recent review that stated that "There is a persistent gap in terms of understanding how a machine's trustworthiness maps to the human–trust variable."(Gebru et al., 2022, p. 959). The TrAM goes half the way by explicating the transition from a system's characteristics to a trustor's PT. But how does PT translate into trust? Mayer et al. (1995) are not explicit about this transition. In their model it seems that trust follows from a combination of perceived trustworthiness and the trustor's propensity to trust. Building on the specifications in the TrAM, we now derive an initial idea on the transition from PT to trust.

Mayer et al.'s (1995) conceptualization of trust as a "willingness to be vulnerable" is reflected in the propositions of the TrAM in at least three ways: First, the trustor cannot perfectly assess the trustee's AT and so they will form their PT without full or accurate knowledge of the trustee's AT. In this sense, the trustor will be vulnerable to a potential mismatch between PT and AT, and vulnerable with respect to characteristics of the system that they have not assessed or about which they are uncertain. Second, although the trustor may perceive a certain degree of presence of trustworthiness in the trustee, they may also perceive a certain absence of trustworthiness. To accept that the trustee is imperfect means being willing to accept vulnerability. Third, even if the trustor was somehow able to perfectly match their PT to the trustee's AT, they would remain vulnerable because the trustee could still disappoint the trustor, for example, due to an unforeseeable event. In this regard, trust can be said to reflect the willingness to accept a potential mismatch between PT and AT and a remaining uncertainty regarding the trustee, a potential known absence of trustworthiness, and the inherent unforseeability of certain events.

Building on these conceptual insights regarding trust and building on the outcome of the trustworthiness assessment as outlined in Fig. 7 (i.e. the trustor's PT), Fig. 8 aims to make the transition from PT to trust graspable.

Trust comes into play depicted by the orange dotted line that represents the trustor's individual threshold for their willingness to become vulnerable (i.e. to trust). This trust threshold indicates the minimum that the area representing the perceived presence of trustworthiness must reach for trust to evolve. If it is reached, a trustor is willing to accept uncertainty (about a trustee's trustworthiness) and the perceived (partial) absence of a trustee's trustworthiness. That means if the area of the circle that is filled with perceptions indicating the presence of trustworthiness passes the trust threshold, the trustor is willing to make themselves vulnerable – *the trustor is in a position to trust.* This also means that the trustor considers trusting behavior toward the trustee (i.e. to actually make themselves vulnerable to the trustee, Fig. 8A, B & D). If perceptions indicating the presence of a system's trustworthiness do not fill enough space to reach the trust threshold, the trustor will be unwilling to be vulnerable to the trustee, and thus will not trust (Fig. 8C). The more the area indicating the presence of trustworthiness surpasses the trust threshold the higher trust might be. This allows trust to be considered as a matter of degree (Loi et al., 2023), i.e. you can trust a trustee to a lower or to a higher degree.

The areas of the circle can be used to describe the relation between PT and trust. If the areas indicating the presence or absence of trustworthiness become larger, uncertainty about the trustee decreases. If the area indicating the presence of trustworthiness becomes larger, it

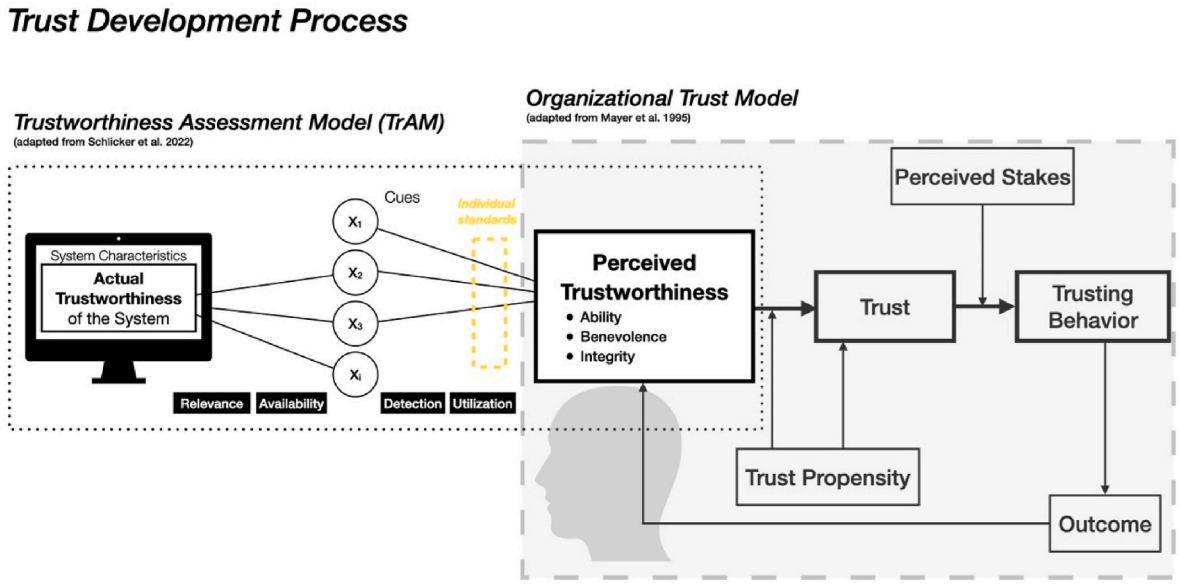

Fig. 6. Overview of the Trust Development Process including the propositions of the TrAM. The left side of the figure shows the Trustworthiness Assessment Model (TrAM) as proposed in this paper. The right side shows an adapted version of the Organizational Trust Model (Mayer et al., 1995).

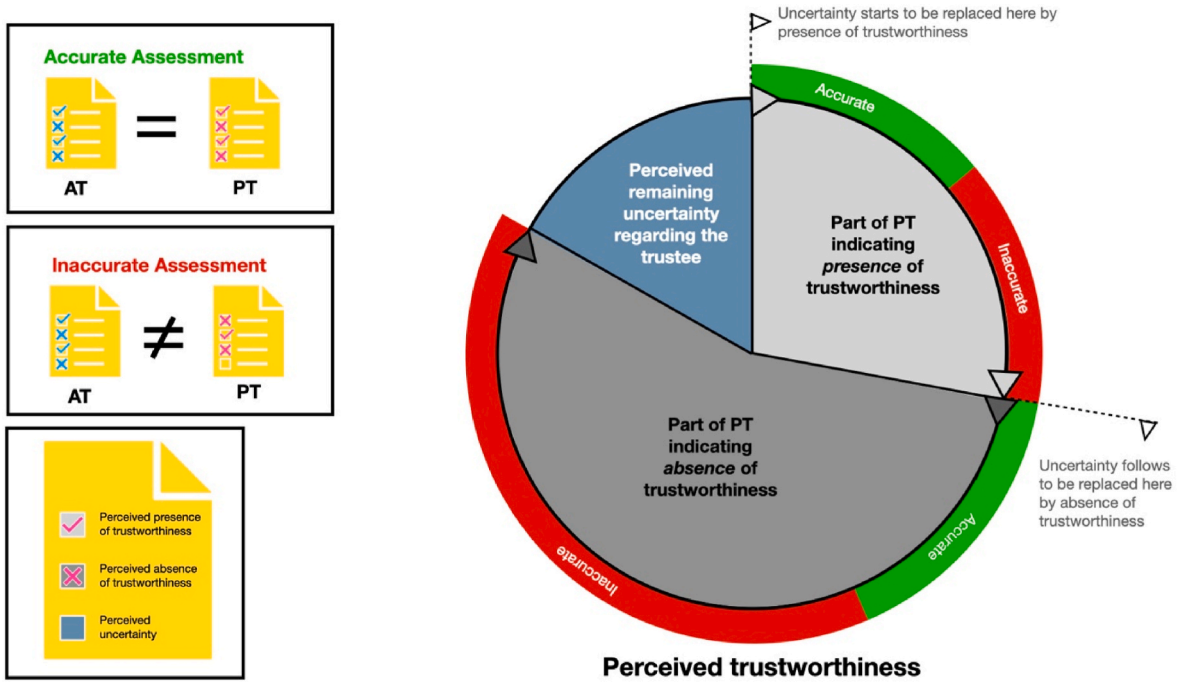

Fig. 7. The circle reflects the trustor's perceived trustworthiness (PT) of a trustee that (in parallel to our requirement list metaphor) consists of the perceived presence of trustworthiness (light grey; checked boxes), perceived absence of trustworthiness (dark grey; crossed out boxes), and the remaining uncertainty regarding system trustworthiness (blue; empty boxes). Before the assessment starts, the circle is filled completely with uncertainty. In this model the uncertainty starts to be replaced clockwise (starting at 12 o'clock) by the part of the PT that indicates the presence of trustworthiness, and is followed by the part of PT that indicates the absence of trustworthiness. The assessments of perceived presence and absence of trustworthiness can be accurate to a certain degree, indicated by the green (accurate) and red (inaccurate) areas at the outside of the circle. The degree of accuracy is the result of the comparison between a trustor's PT and a system's actual trustworthiness (AT). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

becomes more likely that the trustor is willing to accept the remaining uncertainty and any perceived absence of trustworthiness. If the uncertainty and perceived absence of trustworthiness about the trustee increase, it becomes less likely that the trustor will be willing to make themselves vulnerable to the trustee. At this point, we would like to emphasize that the perceived presence and absence of trustworthiness are not just simple additions of cues, but rather a function of cues. The details of this function (e.g., which cues are more important than others, how cues are combined) are ultimately empirical questions, but we suggest that the use of cues is a nonlinear, complex matter, where single cues and the order of those cues could strongly shift the proportions of perceived presence versus absence of trustworthiness, while the uncertainty regarding the system may remain similar.

Trust propensity comes into play when we want to respond to the question: How trustworthy is trustworthy enough to trust? As outlined in section 2 on trust propensity, some individuals are more likely to trust than others (Mayer et al., 1995). In the parlance of the TrAM, this means that some individuals need less evidence indicating the presence of trustworthiness than others. As such, the trustor's propensity to trust can affect the position of the trust threshold in Fig. 8. A high propensity to trust (i.e. a higher propensity to be willing to be vulnerable) moves the trust threshold counterclockwise (see Fig. 8C & D). In other words, the trustor will be more likely to trust, even with a comparatively larger amount of uncertainty about the trustee's trustworthiness and also with a comparatively larger amount of perceived absence of the trustee's trustworthiness.

Finally, trusting behavior comes into play when a trustor actually considers to make themselves vulnerable in a particular situation. At this point, the perceived stakes of the situation (i.e. the risks and benefits) will influence the decision of whether people will actually show trusting behavior (Mayer et al., 1995). The threshold for trustors to actually engage in trusting behavior is indicated by the turquoise dashed line in Fig. 8. A trustor shows trusting behavior, if the trusting behavior threshold is surpassed by the part of PT that indicates the presence of trustworthiness. The position of the trusting behavior threshold is influenced by the perceived stakes. If the likelihood or strength of expected negative outcomes increases, this is represented by moving the trusting behavior threshold clockwise making it less likely that the trustor will engage in trusting behavior; if the likelihood or strength of expected positive outcomes increases, this moves the dashed line counterclockwise making it more likely that the trustor will engage in trusting behavior. In short, if perceptions indicating the presence of trustworthiness reach the trust threshold, a trustor considers to become vulnerable, but if risks are high, they are unlikely to actually engage in trusting behavior.

To conclude, we hope that our description of the relation of the TrAM with the trust development process helps to enhance conceptual clarity in trust research by distinguishing the single components of the trust development process (AT, PT, trust, trust propensity, and trusting behavior). This conceptual clarity might also help to advance the measurement of the single components. Our propositions imply that a measurement of PT should be distinct from a measurement of trust. Trust researchers might be encouraged to revisit existing measurements and examine what exactly those measure. For example, the widely used "trust" in automation scale by Jian et al. (2000) includes items that measure a) trustworthiness (e.g., by using items such as: "The system has integrity" or "I am suspicious of the system's intent, action, or outputs"), b) trust (e.g.: "I can trust the system"), and c) familiarity with a system ("I am familiar with the system") which seems to neither reflect trustworthiness nor trust because one could trust a completely unfamiliar system (which will probably not be associated with an accurate trustworthiness assessment). Potentially, this might have contributed to the finding that the Jian et al. scale can produce biased results (Gutzwiller et al., 2019).

Refining existing scales and exploring new trust measurement approaches will improve the precision of trust research (Kohn et al., 2021).

Fig. 8. The figures depict the result of the trustworthiness assessment process (i.e. the trustor's perceived trustworthiness (PT)) and its relation to trust, trust propensity, and trusting behavior. The orange dotted line denotes the trustor's individual threshold for their willingness to become vulnerable (i.e. trust threshold). Trust is present when the part of PT indicating the presence of trustworthiness surpasses the trust threshold. Trusting behavior depends on the position of the turquoise dashed line (i.e. trusting behavior threshold). The position of this line depends on the perceived risks (which move the dashed line clockwise) and benefits (which move the dashed line counterclockwise) (Fig. 8A & B). The position of the trust threshold might be affected by, for instance, the trustor's propensity to trust (Fig. 8C & D). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Inspiration for the distinct measurement of trustworthiness and trust might be retrieved from organizational trust research. For example, Mayer and Davis (1998) provided a questionnaire that distinguishes between facets of trustworthiness on the one hand (e.g., items such as "Top management is very capable of performing its job." capture ability; items like "Top management is very concerned about my welfare." capture benevolence; and items like "I never have to wonder whether top management will stick to its word." capture integrity), and trust on the other hand (e.g., items such as "I would be willing to let top management have complete control over my future in this company."; "I really wish I had a good way to keep an eye on top management."). Furthermore, it could be considered to measure trust as planned anti-monitoring as suggested by Loi et al. (2023) analyzing how much vulnerability people have accepted toward a system.

5.2. How the TrAM and its propositions relate to calibrated trust

By specifying the trustworthiness assessment process, and by

specifying the proposed relations between AT, PT, trust, trust propensity, and trusting behavior (see Fig. 6), this paper can also help to better understand the concept of "calibrated trust". Calibrated trust has received much attention, but also increasing criticism, (see Wischnewski et al., 2023, for a review). For example, while research on calibrated trust shares the goal of optimizing the joint performance of humans and systems, there is great heterogeneity in defining and measuring calibrated trust (Carter et al., 2023; Merritt, Lee, et al., 2015; Wischnewski et al., 2023).

We do not attempt to redefine calibrated trust, but we can draw conclusions from the TrAM about what calibrated trust may *not* be. Calibrated trust has been defined as (1) a match between AT and PT (de Visser et al., 2020), (2) a match of a system characteristics (often system accuracy) and a trustor's trust in the system (e.g., Laux et al., 2023; Lee & See, 2004), and (3) as correct trusting behavior (e.g., rejecting incorrect and accepting correct system outputs; Okamura & Yamada, 2020; Zhang et al., 2020). Regarding (1), we propose that a strong match between AT and PT can more simply be described as a high accuracy of the trustworthiness assessment, which then supports well-informed and appropriate decisions in further stages of the trust development process. Regarding (2), we argue that this definition of calibrated trust neglects the trustworthiness assessment process – a process that this paper introduces explicitly but that prior research at least implicitly assumed to be a basis of trust development processes (e.g., Lee & See, 2004; Mayer et al., 1995). Additionally, it is unclear how a certain degree of trust could align with specific system characteristics. We think it is easier to follow research on how people assess the characteristics of other entities (Funder, 1995; Hammond & Stewart, 2001) and to propose that trustors try to assess the AT of a system, this leads to their PT (which is accurate to a certain degree), and then PT affects whether people will be in a position to trust and to engage in trusting behavior. Regarding (3) we argue that this definition of calibrated trust may rather reflect calibrated trusting behavior (e.g., appropriate reliance; Schoeffer et al., 2025; Vereschak et al., 2021). Arguably, calibrated trusting behavior is something that we hope for when we try to optimize the accuracy with which people assess the trustworthiness of a system – but following and rejecting systems in the right situation seems to be conceptually downstream of correctly assessing the trustworthiness of a system.

Moreover, the term calibrated trust runs the risk of confusing an accurate trustworthiness assessment with a justified trustworthiness assessment. Whereas an accurate trustworthiness assessment reflects a match between the system's AT and the trustor's PT, a justified trustworthiness assessment results from an appropriate utilization of available and detected cues. In other words, a trustor's PT may be *justified* given the available cues, even if it does not *accurately* reflect the AT. Our model allows for a case where all the cues available to the trustor about the system have been adequately detected and utilized, and yet this perfectly justified PT may not accurately reflect the AT of the system. For example, an otherwise reliable and capable system may perform poorly in a given situation, but cues indicating this absence of trustworthiness may not have been available to the trustor. So, we may need to be careful not to mistakenly accuse the trustor of lacking "calibrated trust" and instead think of ways how to enhance the availability of relevant cues.

5.3. Implications at the micro level

5.3.1. Implications of the relevance and availability of cues on the side of the system

The micro-level trustworthiness assessment process implies that, in order to achieve an accurate assessment of a system's AT, relevant cues must be made available on the side of the system. Knowing which cues are relevant would enable to design systems that are tailored to the trustor's individual standards for trustworthiness. If no cues are available for a trustor to evaluate a particular facet of AT (e.g., fairness) in light of their individual standards, research must investigate what cues might be appropriate.

Regarding cue relevance, it may also be necessary to consider the possible changing relevance of cues. For instance, environmental changes may affect the AT of a system which may also affect the relevance of a cue (Jalaian et al., 2019). Therefore, it may be necessary to consider ways to make trustors aware of possible problems with the declining relevance of certain cues. For example, it may be helpful to provide an "expiration" date for certain cues (e.g., for performance metrics or the representativeness of sample data).

By examining the availability of cues, we may find that some cues to assess the system's AT are more readily available than others, or that some are only available when trustors interact with systems over a long period of time. This type of cue availability analysis has implications for system design (e.g., how to make relevant cues available earlier, how to make those cues more accessible) and for human-system interaction (e.g., extending training time; explicitly producing certain cues, such as erroneous output, during user training; de Visser et al., 2014; de Visser et al., 2020; Hoff & Bashir, 2015).

The relevance and availability of cues for an accurate trustworthiness assessment are also central in the area of explainable AI (XAI). In the terminology of our model, XAI research aims to make relevant cues available (Langer & Landers, 2021). For example, XAI research develops transparent models, e.g., models based on simple rules, where these rules provide relevant cues, or post-hoc explainability approaches that try to generate relevant cues by analyzing the decision processes of a black-box model (Baum et al., 2022, Baum et al., 2023; Biewer et al., 2024; Kenny et al., 2021; Miller, 2019; Wysocki et al., 2023; Lipton, 2018). However, for such black-box models (e.g., using deep neural networks), it may be especially challenging to achieve accurate trustworthiness assessments because these systems may make it more difficult to provide relevant cues. This has also implications for discussions evolving around "effective human oversight", as demanded in the AI Act (European Parliament, 2024) for high-risk application contexts (Baum et al., 2025; Biewer et al., 2024; Green, 2022; Jobin et al., 2019; Sterz et al., 2024). *Effective* human oversight requires that humans can sufficiently reliable assess when to rely on system output and when to reject it. Thus, effective human oversight requires an accurate trustworthiness assessment. In line with other research (Langer et al., 2024; Green, 2022), our model suggests that if we want effective human oversight, we need to focus on making relevant cues available.

5.3.2. Implications of detection and utilization of cues on the side of the trustor

On the side of the trustor, factors that influence the accuracy of trustworthiness assessments include the trustors' individual standards and the detection and utilization of cues. The degree of detail with which trustors can make their individual standards explicit may determine the accuracy of their trustworthiness assessment. Trustors who have clear and explicit individual standards, may be better able to compare different systems, as well as compare their own standards to system's AT. Thus, it may be valuable to examine ways to make trustors more aware of their individual standards. For example, individual standards can be made more accessible by encouraging different stakeholders to express their individual standards. This may help to identify similarities, differences, possible granularities of individual standards, and to explicate these standards.

Cue detection could be influenced by trustors' individual standards, and factors such as attentional guidance and information processing. Experiments (e.g., using eye tracking) or usability testing can provide information about cue detection. Especially, when multiple cues are available (e.g., primary and secondary; conflicting and highly correlated cues), this may provide fruitful insights into cue selection and weighting. Knowledge gained from such studies could be used to make irrelevant cues less salient and relevant cues more salient.

Examining cue utilization provides information about trustor's individual standards, expectations of the system, and trustor's knowledge and weighting of cue relevance. Knowing which cues are frequently

utilized and heavily weighted by users to assess system trustworthiness can further influence system design by adapting cues to make them more adequately utilized (e.g., decreasing saliency of cues that are frequently utilized incorrectly) (Durant et al., 2022). Additionally, users could be trained to correctly utilize detected cues in order to improve the accuracy of their assessment. This training could include, for example, correctly interpreting and understanding accuracy-related information (e.g., precision, recall) of a classifier (Juba & Le, 2019), as well as error training, in which users learn, which cues are present when the system fails and which cues might be relevant to check whether the system is working properly (Bahner et al., 2008).

5.4. Implications at the macro level

At the macro level, the trustworthiness propagation process proposes that the assessment of a system's AT is a net of micro-level assessments. This also implies that trust in a system is often the consequence of trust in other stakeholders and their trustworthiness assessments (see also Ehsan et al., 2021; Kelton et al., 2008; Knowles & Richards, 2021). Thus, the assessment of system trustworthiness involves the combination of the assessment of trust in other people, trust in institutions, and trust in systems (Laux et al., 2023). As a consequence, for an accurate assessment, downstream trustors need to assess not only the AT of a system, but also the AT of providers of secondary cues. Secondary cues result from another stakeholder's trustworthiness assessment and are therefore colored by their individual standards. They may be designed with a specific goal in mind (Ehsan et al., 2021; Knowles & Richards, 2021), such as convincing others of a system's trustworthiness or of embedded moral values (de Visser et al., 2014; Cancro et al., 2022). Individual standards for trustworthiness may differ between the stakeholders (see also Knowles & Richards, 2021), and if these differences are not explicit, incorrect assumptions about cues may lead to an inadequate utilization and thus low accuracy in the assessment of AT. We may even hypothesize that the accuracy of the trustworthiness assessment increases when different trustors providing secondary cues make their individual standards explicit and available for scrutiny by others.

Policy makers could anticipate potential trustworthiness propagation nets in the design and implementation of AI systems. This means identifying the trustors involved in the net, their respective individual standards, and the cues they might use and provide. Policy makers, standardization committees, and certification institutions may then need to ensure that different trustors can determine whether their individual standards are reflected in the system. This is in line with Knowles & Richards (2021) who argue that it should be communicated "how the system ensures that AI is trustworthy *in a way that meaningfully relates to how the public would define trustworthy AI*" (p. 268, emphasis in original).

This also applies to the communication about "trustworthy systems", which needs to emphasize the subjective nature of trustworthiness. For example, imagine that a certification institution develops a "trustworthy AI" label. Given the subjective nature of trustworthiness, trustors who see this label may mistakenly assume that their individual standards are met, and may not examine the system in detail (Ross et al., 1977), which could lead to an inaccurate trustworthiness assessment. Such possible misinterpretations of cues intended to communicate specific characteristics of different products were also shown for "ecolabels", "fair trade labels" (Czarnezki et al., 2014), and "trusted shops" labels (Rüdiger et al., 2022). We are not saying that such labels are always misleading – they can help people to quickly assess whether a product meets their individual standards, a process that would otherwise require a resource-intensive (trustworthiness) assessment process. However, we do say that calling a label "trustworthy AI" may lead to different expectations among different trustors and should therefore be accompanied by a summary of the individual standards that were relevant to a particular institution that assessed the trustworthiness of a system.

5.5. Limitations

There are two limitations of the proposed model that we would like to emphasize. First, the model makes propositions about theoretical processes and concepts that are currently difficult to capture and that may not be readily accessible to the stakeholders involved. For instance, trustors may initially not be able to communicate their individual standards, or may not be aware of the cues they have used in assessing a system's trustworthiness. We do not expect trustors to always explicitly engage in cue detection and utilization, nor may they be motivated to engage in such a process. Despite the challenge of making the concepts of our model measurable, and despite describing processes that may outline a modeled reality, we believe that the concepts we describe are important for better understanding the process by which people arrive at their PT of a system and crucial for better conceptualizing the entire trust development process. Our model aims to inspire research (1) on the side of the system to make relevant cues available, (2) on the side of the trustor to understand what influences the detection and utilization of cues, (3) on the trustworthiness propagation process, e.g., what kind of secondary cues are produced along the trustworthiness propagation process and how do they influence trustworthiness assessments, and (4) that makes the proposed concepts measurable.

Second, we claim that individual standards of system trustworthiness differ across trustors and are task- and context-specific. Thus, at first glance, individual standards may appear to be a concept with limited value for research and practice due to its variability. It is important to emphasize that we do not expect these standards to be arbitrary, but rather expect them to relate to a common understanding of trustworthiness (Jobin et al., 2019; Lee & See, 2004; Mayer et al., 1995). Accordingly, we expect overlap in the individual standards across trustors, but also that the exact combination of features, as well as their weight in forming PT, may be idiosyncratic.

6. Conclusion

The safe and effective use of AI systems requires an accurate assessment of the system, its strengths and weaknesses against a trustor's goals and standards. An accurate assessment of system trustworthiness is therefore a prerequisite for well-grounded trust, trusting behavior and also for what has been termed "alignment" between humans and AI-based systems (Gabriel, 2020). In this paper, we introduced the Trustworthiness Assessment Model, described its micro-level trustworthiness assessment process, which explains how trustors assess a system's AT to arrive at their PT of the system, and we described the model's macro-level trustworthiness propagation process which emphasizes that different stakeholders provide cues that other stakeholders can use to assess system trustworthiness. In doing so, we advance trust theory by explicating a process that focuses on trustworthiness as the basis for trust, and by providing a model that adds the trustee and its characteristics to existing trust models. Additionally, we clarify the concepts of the trust development process as well as their relations. To specify the relations, we provide visualizations of the trust development process. Our practical implications include suggestions on (1) what system designers can do to better enable trustors to accurately assess system trustworthiness, (2) what trustors need to be aware of when assessing system trustworthiness, (3) the role of trustworthiness propagation. Each concept in our model requires empirical evaluation and interdisciplinary theoretical refinement – something we hope will happen in response to the ideas we present in this paper.

Statement on the use of AI tools

During the preparation of this work the authors used DeepL and DeepL Write BETA in order to increase language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Nadine Schlicker: Conceptualization, Visualization, Writing – original draft, Writing – review & editing. **Kevin Baum:** Writing – review & editing, Conceptualization, Funding acquisition. **Alarith Uhde:** Writing – review & editing. **Sarah Sterz:** Writing – review & editing. **Martin C. Hirsch:** Supervision, Writing – review & editing. **Markus Langer:** Writing – review & editing, Conceptualization, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is partially funded by the German Research Foundation (DFG) in the project 389792660 as part of the Transregional Collaborative Research Center TRR 248 – Foundations of Perspicuous Software Systems (CPEC, Project 389792660), and by the German Federal Ministry of Education and Research (BMBF) as part of the project "MACMERLin" (Grant Agreement No. 01IW24007), and the project "Ophthalmo-AI" (Grant Agreement No. 16SV8640). Additionally, this work received support from the European Regional Development Fund (ERDF) and the Saarland within the scope of the (To)CERTAIN project as part of the Center for European Research in Trusted Artificial Intelligence (CERTAIN) as well as from the Volkswagen Stiftung in the project "Explainable Intelligent System" (AZ 98512, 98513, and 98514), and from the Daimler and Benz Stiftung in the project "TITAN" (Grant No. 45-06/24). We want to thank Kilian Krug for his support in the design of Figure 5.

Data availability

No data was used for the research described in the article.

References

Alam, L., & Mueller, S. (2021). Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *BMC Medical Informatics and Decision Making, 21*(1), 178. https://doi.org/10.1186/s12911-021-01542-6

Arnold, M., Bellamy, R.K.E., Hind, M., Houde, S., Mehta, S., & Mojsilovic, A., et al. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, *63*(4/5), 6:1-6:13. https://doi.org/10.1147/JRD.2019.2942288.

Asch, S. E. (1955). Opinions and social pressure. *Scientific American, 193*(5), 31–35.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., et al. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Baer, M. D., & Colquitt, J. A. (2018). Why do people trust?: Moving toward a more comprehensive consideration of the antecedents of trust. In R. H. Searle, I. Ann-Marie, Nienaber, & S. B. Sitkin (Eds.), *The Routledge companion to trust* (1st ed., pp. 163–182). Routledge. https://doi.org/10.4324/9781315745572-12.

Bahner, J. E., Elepfandt, M. F., & Manzey, D. (2008). Misuse of diagnostic aids in process control: The effects of automation misses on complacency and automation bias. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting, 52*(19), 1330–1334. https://doi.org/10.1177/154193120805201906

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. *Proceedings of the AAAI conference on human computation and crowdsourcing, 7*, 2–11). https://doi.org/10.1609/hcomp.v7i1.5285

Baracaldo, N., Ali, A., Purcell, M., Rawat, A., Sinn, M., Altakrouri, B., et al. (2022). Towards an accountable and reproducible federated learning: A FactSheets approach. http://arxiv.org/abs/2202.12443.

Bærøe, K., Miyata-Sturm, A., & Henden, E. (2020). How to achieve trustworthy artificial intelligence for health. *Bulletin of the World Health Organization, 98*(4), 257–262. https://doi.org/10.2471/BLT.19.237289

Baum, K. (2016). In T. Margaria, & B. Steffen (Eds.), *Leveraging Applications of Formal Methods, Verification and Validation: Discussion, Dissemination* (pp. 633–647). Springer International Publishing. https://doi.org/10.1007/978-3-319-47169-3_49.

Baum, D., Baum, K., Gros, T.P., & Wolf, V. (2023). XAI requirements in smart production processes: A case study. In L. Longo (Ed.), *Explainable artificial intelligence*, 2023 (pp. 3–24). Springer Nature Switzerland. doi:10.1007/978-3-031-44064-9_1.

Baum, K., Biewer, S., Hermanns, H., Hetmank, S., Langer, M., Lauber-Rönsberg, A., & Sterz, S. (2025). In T. Neele, & A. Wijs (Eds.), *Model Checking Software* (Vol. 14624, pp. 3–25). Springer Nature Switzerland.

Baum, K., Mantel, S., Schmidt, E., & Speith, T. (2022). From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology, 35*(1), 12. https://doi.org/10.1007/s13347-022-00510-w

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275*(5304), 1293–1295. https://doi.org/10.1126/science.275.5304.1293

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., et al. (2019). AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *IBM Journal of Research and Development, 63*(4/5), 4:1–4:15. https://doi.org/10.1147/JRD.2019.2942287

Benčević, M., Habijan, M., Galić, I., Babin, D., & Pižurica, A. (2024). Understanding skin color bias in deep learning-based skin lesion segmentation. *Computer Methods and Programs in Biomedicine, 245*, Article 108044. https://doi.org/10.1016/j.cmpb.2024.108044

Biewer, S., Baum, K., Sterz, S., Hermanns, H., Hetmank, S., Langer, M., et al. (2024). Software doping analysis for human oversight. *Formal Methods in System Design*. https://doi.org/10.1007/s10703-024-00445-2

Birhane, A. (2022). The unseen Black faces of AI algorithms. *Nature, 610*(7932), 451–452. https://doi.org/10.1038/d41586-022-03050-7

Bolton, M. L. (2022). Trust is not a virtue: Why we should not trust trust. *Ergonomics in Design.* , Article 10648046221130171. https://doi.org/10.1177/10648046221130171

Borsboom, D., Mellenbergh, G. J., & Heerden, J.van (2003). The theoretical status of latent variables. *Psychological Review, 110*, 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Bower, G. H. (1981). Mood and memory. *American Psychologist, 36*, 129–148. https://doi.org/10.1037/0003-066X.36.2.129

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. University of California Press. https://doi.org/10.1525/9780520350519

Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces,* (pp. 454–464). https://doi.org/10.1145/3377325.3377498

Cancro, G.J., Pan, S., & Foulds, J. (2022). Tell me something that will help me trust you: A survey of trust calibration in human-agent interaction (arXiv:2205.02987). arXiv. http://arxiv.org/abs/2205.02987.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Carter, O. B. J., Loft, S., & Visser, T. A. W. (2023). Meaningful communication but not superficial anthropomorphism facilitates human-automation trust calibration: The Human-Automation Trust Expectation Model (HATEM). *Human Factors, 66*(11), 2485–2502. https://doi.org/10.1177/00187208231218156

Chiou, E. K., & Lee, J. D. (2021). Trusting automation: Designing for responsivity and resilience. *Human Factors, 65*, 137–165. https://doi.org/10.1177/00187208211009995

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., & Usunier, N. (2017). Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning , 70,* (pp. 854–863). https://proceedings.mlr.press/v70/cisse17a.html.

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology, 92*(4), 909–927. https://doi.org/10.1037/0021-9010.92.4.909

Conway, D., Chen, F., Yu, K., Zhou, J., & Morris, R. (2016). Misplaced trust: A bias in human-machine trust attribution - In contradiction to learning theory. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems,* (pp. 3035–3041). https://doi.org/10.1145/2851581.2892433

Czarnezki, J., Homan, A., & Jeans, M. (2014). Creating order amidst food eco-label chaos. *Duke Environmental Law and Policy Forum, 25*(2014), 281. https://scholarship.law.duke.edu/delpf/vol25/iss2/2.

de Visser, E. J., Cohen, M., Freedy, A., & Parasuraman, R. (2014). In R. Shumaker, & S. Lackey (Eds.), *Virtual, Augmented and mixed reality. Designing and developing virtual and augmented environments, (Vol. 8525* pp. 251–262). Springer International Publishing. https://doi.org/10.1007/978-3-319-07458-0_24.

de Visser, E. J., Peeters, M. M. M., Jung, M., Kohn, S., Tyler, S., Pak, R., et al. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics, 12*. https://doi.org/10.1007/s12369-019-00596-x

Deutsch, M. (1975). Equity, equality, and need: What determines which value will be used as the basis of distributive justice? *Journal of Social Issues, 31*(3), 137–149. https://doi.org/10.1111/j.1540-4560.1975.tb01000.x

Dietvorst, B. J., & Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science, 31*(10), 1302–1314. https://doi.org/10.1177/0956797620948841

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietz, G., & Den, H. D. N. (2006). Measuring trust inside organisations. *Personnel Review, 35*(5), 557–588. https://doi.org/10.1108/00483480610682299

DIN, DKE. (2020). *German standardization roadmap on artificial intelligence*. https://www.din.de/resource/blob/772610/8bfea3055c03aa1e2563afc16001b06f/normungsroadmap-en-data.pdf.

Drobotowicz, K., Kauppinen, M., & Kujala, S. (2021). In *Trustworthy AI services in the public sector: What are citizens saying about It?, (Vol. 12685* pp. 99–115). Engineering: Foundation for Software Quality. https://doi.org/10.1007/978-3-030-73128-1_7.

Durant, S., Wilkins, B., Woods, C., Uliana, E., & Stathis, K. (2022). Attention guidance agents with eye-tracking. In N. Alechina, M. Baldoni, & B. Logan (Eds.), *Engineering multi-agent systems* (pp. 92–113). Springer International Publishing. https://doi.org/10.1007/978-3-030-97457-2_6.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors, 44*(1), 79–94. https://doi.org/10.1518/0018720024494856

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems,* (pp. 1–19). https://doi.org/10.1145/3411764.3445188

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. *Human Factors, 59*(1), 5–27. https://doi.org/10.1177/0018720816681350

European Parliament. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) http://data.europa.eu/eli/reg/2024/1689/oj.

Flathmann, C., Schelble, B. G., Zhang, R., & McNeese, N. J. (2021). Modeling and guiding the creation of ethical human-AI teams. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 469–479). https://doi.org/10.1145/3461702.3462573

Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence, 1*(6), 261–262. https://doi.org/10.1038/s42256-019-0055-y

Forgas, J. P. (1995). Mood and judgment: The affect infusion model (AIM). *Psychological Bulletin, 117*, 39–66. https://doi.org/10.1037/0033-2909.117.1.39

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (Im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM, 64*(4), 136–143. https://doi.org/10.1145/3433949

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*(4), 652–670. https://doi.org/10.1037/0033-295X.102.4.652

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Gebru, B., Zeleke, L., Blankson, D., Nabil, M., Nateghi, S., Homaifar, A., et al. (2022). A review on human–machine trust evaluation: Human-centric and machine-centric perspectives. *IEEE Transactions of Human Machine Systems, 52*(5), 952–962. https://doi.org/10.1109/THMS.2022.3144956

Geyer, T., & Müller, H. J. (2009). Distinct, but top-down modulable color and positional priming mechanisms in visual pop-out search. *Psychology Research, 73*(2), 167–176. https://doi.org/10.1007/s00426-008-0207-x

Glikson, E., & Woolley, A. (2020). Human trust in artificial intelligence: Review of empirical research. *The Academy of Management Annals*, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057.

Goffman, E. (2006). The presentation of self. In *Life as theater: A dramaturgical sourcebook* (pp. 129–139). New Brunswick, NJ: Aldine Transaction. https://bpb-us-e2.wpmucdn.com/sites.middlebury.edu/dist/c/2414/files/2012/09/Goffman2.pdf.

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Report, 45*, Article 105681. https://doi.org/10.1016/j.clsr.2022.105681

Guffey, M. E., & Loewy, D. (2012). *Essentials of business communication.* Cengage Learning.

Gutzwiller, R. S., Chiou, E. K., Craig, S. D., Lewis, C. M., Lematta, G. J., & Hsiung, C.-P. (2019). Positive bias in the 'Trust in Automated Systems Survey'? An examination of the Jian et al. (2000) scale. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting, 63*(1), 217–221. https://doi.org/10.1177/1071181319631201

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice.* Oxford University Press.

Hammond, K. R., & Stewart, T. R. (2001). *The essential Brunswik: Beginnings, explications, applications.* Oxford University Press.

Hardré, P. L. (2016). When, how, and why do we trust technology too much? In S. Y. Tettegah, & D. L. Espelage (Eds.), *Emotions, Technology, and Behaviors* (pp. 85–106). Academic Press. https://doi.org/10.1016/B978-0-12-801873-6.00005-4.

Hauer, M. P., Adler, R., & Zweig, K. (2021). Assuring fairness of algorithmic decision making. In *2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)* (pp. 110–113). https://doi.org/10.1109/ICSTW52544.2021.00029

Hauschke, A., Puntschuh, M., Hallensleben, S., & Loh, W.. VDE SPEC 90012 V1.0 - VCIO based description of systems for AI trustworthiness characterisation. https://www.vde.com/resource/blob/2177870/a24b13db01773747e6b7bba4ce20ea60/vde-spec-90012-v1-0–en–data.pdf.

Hawkins, H., Hillyard, S., Luck, S., Mouloua, M., Downing, C., & Woodward, D. (1990). Visual attention modulates signal detectability. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 802–811. https://doi.org/10.1037/0096-1523.16.4.802

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI.* https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

High-Level Expert Group on Artificial Intelligence. (2020). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

Hoens, T. R., Polikar, R., & Chawla, N. V. (2012). Learning from streaming data with concept drift and imbalance: An overview. *Progress in Artificial Intelligence, 1*(1), 89–101. https://doi.org/10.1007/s13748-011-0008-0

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Human, L. J., & Biesanz, J. C. (2013). Targeting the good target: An integrative review of the characteristics and consequences of being accurately perceived. *Personality and Social Psychology Review, 17*(3), 248–272. https://doi.org/10.1177/1088868313495593

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* (pp. 624–635). https://doi.org/10.1145/3442188.3445923

Jalaian, B., Lee, M., & Russell, S. (2019). Uncertain context: Uncertainty quantification in machine learning. *AI Magazine, 40*(4), 40–49. https://doi.org/10.1609/aimag.v40i4.4812

Jensen, M. S., Yao, R., Street, W. N., & Simons, D. J. (2011). Change blindness and inattentional blindness. *WIREs Cognitive Science, 2*(5), 529–546. https://doi.org/10.1002/wcs.130

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). In J. Y. C. Chen, & G. Fragomeni (Eds.), *Virtual, augmented and mixed reality. Applications and case studies* (Vol. 11575, pp. 476–489). Springer International Publishing. https://doi.org/10.1007/978-3-030-21565-1_32.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Juba, B., & Le, H. S. (2019). Precision-recall versus accuracy and the role of large data sets. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 4039–4048. https://doi.org/10.1609/aaai.v33i01.33014039

Jung, J. C., & Sharon, E. (2019). The Volkswagen emissions scandal and its aftermath. *Global Business Organizational Excellence, 38*(4), 6–15. https://doi.org/10.1002/joe.21930

Kaplan, A. D., Kessler, T. T., Christopher Brill, J., & Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human Factors, 65*(2), 337–359. https://doi.org/10.1177/00187208211013988

Kaufman, A. S., Flanagan, D. P., Alfonso, V. C., & Mascolo, J. T. (2006). Test review: Wechsler intelligence scale for children, (WISC-IV). *The Journal of Psychoeducational Assessment, 24*(3), 278–295. https://doi.org/10.1177/0734282906288389

Kay, M., Patel, S. N., & Kientz, J. A. (2015). How good is 85%? A survey tool to connect classifier evaluation to acceptability of accuracy. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems,* (pp. 347–356). https://doi.org/10.1145/2702123.2702603

Keller, M. D., Harrison-Smith, B., Patil, C., & Arefin, M. S. (2022). Skin colour affects the accuracy of medical oxygen sensors. *Nature, 610*(7932), 449–451. https://doi.org/10.1038/d41586-022-03161-1

Kelp, C., & Simion, M. (2022). What is trustworthiness? *Noûs, 57*, 667–683. https://doi.org/10.1111/nous.12448.

Kelton, K., Fleischmann, K. R., & Wallace, W. A. (2008). Trust in digital information. *Journal of the American Society for Information Science and Technology, 59*(3), 363–374. https://doi.org/10.1002/asi.20722

Kenny, E. M., Ford, C., Quinn, M., & Keane, M. T. (2021). Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence, 294*, Article 103459. https://doi.org/10.1016/j.artint.2021.103459

Kirlik, A. (2006). *Adaptive perspectives on human-technology interaction: Methods and models for cognitive engineering and human-computer interaction.* USA: Oxford University Press.

Knowles, B., & Richards, J. T. (2021). The sanction of authority: Promoting public trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* (pp. 262–271). https://doi.org/10.1145/3442188.3445890

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19),* (pp. 1–14). https://doi.org/10.1145/3290605.3300641

Koh, Y. J., & Sundar, S. S. (2010). Effects of specialization in computers, web sites, and web agents on e-commerce trust. *International Journal of Human-Computer Studies, 68*(12), 899–912. https://doi.org/10.1016/j.ijhcs.2010.08.002

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology, 12*, Article 604977. https://doi.org/10.3389/fpsyg.2021.604977

Körber, M. (2019). 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),* (pp. 13–30.). Springer International Publishing. https://doi.org/10.1007/978-3-319-96074-6_2.

Kuncel, N. R. (2018). Judgment and decision making in staffing research and practice. In *The SAGE handbook of industrial, work and organizational psychology: Personnel psychology and employee performance* (pp. 474–487). SAGE Publications Ltd.. https://doi.org/10.4135/9781473914940.n17

Langer, M., Baum, K., & Schlicker, N. (2024). Effective Human Oversight of AI-Based Systems: A Signal Detection Perspective on the Detection of Inaccurate and Unfair Outputs. *Minds and Machines, 35*. https://doi.org/10.1007/s11023-024-09701-0.

Langer, M., König, C. J., Back, C., & Hemsing, V. (2022). Trust in artificial intelligence: Comparing trust processes between human and automated trustees in light of unfair bias. *Journal of Business and Psychology*. https://doi.org/10.1007/s10869-022-09829-9

Langer, M., König, C. J., & Busch, V. (2021). Changing the means of managerial work: Effects of automated decision support systems on personnel selection tasks. *Journal of Business and Psychology, 36*(5), 751–769. https://doi.org/10.1007/s10869-020-09711-6

Langer, M., & Landers, R. N. (2021). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior, 123*, Article 106878. https://doi.org/10.1016/j.chb.2021.106878

Laurent, J., Swerdlik, M., & Ryburn, M. (1992). Review of validity research on the Stanford-Binet Intelligence Scale: Fourth edition. *Psychological Assessment, 4*(1), 102–112. https://doi.org/10.1037/1040-3590.4.1.102

Laux, J., Wachter, S., & Mittelstadt, B. (2023). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Government*. https://doi.org/10.1111/rego.12512

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243–1270. https://doi.org/10.1080/00140139208967392

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, M., Alarcon, G., & Capiola, A. (2022). I think you are trustworthy, need I say more?" The factor structure and practicalities of trustworthiness assessment. *Frontiers in Psychology, 13*, 797443. https://doi.org/10.3389/fpsyg.2022.797443

Li, X., Cui, Z., Wu, Y., Gu, L., & Harada, T. (2021). Estimating and improving fairness with adversarial learning. http://arxiv.org/abs/2103.04243.

Liao, Q. V., & Sundar, S. S. (2022). In *Designing for responsible trust in AI systems: A communication perspective. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1257–1268). https://doi.org/10.1145/3531146.3533182.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *ACM Queue, 16*(3), 31–57. https://doi.org/10.1145/3233231

Loi, M., Ferrario, A., & Viganò, E. (2023). How much do you trust me? A logico-mathematical analysis of the concept of the intensity of trust. *Synthese, 201*(6), 186. https://doi.org/10.1007/s11229-023-04169-4

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In *11th Australasian Conference on Information Systems, 53* (pp. 6–8). https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b8eda9593fbcb63b7ced1866853d9622737533a2.

Mayer, R. C., Davis, J. H., & David Schoorman, F. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734. https://doi.org/10.2307/258792

Mayer, R. C., & Davis, J. H. (1998). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology, 84*(1), 123–136. https://doi.org/10.1037/0021-9010.84.1.123

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal, 38*(1), 24–59. https://doi.org/10.2307/256727

McLeod, C. (2021). Trust. In E. N. Zalta (Ed.), *The stanford Encyclopedia of philosophy (fall 2021)*. Metaphysics Research Lab, Stanford University. Retrieved September 7, 2022 from https://plato.stanford.edu/archives/fall2021/entries/rust/.

Melms, L., Ilesan, R. R., Ulrich, K., Hildebrandt, O., Conradt, R., Eckstein, J., et al. (2023). Training one model to detect heart and lung sound events from single point auscultations. http://arxiv.org/abs/2301.06078.

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors, 53*(4), 356–370. https://doi.org/10.1177/0018720811411912

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust It, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors, 55*(3), 520–534. https://doi.org/10.1177/0018720812465081

Merritt, S. M., Lee, D., Unnerstall, J. L., & Huber, K. (2015). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors, 57*(1), 34–47. https://doi.org/10.1177/0018720814561675

Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring individual differences in the perfect automation schema. *Human Factors, 57*(5), 740–753. https://doi.org/10.1177/0018720815581247

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., et al. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19),* (pp. 220–229). https://doi.org/10.1145/3287560.3287596

Möllering, G. (2006). *Trust: Reason, routine, reflexivity*. Emerald Group Publishing.

Morik, K., Kotthaus, H., Heppe, L., Heinrich, D., Fischer, R., & Pauly, A., et al. (2021). The care label concept: A certification suite for trustworthy and resource-aware machine learning *arXiv:2106.00512 [Cs]*. http://arxiv.org/abs/2106.00512.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies, 27*(5), 527–539. https://doi.org/10.1016/S0020-7373(87)80013-5

Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. *PLoS One, 15*(2), Article e0229132. https://doi.org/10.1371/journal.pone.0229132

Oliveira, J., Renna, F., Dias Costa, P., Nogueira, M., Oliveira, C., Ferreira, C., et al. (2021). The CirCor DigiScope dataset: From murmur detection to murmur classification. *IEEE Journal of Biomedical Health Information, 2021*. https://doi.org/10.1109/JBHI.2021.3137048

Papagni, G., de Pagter, J., Zafari, S., Filzmoser, M., & Koeszegi, S. T. (2022). Artificial agents' explainability to support trust: Considerations on timing and context. *AI & Society, 38*(2), 947–960. https://doi.org/10.1007/s00146-022-01462-7

Papenmeier, A., Kern, D., Hienert, D., Kammerer, Y., & Seifert, C. (2022). How accurate does it feel? Human perception of different types of classification mistakes. In *CHI Conference on Human Factors in Computing Systems (CHI '22),* (pp. 1–13). https://doi.org/10.1145/3491102.3501915

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Parikh, R. B., Teeple, S., & Navathe, A. S. (2019). Addressing bias in artificial intelligence in health care. *JAMA, 322*(24), 2377–2378. https://doi.org/10.1001/jama.2019.18058

Prewett, P. N. (1995). A comparison of two screening tests (the Matrix Analogies test—short form and the Kaufman Brief intelligence test) with the WISC-III. *Psychological Assessment, 7*(1), 69–72. https://doi.org/10.1037/1040-3590.7.1.69

Rechkemmer, A., & Yin, M. (2022). When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems,* (pp. 1–14). https://doi.org/10.1145/3491102.3501967

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology, 49*(1), 95–112. https://doi.org/10.1037/0022-3514.49.1.95

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Rieger, T., & Manzey, D. (2022). Human performance consequences of automated decision aids: The impact of time pressure. *Human Factors, 64*(4), 617–634. https://doi.org/10.1177/0018720820965019

Rieger, T., Roesler, E., & Manzey, D. (2022). Challenging presumed technological superiority when working with (artificial) colleagues. *Scientific Reports, 12*(1), 3768. https://doi.org/10.1038/s41598-022-07808-x

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The Urban Review, 3*(1), 16–20. https://doi.org/10.1007/BF02322211

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13*(3), 279–301. https://doi.org/10.1016/0022-1031(77)90049-X

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Introduction to special topic forum: Not so different after all: A cross-discipline view of trust. *Academy of Management Review, 23*(3), 393–404. https://doi.org/10.5465/amr.1998.926617

Roy, M. C., Dewit, O., & Aubert, B. A. (2001). The impact of interface usability on trust in web retailers. *Internet Research, 11*(5), 388–398. https://doi.org/10.1108/10662240110410165

Rüdiger, K., Cabbidu, F., Lorenz, S.-C., & Hartman, H. (2022). In *Intercultural or international perspectives in human resource management. Cross-cultural business conference 2022* (pp. 154–164). https://www.shaker.eu/en/content/catalogue/index.asp?lang=en&ID=8&ISBN=978-3-8440-8625-6&search=yes.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Saßmannshausen, T., Burggräf, P., Hassenzahl, M., & Wagner, J. (2023). Human trust in otherware – a systematic literature review bringing all antecedents together. *Ergonomics, 66*(7), 976–998. https://doi.org/10.1080/00140139.2022.2120634

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors, 58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Scharowski, N., Benk, M., Kühne, S., Wettstein, L., & Brühlmann, F. (2023). Certification labels for trustworthy AI: Insights from an empirical mixed-method study. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 248–260. https://doi.org/10.1145/3593013.359399

Schelble, B. G., Lopez, J., Textor, C., Zhang, R., McNeese, N. J., Pak, R., et al. (2022). Towards ethical AI: Empirically investigating dimensions of AI ethics, trust repair, and performance in human-AI teaming. *Human Factors.* , Article 00187208221116952. https://doi.org/10.1177/00187208221116952

Schlicker, N., & Langer, M. (2021). Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. In *Mensch und Computer 2021,* (pp. 325–329). https://doi.org/10.1145/3473856.3474018

Schoeffer, J., Jakubik, J., Vössing, M., Kühl, N., & Satzger, G. (2025). AI reliance and decision quality: Fundamentals, interdependence, and the effects of interventions. *Journal of Artificial Intelligence Research, 82*, 471–501. https://doi.org/10.1613/jair.1.15873.

Serna, I., Morales, A., Fierrez, J., & Obradovich, N. (2022). Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *Artificial Intelligence, 305*, Article 103682. https://doi.org/10.1016/j.artint.2022.103682

Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). GLocalX - from local to global explanations of black box AI models. *Artificial Intelligence, 294*, Article 103457. https://doi.org/10.1016/j.artint.2021.103457

Shin, D. (2019). Blockchain: The emerging technology of digital trust. *Telematics and Informatics, 45*(December 2019), Article 101278. https://doi.org/10.1016/j.tele.2019.101278

Spearman, C. (1961). "General intelligence" objectively determined and measured. *Appleton-Century-Crofts*. https://doi.org/10.1037/11491-006

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., et al. (2024). On the quest for effectiveness in human oversight: Interdisciplinary perspectives. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24),* (pp. 2495–2507). https://doi.org/10.1145/3630106.3659051

Textor, C., Zhang, R., Lopez, J., Schelble, B. G., McNeese, N. J., Freeman, G., et al. (2022). Exploring the relationship between ethics and trust in human-artificial intelligence teaming: A mixed methods approach. *Journal of Cognitive Engineering and Decision Making, 16*(4), 252–281. https://doi.org/10.1177/15553434221113964

Thielmann, I., & Hilbig, B. E. (2015). Trust: An integrative review from a person–situation perspective. *Review of General Psychology, 19*(3), 249–277. https://doi.org/10.1037/gpr0000046

Thompson, C., Dalgleish, L., Bucknall, T., Estabrooks, C., Hutchinson, A. M., Fraser, K., et al. (2008). The effects of time pressure and experience on nurses' risk assessment decisions: A signal detection analysis. *Nursing Research, 57*(5), 302–311. https://doi.org/10.1097/01.NNR.0000313504.37970.f9

Thompson, W. R., Reinisch, A. J., Unterberger, M. J., & Schriefl, A. J. (2019). Artificial intelligence-assisted auscultation of heart murmurs: Validation by virtual clinical trial. *Pediatric Cardiology, 40*(3), 623–629. https://doi.org/10.1007/s00246-018-2036-z

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & van Moorsel, A. (2020). The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,* (pp. 272–283). https://doi.org/10.1145/3351095.3372834

UNESCO. *UNESCO's Recommendation on the Ethics of Artificial Intelligence: Key Facts*. United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000385082.

van der Werff, L., Blomqvist, K., & Koskinen, S. (2021). Trust cues in artificial intelligence: A multilevel case study in a service organization. In *Understanding trust in organizations*. Routledge.

Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Computer-Human Interaction, 5*, 1–39. https://doi.org/10.1145/3476068. CSCW2.

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Bias preservation in machine learning: The legality of fairness metrics under EU non-discrimination law. *West Virginia Law Review, 123*(3), 735. https://doi.org/10.2139/ssrn.3792772

Wei, Q., Zeng, S.-E., Wang, L.-P., Yan, Y.-J., Wang, T., Xu, J.-W., et al. (2022). The added value of a computer-aided diagnosis system in differential diagnosis of breast lesions by radiologists with different experience. *Journal of Ultrasound in Medicine, 41*(6), 1355–1363. https://doi.org/10.1002/jum.15816

Wischnewski, M., Krämer, N., & Müller, E. (2023). 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 ,* (pp. 1–16). https://doi.org/10.1145/3544548.3581197

Wysocki, O., Davies, J. K., Vigo, M., Armstrong, A. C., Landers, D., Lee, R., et al. (2023). Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence, 316*, Article 103839. https://doi.org/10.1016/j.artint.2022.103839

Xie, C., & Wu, Y. (2019). Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 501–509). https://doi.org/10.1109/CVPR.2019.00059

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* (pp. 1–12). https://doi.org/10.1145/3290605.3300509

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns.* , Article 100455. https://doi.org/10.1016/j.patter.2022.100455

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20),* (pp. 295–305). https://doi.org/10.1145/3351095.3372852

Zicari, R. V., Brodersen, J., Brusseau, J., Düdder, B., Eichhorn, T., Ivanov, T., et al. (2021). Z-Inspection®: A process to assess trustworthy AI. *IEEE Transaactions on Technology and Society, 2*(2), 83–97. https://doi.org/10.1109/TTS.2021.3066209