Introducing the trustworthiness assessment model and its implications for research on trust in artificial intelligence

The trustworthiness assessment model

Nadine Schlicker*

Institute for AI in Medicine, Philipps-University of Marburg, nadine.schlicker@uni-marburg.de

Markus Langer

Psychology and Digitalization, Philipps-University of Marburg, markus.langer@uni-marburg.de

To advance research on trust in artificial intelligence (AI), there is a need to advance conceptual clarity regarding the main concepts in trust research. In this paper, we first distinguish trust propensity, actual and perceived trustworthiness, trust, and trusting behavior. We propose that the differentiation between actual trustworthiness of a system and perceived trustworthiness of a trustor with respect to the system is crucial to understand trust development. To describe how human trustors assess a system's actual trustworthiness to arrive at their perceived trustworthiness, we build on psychological models about how humans assess characteristics of other humans and outline the trustworthiness assessment model. We propose that trustors assess system trustworthiness based on cues associated with the system. The accuracy of this assessment depends on the relevance and availability of cues on the side of the system, and on the detection and utilization of these cues on the side of the human trustor. This paper contributes a conceptual model that advances understanding of trustworthiness assessment processes when confronted with AI systems. Additionally, we contribute a discussion on the conceptual implications of our model regarding (calibrated) trust and trusting behavior.

CCS CONCEPTS • Human-centered computing • Human computer interaction (HCI) • HCI theory, concepts and models

Additional Keywords and Phrases: Trust in automation, trustworthiness, trustworthy AI, calibrated trust, trustworthy AI

1 INTRODUCTION

Due to the rise of powerful, but imperfect artificial intelligent (AI) systems in high-risk contexts [9,25] the concept of appropriate, and calibrated trust has received increasing attention. At the same time, the concept has been criticized for being unclearly defined, operationalized, and for being imprecise about what exactly has to be "calibrated" [91]. Calibrated trust has been described as trust that matches the capabilities of the system [61] or as a match between system's actual and perceived trustworthiness [87]. In fact, many conceptualizations of calibrated trust have in common that they refer to the actual capabilities or the actual trustworthiness of systems. However, models of trust often only include a trustor's perceptions regarding the trustee's trustworthiness without explicitly stating how people arrive at those perceptions [47,52].

In the quest for a better understanding of calibrated trust, existing models on trust may thus need to be expanded by integrating the actual trustworthiness of the trustee (i.e., in our case: the system). In this paper, we provide an overview on the trustworthiness assessment model that we describe in more detail in [78]. This model adds the trustee and their actual

trustworthiness to existing trust models and describes how trustors might develop their perceived trustworthiness of a trustee. By inspecting the outcome of the trustworthiness assessment process (i.e., perceived trustworthiness), we provide ideas on how trust may be conceptualized and additionally describe potential targets for "calibration".

2 RELATED WORK

Trust needs at least two parties: one party who trusts (the trustor – in our case a human) and one party who is trusted (the trustee – in our case any (AI) system). Based on the most popular models of trust [47,52], we assume that four concepts are particularly important to trust development processes: *propensity to trust, trustworthiness*, *trust,* and *trusting behavior.*

Propensity to trust was defined as a stable disposition and "the general willingness to trust others" in human interactions [52], and similarly in human computer interactions [30,34,47]. It reflects the "generalized expectancy that others can be relied upon" [5]. Propensity to trust develops over the lifespan and may be especially influential in novel situations, in which trustees are unknown [5].

Trustworthiness has been used to refer to two sides of one coin, e.g. [36,47,49,52,77,87,88]. First, trustworthiness has been referred to as an „objective attribute of the trustee" [94]; see also [24,32,37,54] for a similar conceptualization of trustworthiness as a property of the trustee). Second, *trustworthiness* has been referred to as a subjective perception of attributes of a trustee [52]. Here, trustors need to assess the trustworthiness of a trustee to arrive at such perceptions of trustworthiness [30]. Trustworthiness perceptions reflect a trustor's expectation about the abilities, principles, and intentions of the trustee and a trustor's beliefs about whether the trustee might help the trustor to achieve their goals [5,15,46,47,52]. Perceptions of trustworthiness are situation specific and subjective [5,30,52]. Situation specific means that a trustor assesses trustworthiness of a trustee with respect to a specific task, within a specific context, and at a specific point in time [52]. Subjective means that whether a trustee is considered to be trustworthy depends on the trustor's individual cognitive and affective assessment of a system in light of the trustor's individual goals, values, and abilities in a specific situation where they consider to rely on the trustee [13,53,61].

In both specifications, trustworthiness is assumed to consist of various facets [5,18,28,32,47,52]. For instance, Mayer et al. [52] propose that ability, benevolence, and integrity, subsume all aspects of (perceived) trustworthiness in interpersonal trust. Lee and See [47] translated those trustworthiness facets to human-computer trust (referring to performance, purpose, and process) and called them "bases of trust". To clearly distinguish between those two uses of the term trustworthiness, we propose that on the side of the trustee, there exists an *actual* trustworthiness (AT) reflecting the characteristics of the system. On the side of the trustor, there exists a *perceived* trustworthiness (PT) of the system.

Trust is based on the trustor's PT [5,15,52,60]. Acknowledging the variety of trust definitions [47,52,61,68,73], most of them have three core components in common [83]: First, trust "implies uncertainty and risk, given the absence of control on the part of the trustor". Second, trust "is based on an expectation that […] the trustee will act in the trustor's interest". Third, trust "requires accepting personal vulnerability". For the purpose of the current work, we follow Mayer et al. [52] who defines 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". Trust is a downstream consequence of the trustworthiness assessment process. Specifically, PT as the outcome of the trustworthiness assessment process reflects a trustor's inferred knowledge about a trustee. However, the trustworthiness assessment process always leaves the trustor with remaining uncertainty about the trustee, for instance regarding the trustee's abilities or intentions. This also implies that the trustor will never be able to perfectly predict the trustee's behavior. Trust helps to bridge this knowledge gap and the remaining uncertainty. If the trustworthiness assessment led to perfect knowledge without uncertainty about the trustee, there would be no need for trust.

Trusting behavior as the behavioral manifestation of trust (e.g., reliance, compliance [86,87]) is positively related to a trustor's perceived trustworthiness of a trustee and their trust in the trustee [41] in situations of perceived risk [52]. To be more specific trusting behavior depends on perceived stakes, meaning that any decision to actually show trusting behavior will be coupled with an expected positive outcome, but also a potential negative outcome that reflects the trustor's vulnerability. The translation from trust to trusting behavior is thus influenced by the perceived stakes. Consequently, trust does not directly translate into trusting behavior because, for instance, if perceived risk is too high trustors may decide to do a task by themselves [10,93,95]. Conversely, sometimes behavior might appear to be trusting behavior, but might actually rather be the consequence of situational factors, such as time pressure [70], or social conformity [3]. Nevertheless, if we assume that perceptions (i.e., of trustworthiness) and intentions (i.e., trust) affect behavior, trust and trusting behavior should still be positively related [86].

To conclude, understanding trust propensity, actual trustworthiness, perceived trustworthiness, trust, and trusting behavior as well as their relations is crucial to understand trust development processes. Whereas all but actual trustworthiness are concepts that lie within the trustor, the question of whether trust is appropriate depends heavily on the characteristics of the trustee with regard to the trustor's goals. Identifying the degree of overlap of actual and perceived trustworthiness is essential for an informed trust development process. This is in line with more recent work stating that calibrated trust requires an alignment of "the perception of an actor's trustworthiness with its actual trustworthiness so that the prediction error is minimized" [87]. Consequently, when we aim for appropriate trust, we might also aim to improve the user's ability to accurately assess a decision aid's trustworthiness [61]. To do so, we propose that it is necessary to better understand the process through which AT translates into PT. In the following sections, we shed light on the process that links AT and PT and the factors that might influence the accuracy of the trustworthiness assessment.

3 THE TRUSTWORTHINESS ASSESSMENT MODEL

The trustworthiness assessment model describes the process that takes place before the existing models – i.e. the formation of trustor's PT through their assessment of a trustee's AT. Figure 1 shows how our trustworthiness assessment process relates to existing models regarding trust processes [52].

To specify the relation between AT and PT, we build on models from psychology that describe how humans assess characteristics of other humans. Specifically, we propose that Brunswik's Lens Model [26,38,42] and the Realistic Accuracy Model [22] help to understand the relation between AT and PT. These models propose processes specifying how people assess constructs that are not directly observable. For instance, mental ability and personality describe latent constructs for which there is no way to observe or measure their "true value" (i.e. their ground truth) [8]. However, we, as a society, define these constructs (e.g., early concepts for cognitive abilities [81]), refine our definition of these constructs (e.g., later concepts for cognitive abilities [12]), operationalize these constructs (e.g., in written tests or questionnaires for cognitive abilities [35,44,66]) and thus assess these constructs more or less reliably.

We also assess these constructs in interactions with people [22]. For instance, we observe that a person with a colorful wardrobe is talkative and assess that this person seems to be comparably extraverted. We thus interpret available cues (e.g., a person's behavior and their clothing) to assess a person's actual extraversion resulting in our perceived extraversion of this person. This assessment varies in its accuracy depending on a) attributes of the target whose characteristics are assessed, b) on the cues available to assess the target's characteristics, and on c) the observer who assesses the target's characteristics [22]. Regardless of the accuracy of the assessment, this assessment and the corresponding expectations can influence the assessor's subsequent assessments and the assessor's behavior towards the target [72]. In fact, a miscalibration

between a target's actual personality and an observer's perceived personality of the target may lead to challenges in the interaction between target and observer [22,31].

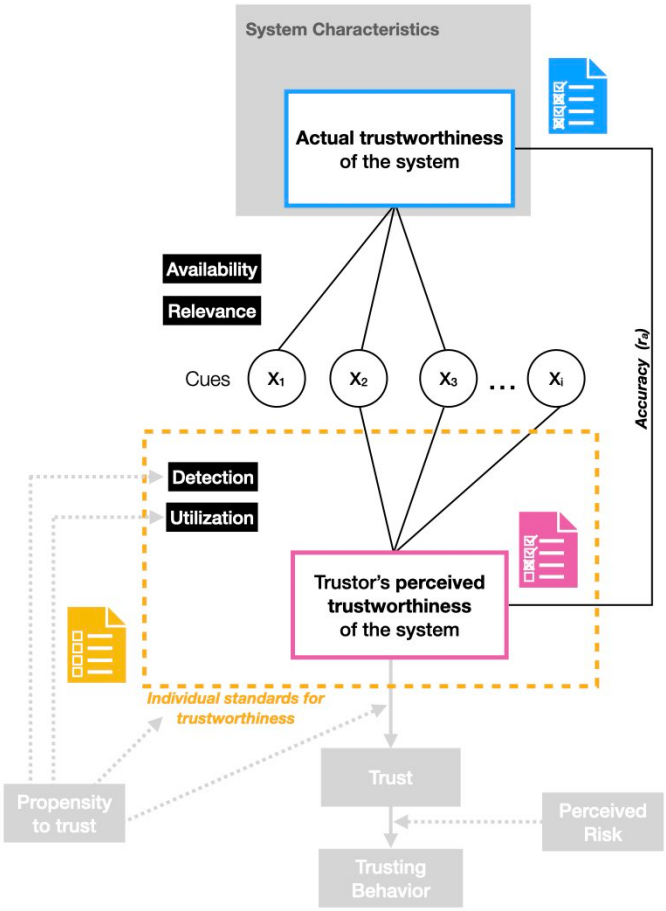

Figure 1: The trustworthiness assessment model showing the relation between actual trustworthiness (AT) and perceived trustworthiness (PT). Cues are everything a trustor uses to infer the actual trustworthiness of a system; they are thus not necessarily correlated with the AT. The lines between cues and actual trustworthiness represent correlational relations (negative and positive). The absence of lines indicates a correlation of zero. The list icons reflect our requirements list metaphor indicating that individual standards (bottom left, yellow), actual trustworthiness (top right, blue), and perceived trustworthiness (center right, purple) stand in relation to each other as depicted in Figure 2. The grey boxes at the bottom indicate how our trustworthiness assessment model relates to Mayer et al.'s [52] conceptualization of trust (trust, trust propensity, trusting behavior, and perceived risk)

We translate this to a trustor's assessment of system trustworthiness. Specifically, we propose in the trustworthiness assessment model (Figure 1) that trustors cannot directly assess a system's AT. Instead, they assess the AT of a system on the basis of available cues. Those cues are, to a certain degree, relevant to assess the system's AT. Human trustors need to detect and utilize (possibly various) cues to arrive at their PT. The degree to which a trustor's PT matches a system's AT 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 human.

3.1 Main Concepts of the Trustworthiness Assessment Model

3.1.1 Actual trustworthiness (AT).

We define a system's AT as a latent, thus not directly observable, construct that indicates the true value of a system's trustworthiness (in the sense of the realistic accuracy model [22]). AT consists of several facets that contribute to higher or lower trustworthiness and all of them could potentially be measured, but only indirectly [30,47]). For instance, we can measure system performance (e.g., predictive accuracy) or system fairness but neither can we perfectly measure these single facets (e.g., due to only being able to access a subset of all available data to assess system performance), nor can we perfectly assess how they combine to AT [30,47]). As a consequence of AT not being directly measurable, and as being user- and context-specific, trustors must continuously assess the system's AT in light of their individual standards to form their PT of the system. To further specify the concept of 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 [39,61]. Individual standards answer the question: "*What makes a system trustworthy for me*?" As a metaphor, which we will use throughout the manuscript, we can think of individual standards as a requirement list that contains all factors constituting a perfectly trustworthy system for the trustor (see Figure 1 and Figure 2). Individual standards are not arbitrary, they belong to a common understanding of the concept of trustworthiness of systems, which is also influenced by the public discussion about trustworthy AI [19,28]. Thus, individual standards might be clustered by the factors that are assumed to constitute perceived trustworthiness (e.g. ability, benevolence, integrity [52]). However, the specification of these facets of trustworthiness likely differs between different (groups of) individuals. For instance, the specification of the trustworthiness facet ability of a system depends on users' goals (e.g. screening vs. decision making), interests (e.g. saving money vs. data privacy), their personal alternatives to the system (e.g., their own ability to perform a task) [30]. Similarly, differences might appear due to different ethical and moral values reflected in trustors' individual standards (e.g., cooperation vs. competition) [82], as well as due to the normative and regulatory frame in which trustors operate (e.g., under the influence of the GDPR). For instance, referring to integrity as a set of principles the trustor finds acceptable [52], the moral machine experiment in the context of the trolley-problem in autonomous driving, showed that different cultures have different moral principles (e.g. different strengths of preferences for sparing high over low status people, or humans over pets in Eastern vs. Southern culture clusters) – in the terminology of our model this would reflect varying individual standards [4].

We refer to *system characteristics* as (only theoretically available) context-free facts that subsume everything that could theoretically be ascertained about a system. System characteristics answer the question: "*What are the characteristics of this system?*" on a descriptive level. For example, system characteristics might hold information such as the system's functionality, robustness, and reliability (in a specific dataset), but also the UI design, the company behind it, and the data used for training. Importantly, system characteristics in this (only theoretically existing) context-free space are also *objective* as they do not depend on the trustor. In contrast, AT is always subjective and thus depends on the trustor.

To bring this together, the AT of a system 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 specific task and a specific point in time. Consequently, AT answers the question: "*How trustworthy is the system actually with respect to my individual standards*?" In the requirement list metaphor, AT reflects how many checkboxes are ticked on the "individual standards requirement list" if perfect assessment was possible - which it never is, which gives rise to the need to think about PT (see Figure 2). To further clarify the

distinction, *system characteristics* contain information such as "what is the accuracy of an ADM system in a specific data set?"; *AT* answers questions such as, "is the ADM system's accuracy high enough?" What is high enough might differ for different trustors. For example, in medical diagnostics the evaluation of system accuracy may require a comparison with how well human experts perform in the same task. Whereas a novice user will perform significantly worse without the system, an expert user might perform better without the system [89]. Consequently, if both trustors assess system trustworthiness by comparing system accuracy to their own, a system with 80 % accuracy might only be judged trustworthy by the novice, not by the expert.

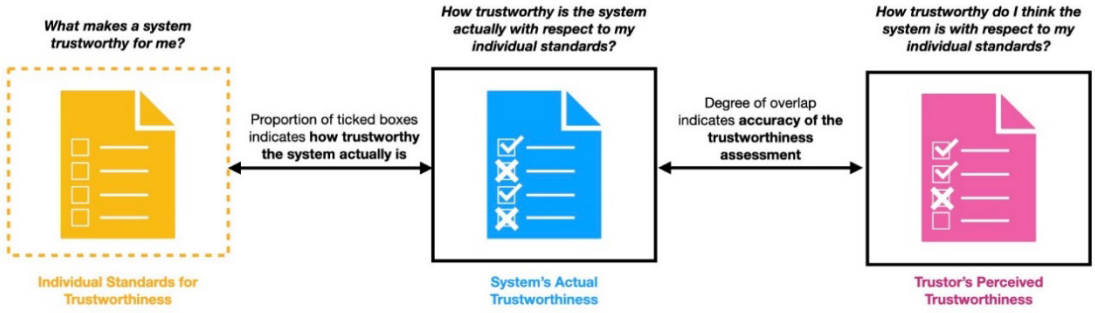

Figure 2: The requirement list metaphor describes the relationship between individual standards (left, yellow), actual trustworthiness (center, blue), and perceived trustworthiness (right, purple). In this case, the perception reflected in the first checkbox was correct, while the ones of the second and third boxes were incorrect. Regarding the fourth box the trustor was not able to assess whether the trustee matches their individual standards of trustworthiness, leaving the trustor with uncertainty regarding the trustee. Overall, the accuracy of trustworthiness assessment is suboptimal.

Over time a system's AT can be influenced by changes in the system characteristics and by changes in the individual standards [21,30,82]). Changes in the system characteristics can result from improving the system. For instance, this may relate to optimizing its technical functionalities (e.g., algorithmic fairness [7,48,79]); robustness [14,92]), explainability [69,80], and safety (for a review see [85])). Changes in the system characteristics can also be due to changes in the environment. Imagine a machine learning based decision-support system to detect different types of viruses. As the virus mutates, this reduces the algorithm's accuracy and thus changes its system characteristics [29]. Changes in the individual standards can be due to new experiences or knowledge the trustor gains. For instance, a trustor's individual standard with respect to system accuracy may increase because they increased their expertise in a task (e.g., a novice who increases their diagnostic accuracy). Trustors could also lower their individual standards which would increase the system's AT [85].

3.1.2 Perceived trustworthiness (PT).

PT describes a trustor's resulting assessment of actual system trustworthiness. Consequently, PT answers the question "*How trustworthy do I think the system is with respect to my individual standards*?" PT is another latent construct that is not directly observable but that can 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 [1,71]) or by observing people's interactions with systems [76,95]). PT is a trustor's assessment of the AT of the system based on a cognitive and affective evaluation of cues presumably associated with the system [5,47,51]. This assessment takes place on various facets that contribute to overall PT. Research indicates that these facets could include, for example, the functionality of a system, fairness of the system, or the purpose for which the system was developed [5,47,55]. Since AT

is not directly accessible, trustors use available cues to form their PT. Speaking in our requirement list metaphor, PT concerns how many boxes a trustor marks or leaves blank. Perceptions that indicate trustworthiness are checked boxes, perceptions that indicate lack of trustworthiness are crossed out boxes, and uncertainty regarding the trustee is reflected as empty boxes. It is important to emphasize again that there remains uncertainty regarding the trustee after the trustworthiness assessment. Metaphorically speaking, trustors might not be able to mark some of the boxes after the assessment or may remain unsure about whether they should mark one of the boxes. To be clear, PT does not concern which boxes would have to be marked with perfect knowledge of the system characteristics – this would correspond to the AT of the system.

3.1.3 Cues.

Cues form the interface between AT and PT. A cue is an "information element that can be used to make a trust assessment about an agent" [88]. Cues are pieces of information that presumably provide insights regarding the AT of a system [11,49,88]. Single cues may only provide narrow or even misleading insights regarding a system's AT, and each cue relates to a certain degree to the AT. Thus, trustors are constantly searching for, confronted with, (consciously or unconsciously) using, and interpreting cues to assess systems' AT. The trustor's PT is determined by how they actively and passively select and weight cues. In line with this, research investigated how different cues are weighted and associated with trust in decision aids [16].

Many aspects can be considered as cues to assess system trustworthiness. For example, cues can be the aesthetics of a user interface, any information included in the system's user manual, information about the inputs a system uses, single outputs of a system, the indicated predictive power of a classifier, information about uncertainty accompanying a classification output, a displayed or communicated rationale for the system's recommendation, a seal indicating trustworthy AI, or the logo of a company [11,49,88,90]. Cues can also stem from other people (e.g. testimonies of co-workers).[1]

3.2 Relations Between the Model Components and Influencing Factors

Trustworthiness assessment is accurate when the trustor's PT matches the system's AT (Figure 1). According to Funder, this accuracy depends on relevance and availability of cues on the system's side, and on detection and utilization of cues on the trustor's side.

On the side of the system, *cue relevance* defines how indicative a cue is for the AT of a system. For instance, a relevant cue for AT can be information on a system's performance indicators (e.g., precision, recall, F1-Score [58]). A less relevant cue may be the popularity of a brand [32,74]. Cues may also not be in any relation with AT although trustors might use them to form their PT. For example, anthropomorphization was often found to relate to people's trust in systems [75]. However, since every system can be anthropomorphized without changing its actual capabilities, anthropomorphization may actually not be relevant to AT.

Cue availability refers to the fact that relevant cues need to be accessible to the trustor. For example, the quality of a training data set might be strongly related to AT of a system but users might not have access to such information without digging deep into the technical documentation of the system. Cue availability might be increased by, e.g., model cards [59] and fact sheets that highlight, i.a. information on the purpose, algorithm, training and test dataset, potential bias, and model development of the system [2,6].

[1] In [78] we also describe the macro level of the overall trustworthiness assessment model and provide a more detailed description on how cues may be propagated between stakeholders, i.e. how one stakeholder's trustworthiness assessment of a system affects others' trustworthiness assessments of the same system. We differentiate cues that stem directly from the system (primary cues) and those that stem from other trustors (secondary cues).

On the side of the trustor, *cue detection* means that relevant and available cues must be detected by the trustor. Factors that may influence the detection of cues are attention [27], situation awareness [20], time pressure [70], or experience with a system [84]. Furthermore, a trustor's individual standards could top-down guide user's attention and detection of trustworthiness cues [23,33]. Beyond that, user interface properties such as low contrast could make cue detection more difficult.

Cue utilization means that trustors need to correctly interpret a relevant, available, and detected cue. In other words, even if relevant information is available and detected, trustors need to weigh this information appropriately. For instance, inaccuracies can be due to implicit attitudes, [43,56,57], cognitive biases [50,64], contextual misinterpretation [63], little domain-, task-, and system-knowledge [57,95], as well as due to an too strong weighting of a small sample of personal experience compared to information about a large, potentially more representative data set [17,67,93]. Further evidence for inadequate cue utilization [10] showed that although relevant cues (in their case, example-based explanations that actually increase joint performance of human-AI teams) were available, participants had higher preference for and higher trust in less relevant cues (rule-based explanations, which did not lead to better joint performance of human-AI teams). To conclude, when incorrect assumptions about the relevance of individual cues exist this can lead to inadequate assessments of system's AT (e.g., the assumption that a high-quality user interface or the logo of a popular company indicates high system-performance [32,40].

Funder [22] underlines that all factors, i.e. relevance, availability, detection, and utilization, need to be considered to evaluate whether it is possible to achieve an accurate assessment of a target's characteristics. In our case, if only irrelevant cues are available, this will prevent accurate assessment of AT. If no cues are available or detected, it is hard to accurately assess system trustworthiness. Finally, the incorrect utilization of relevant, available and detected cues will also lead to a low accuracy of the trustworthiness assessment.

To sum up, the result of the trustworthiness assessment process under the influence of availability, relevance, detection, and utilization of cues, is the trustor's perception of the trustee's trustworthiness, the trustee's lack of trustworthiness, and uncertainty regarding the trustee.

4 CONCEPTUAL IMPLICATIONS

4.1 Trustworthiness assessment as an integral part of trust processes

In seminal models on trust (i.e. [47,52]) there formerly existed a gap as those models only included trustors' PT without describing how PT may be affected by the trustees AT. In [78] and in this paper, we describe the trustworthiness assessment model that augments existing trust models (e.g.[47,52]) by adding the trustee (and its characteristics), spelling out the concepts that are 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. Including the trustworthiness assessment as an explicit part of trust development processes may help to guide future research in understanding the transition from PT to trust and trusting behavior. Building on the insights of the trustworthiness assessment model we now want to sketch how PT, trust, and trusting behavior may be related.

Figure 3 serves to make those relations graspable. The circle is filled with perceptions that indicate a system's trustworthiness (light grey; checked boxes in the perceived trustworthiness list), perceptions that reflect a lack of trustworthiness (dark grey; crossed out boxes), and uncertainty regarding trustworthiness of a trustee (blue; empty boxes). The assessments of perceived trustworthiness and perceived lack of trustworthiness can be accurate to a certain degree, indicated by the green (accurate) and red (inaccurate) areas at the outside of the circle. Note that knowing about the (lack

of) trustworthiness of a trustee reduces uncertainty regarding the trustee. Perception is the result of a weighted assessment of cues. The weight reflects the subjective importance with which a cue influences perceptions. For example, there might be red flag cues that more strongly influence perceptions than others. This could strongly increase the perception of lack of trustworthiness and at the same time decrease the proportion of perceived trustworthiness. To conclude, single cues might heavily shift the proportions of trustworthiness versus lack of trustworthiness perceptions and at the same time the remaining uncertainty regarding the system may remain nearly similar.

Trust, defined as the willingness to be vulnerable to the trustee, in this concept is depicted as the dashed line. This dashed line depicts how much uncertainty (regarding the trustworthiness of a trustee) a trustor is willing to accept. If the area of the circle that is filled with perceptions indicating trustworthiness passes the threshold indicated by the dashed line, the trustor will be in a position to consider taking the risk and engaging in trusting behavior towards the trustee, i.e. actually make themselves vulnerable to the trustee. If perceptions that indicate a system's trustworthiness do not fill enough space to reach the threshold, the trustor will not be willing to be vulnerable to the trustee.

The position of the dashed line might be affected by the trustor's propensity to trust. A high propensity to trust (i.e., a higher propensity to be willing to be vulnerable) moves this dashed line counterclockwise. In other words, the trustor will be more likely to become vulnerable, even with comparably low perceived knowledge about the trustee's trustworthiness.

In case a trustor is actually willing to be vulnerable, in case the trustor considers to engage in trusting behavior, stakes of the situation come into play (i.e., possible positive as well as possible negative outcomes). If the likelihood or strength of negative outcomes increases, this moves the dashed line clockwise making it less likely that the trustor will actually show trusting behavior; if the likelihood or strength of positive outcomes increases, this moves the dashed line counterclockwise increasing the likelihood of the trustor to show trusting behavior. The proportions of correct and incorrect assessments reflect how accurately a trustor assesses the AT of a system and indicate whether trust and potential trusting behavior are well-informed and appropriate given trustor's individual standards.

What is sketched in Figure 3 and what was outlined before is not intended as a full-fledged theory on the relations between perceived trustworthiness, trust, and trusting behavior. Instead, we hope that it stimulates discussions on the relation between those concepts. Future research could build on these ideas to spell out concrete propositions.

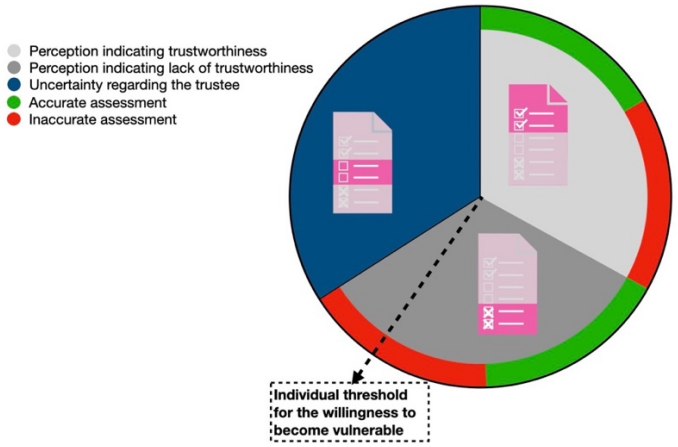

Figure 3: The graph sketches the expected outcome of the trustworthiness assessment and its relation to trust. The circle is filled with perceptions that indicate a system's trustworthiness (light grey), perceptions that reflect a lack of trustworthiness (dark grey), and uncertainty regarding trustworthiness of a trustee (blue). In our requirement list metaphor, perceptions that indicate trustworthiness are

checked boxes, perceptions that indicate lack of trustworthiness are crossed out boxes, and uncertainty regarding the trustee is reflected as empty boxes. The assessments of perceived trustworthiness and perceived lack 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 dashed line indicates a trustor's individual threshold for the willingness to become vulnerable.

4.2 Implications for Calibrated Trust

By spelling out the trustworthiness assessment process, and by spelling out proposed relations between AT, PT, trust, and trusting behavior, this paper may contribute to advance our understanding of the concept of "calibrated trust".

Up front, we cannot provide a clear definition of calibrated trust, we can only provide conclusions from the trustworthiness assessment model about what calibrated trust may *not be* and how *not to measure* it. Calibrated trust has been defined as a match of a system's trustworthiness and a trustor's trust in the system (e.g., [45,47]). As we have outlined in this paper, there may be a conceptual gap if we assume that AT directly translates into trust. It may be more adequate to assume that AT will be assessed and leads to PT and that PT then affects people's trust in systems. In other work, calibrated trust has been defined as a match of AT and PT [87]. We propose that a strong match between AT and PT may more simply be described as a high accuracy of the trustworthiness assessment. Finally, instead of trying to capture anything related to trustworthiness or trust, most commonly, calibrated trust was measured by observing people's trusting behavior with respect to a system (e.g., rejecting vs. accepting system outputs; [62,95]). Thus, many calibrated trust measures to date might rather reflect "calibrated trusting behavior" or even just "behavior that reflects how well people can differentiate correct and incorrect outputs". In line with this, research sometimes refers to calibrated trust in a way that sounds as if perfectly calibrated trust is theoretically possible [47,87]. At this point, we must be careful not to be tempted to think about calibrated trust as a state of being able to tell exactly when to rely on a system and when not to. This may work for hard-coded, highly predictable systems [30,65], but not for AI systems for which uncertainty is inherent. If system behavior is completely predictable, if people always know whether to reject or accept system outputs, there would be no uncertainty left, and consequently trust would be obsolete. This raises the question whether trust really is miscalibrated if it is based on a correct trustworthiness assessment (reflecting everything that the trustor could have known about the system to this point), but the outcome in a specific situation is undesirable (e.g., the system output is incorrect)?

We propose at this point that the accuracy of the trustworthiness assessment is a prerequisite to whatever calibrated trust actually is. An inaccurate trustworthiness assessment might wrongfully increase or decrease the perceived trustworthiness and thereby increase the probability for a violation of expectations or even perceived betrayal with respect to the trustee. In Figure 3 we identified three factors that might be *calibrated* to achieve an "optimal" human-system interaction: the accuracy of the trustworthiness assessment, the uncertainty regarding the trustee, and the trustor's threshold for willingness to become vulnerable. Across the paper, we mainly provide ideas how to influence the former two variables. Specifically, factors that affect the accuracy of the trustworthiness assessment and the uncertainty regarding the trustee on the system side are relevance and availability of cues, and on the trustor's side detection and utilization of cues.

For future research it becomes paramount to clearly distinguish the aspects that contribute to calibrated trust beyond an accurate trustworthiness assessment, the factors that influence calibrated trust and hence understand how it can be defined in a way that allows its measurement. We hope that our trustworthiness assessment model and the related thoughts regarding its outcome (see Figure 3) might advance conceptual clarity regarding trust in AI systems and provides nutritious ground for discussions of the nature of trust and the related concepts.

REFERENCES

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

[2] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Darrell Reimer, Alexandra Olteanu, David Piorkowski, Jason Tsay, and Kush R. Varshney. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. Retrieved June 2, 2022 from http://arxiv.org/abs/1808.07261

[3] Solomon E. Asch. 1955. Opinions and Social Pressure. *Scientific American* 193, 5 (1955), 31–35.

[4] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine experiment. *Nature* 563, 7729 (November 2018), 59–64. DOI:https://doi.org/10.1038/s41586-018-0637-6

[5] Michael D. Baer and Jason A. Colquitt. 2018. Why do people trust?: Moving toward a more comprehensive consideration of the antecedents of trust. In *The Routledge companion to trust* (1st ed.), Rosalind H. Searle, Ann-Marie I. Nienaber and Sim B. Sitkin (eds.). Routledge, New York : Routledge, 2017, 163–182. DOI:https://doi.org/10.4324/9781315745572-12

[6] Nathalie Baracaldo, Ali Anwar, Mark Purcell, Ambrish Rawat, Mathieu Sinn, Bashar Altakrouri, Dian Balta, Mahdi Sellami, Peter Kuhn, Ulrich Schopp, and Matthias Buchinger. 2022. Towards an accountable and reproducible federated learning: A FactSheets approach. Retrieved June 2, 2022 from http://arxiv.org/abs/2202.12443

[7] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv:1810.01943* (October 2018). Retrieved from http://arxiv.org/abs/1810.01943

[8] Denny Borsboom, Gideon J. Mellenbergh, and Jaap van Heerden. 2003. The theoretical status of latent variables. *Psychological Review* 110, (2003), 203–219. DOI:https://doi.org/10.1037/0033-295X.110.2.203

[9] Titus J. Brinker, Achim Hekler, Alexander H. Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Dirk Schadendorf, Tim Holland-Letz, Jochen S. Utikal, Christof von Kalle, Wiebke Ludwig-Peitsch, Judith Sirokay, Lucie Heinzerling, Magarete Albrecht, Katharina Baratella, Lena Bischof, Eleftheria Chorti, Anna Dith, Christina Drusio, Nina Giese, Emmanouil Gratsias, Klaus Griewank, Sandra Hallasch, Zdenka Hanhart, Saskia Herz, Katja Hohaus, Philipp Jansen, Finja Jockenhöfer, Theodora Kanaki, Sarah Knispel, Katja Leonhard, Anna Martaki, Liliana Matei, Johanna Matull, Alexandra Olischewski, Maximilian Petri, Jan-Malte Placke, Simon Raub, Katrin Salva, Swantje Schlott, Elsa Sody, Nadine Steingrube, Ingo Stoffels, Selma Ugurel, Anne Zaremba, Christoffer Gebhardt, Nina Booken, Maria Christolouka, Kristina Buder-Bakhaya, Therezia Bokor-Billmann, Alexander Enk, Patrick Gholam, Holger Hänßle, Martin Salzmann, Sarah Schäfer, Knut Schäkel, Timo Schank, Ann-Sophie Bohne, Sophia Deffaa, Katharina Drerup, Friederike Egberts, Anna-Sophie Erkens, Benjamin Ewald, Sandra Falkvoll, Sascha Gerdes, Viola Harde, Axel Hauschild, Marion Jost, Katja Kosova, Laetitia Messinger, Malte Metzner, Kirsten Morrison, Rogina Motamedi, Anja Pinczker, Anne Rosenthal, Natalie Scheller, Thomas Schwarz, Dora Stölzl, Federieke Thielking, Elena Tomaschewski, Ulrike Wehkamp, Michael Weichenthal, Oliver Wiedow, Claudia Maria Bär, Sophia Bender-Säbelkampf, Marc Horbrügger, Ante Karoglan, Luise Kraas, Jörg Faulhaber, Cyrill Geraud, Ze Guo, Philipp Koch, Miriam Linke, Nolwenn Maurier, Verena Müller, Benjamin Thomas, Jochen Sven Utikal, Ali Saeed M. Alamri, Andrea Baczako, Carola Berking, Matthias Betke, Carolin Haas, Daniela Hartmann, Markus V. Heppt, Katharina Kilian, Sebastian Krammer, Natalie Lidia Lapczynski, Sebastian Mastnik, Suzan Nasifoglu, Cristel Ruini, Elke Sattler, Max Schlaak, Hans Wolff, Birgit Achatz, Astrid Bergbreiter, Konstantin Drexler, Monika Ettinger, Sebastian Haferkamp, Anna Halupczok, Marie Hegemann, Verena Dinauer, Maria Maagk, Marion Mickler, Biance Philipp, Anna Wilm, Constanze Wittmann, Anja Gesierich, Valerie Glutsch, Katrin Kahlert, Andreas Kerstan, Bastian Schilling, and Philipp Schrüfer. 2019. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer* 113, (May 2019), 47–54. DOI:https://doi.org/10.1016/j.ejca.2019.04.001

[10] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, ACM, Cagliari Italy, 454–464. DOI:https://doi.org/10.1145/3377325.3377498

[11] George J. Cancro, Shimei Pan, and James Foulds. 2022. Tell me something that will help me trust you: A survey of trust calibration in human-agent interaction. Retrieved July 5, 2022 from http://arxiv.org/abs/2205.02987

[12] John B. Carroll. 1993. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge University Press.

[13] Erin K. Chiou and John D. Lee. 2021. Trusting automation: Designing for responsivity and resilience. *Hum Factors* (April 2021), 001872082110099. DOI:https://doi.org/10.1177/00187208211009995

[14] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. 2017. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning* (Proceedings of Machine Learning Research), PMLR, 854–863. Retrieved from https://proceedings.mlr.press/v70/cisse17a.html

[15] Jason A. Colquitt, Brent A. Scott, and Jeffery A. LePine. 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 (2007), 909–927. DOI:https://doi.org/10.1037/0021-9010.92.4.909

[16] Dan Conway, Fang Chen, Kun Yu, Jianlong Zhou, and Richard Morris. 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*, ACM, San Jose California USA, 3035–3041. DOI:https://doi.org/10.1145/2851581.2892433

[17] Berkeley J. Dietvorst and Soaham Bharti. 2020. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychol Sci* 31, 10 (October 2020), 1302–1314. DOI:https://doi.org/10.1177/0956797620948841

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

[19] Karolina Drobotowicz, Marjo Kauppinen, and Sari Kujala. 2021. Trustworthy AI services in the public sector: What are citizens saying about It? In *Requirements Engineering: Foundation for Software Quality*, Fabiano Dalpiaz and Paola Spoletini (eds.). Springer International Publishing, Cham, 99–115. DOI:https://doi.org/10.1007/978-3-030-73128-1_7

[20] Mica R. Endsley. 2017. From here to autonomy: Lessons learned from human–automation research. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 59, 1 (February 2017), 5–27. DOI:https://doi.org/10.1177/0018720816681350

[21] Christopher Flathmann, Beau G. Schelble, Rui Zhang, and Nathan J. McNeese. 2021. Modeling and Guiding the Creation of Ethical Human-AI Teams. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, ACM, Virtual Event USA, 469–479. DOI:https://doi.org/10.1145/3461702.3462573

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

[23] Thomas Geyer and Hermann J. Müller. 2009. Distinct, but top-down modulable color and positional priming mechanisms in visual pop-out search. *Psychological Research* 73, 2 (March 2009), 167–176. DOI:https://doi.org/10.1007/s00426-008-0207-x

[24] Ben Green. 2022. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45, (July 2022), 105681. DOI:https://doi.org/10.1016/j.clsr.2022.105681

[25] Nina Grgić-Hlača, Christoph Engel, and Krishna P. Gummadi. 2019. Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (November 2019), 1–25. DOI:https://doi.org/10.1145/3359280

[26] Kenneth R. Hammond. 1996. *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice*. Oxford University Press.

[27] Harold Hawkins, Steven Hillyard, Steven Luck, Mustapha Mouloua, Cathryn Downing, and Donald Woodward. 1990. Visual attention modulates signal detectability. *Journal of experimental psychology. Human perception and performance* 16, (December 1990), 802–11. DOI:https://doi.org/10.1037/0096-1523.16.4.802

[28] High-Level Expert Group on Artificial Intelligence. 2019. Ethics guidelines for trustworthy AI.

[29] T. Ryan Hoens, Robi Polikar, and Nitesh V. Chawla. 2012. Learning from streaming data with concept drift and imbalance: an overview. *Prog Artif Intell* 1, 1 (April 2012), 89–101. DOI:https://doi.org/10.1007/s13748-011-0008-0

[30] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, 3 (May 2015), 407–434. DOI:https://doi.org/10.1177/0018720814547570

[31] Lauren J. Human and Jeremy C. Biesanz. 2013. Targeting the good target: An integrative review of the characteristics and consequences of being accurately perceived. *Pers Soc Psychol Rev* 17, 3 (August 2013), 248–272. DOI:https://doi.org/10.1177/1088868313495593

[32] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ACM, Virtual Event Canada, 624–635. DOI:https://doi.org/10.1145/3442188.3445923

[33] Melinda S. Jensen, Richard Yao, Whitney N. Street, and Daniel J. Simons. 2011. Change blindness and inattentional blindness. *WIREs Cognitive Science* 2, 5 (2011), 529–546. DOI:https://doi.org/10.1002/wcs.130

[34] Sarah A. Jessup, Tamera R. Schneider, Gene M. Alarcon, Tyler J. Ryan, and August Capiola. 2019. The Measurement of the Propensity to Trust Automation. In *Virtual, Augmented and Mixed Reality. Applications and Case Studies*, Jessie Y.C. Chen and Gino Fragomeni (eds.). Springer International Publishing, Cham, 476–489. DOI:https://doi.org/10.1007/978-3-030-21565-1_32

[35] Alan S. Kaufman, Dawn P. Flanagan, Vincent C. Alfonso, and Jennifer T. Mascolo. 2006. Test review: Wechsler intelligence scale for children, (WISC-IV). *Journal of Psychoeducational Assessment* 24, 3 (September 2006), 278–295. DOI:https://doi.org/10.1177/0734282906288389

[36] Matthew Kay, Shwetak N. Patel, and Julie A. Kientz. 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* (CHI '15), Association for Computing Machinery, New York, NY, USA, 347–356. DOI:https://doi.org/10.1145/2702123.2702603

[37] Christoph Kelp and Mona Simion. 2022. What is trustworthiness? *Noûs* (May 2022). Retrieved September 7, 2022 from http://eprints.gla.ac.uk/270906/

[38] Alex Kirlik and Aviation Human Factors Division Alex Kirlik. 2006. *Adaptive Perspectives on Human-Technology Interaction: Methods and Models for Cognitive Engineering and Human-Computer Interaction*. Oxford University Press, USA.

[39] Bran Knowles and John T. Richards. 2021. The sanction of authority: Promoting public trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ACM, Virtual Event Canada, 262–271. DOI:https://doi.org/10.1145/3442188.3445890

[40] Yoon Jeon Koh and S. Shyam Sundar. 2010. Effects of specialization in computers, web sites, and web agents on e-commerce trust. *International Journal of Human-Computer Studies* 68, 12 (December 2010), 899–912. DOI:https://doi.org/10.1016/j.ijhcs.2010.08.002

[41] Moritz Körber. 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)*, Springer International Publishing, Cham, 13–30. DOI:https://doi.org/10.1007/978-3-319-96074-6_2

[42] Nathan R. Kuncel. 2018. Judgment and decision making in staffing research and practice. In *The SAGE handbook of industrial, work & organizational psychology: Personnel psychology and employee performance* (2nd ed.). Sage Reference, 474–488.

[43] Markus Langer, Cornelius J. König, Caroline Back, and Victoria Hemsing. 2022. Trust in Artificial Intelligence: Comparing Trust Processes Between Human and Automated Trustees in Light of Unfair Bias. *J Bus Psychol* (June 2022). DOI:https://doi.org/10.1007/s10869-022-09829-9

[44] Jeff Laurent, Mark Swerdlik, and Mary Ryburn. 1992. Review of validity research on the Stanford-Binet Intelligence Scale: Fourth Edition. *Psychological Assessment* 4, 1 (March 1992), 102–112. DOI:https://doi.org/10.1037/1040-3590.4.1.102

[45] Johann Laux, Sandra Wachter, and Brent Mittelstadt. 2023. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance* n/a, n/a (2023). DOI:https://doi.org/10.1111/rego.12512

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

[47] John D. Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80. DOI:https://doi.org/10.1518/hfes.46.1.50_30392

[48] Xiaoxiao Li, Ziteng Cui, Yifan Wu, Lin Gu, and Tatsuya Harada. 2021. Estimating and improving fairness with adversarial learning. Retrieved June 7, 2021 from http://arxiv.org/abs/2103.04243

[49] Q. Vera Liao and S. Shyam Sundar. 2022. *Designing for responsible trust in AI systems: A communication perspective*. DOI:https://doi.org/10.1145/3531146.3533182

[50] Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. 2019. Resistance to medical artificial intelligence. *Journal of Consumer Research* 46, 4 (December 2019), 629–650. DOI:https://doi.org/10.1093/jcr/ucz013

[51] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Citeseer, 6–8.

[52] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *The Academy of Management Review* 20, 3 (1995), 709–734. DOI:https://doi.org/10.2307/258792

[53] Daniel J. McAllister. 1995. Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations. *The Academy of Management Journal* 38, 1 (1995), 24–59. DOI:https://doi.org/10.2307/256727

[54] Carolyn McLeod. 2021. Trust. In *The Stanford Encyclopedia of Philosophy* (Fall 2021), Edward N. Zalta (ed.). Metaphysics Research Lab, Stanford University. Retrieved September 7, 2022 from https://plato.stanford.edu/archives/fall2021/entriesrust/

[55] Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. 2021. More similar values, more trust? - the effect of value similarity on trust in human-agent interaction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, ACM, Virtual Event USA, 777–783. DOI:https://doi.org/10.1145/3461702.3462576

[56] Stephanie M. Merritt, Heather Heimbaugh, Jennifer LaChapell, and Deborah Lee. 2013. I Trust It, but I Don't Know Why: Effects of Implicit Attitudes Toward Automation on Trust in an Automated System. *Hum Factors* 55, 3 (June 2013), 520–534. DOI:https://doi.org/10.1177/0018720812465081

[57] Stephanie M. Merritt, Jennifer L. Unnerstall, Deborah Lee, and Kelli Huber. 2015. Measuring Individual Differences in the Perfect Automation Schema. *Hum Factors* 57, 5 (August 2015), 740–753. DOI:https://doi.org/10.1177/0018720815581247

[58] Jiaju Miao and Wei Zhu. 2022. Precision–recall curve (PRC) classification trees. *Evol. Intel.* 15, 3 (September 2022), 1545–1569. DOI:https://doi.org/10.1007/s12065-021-00565-2

[59] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (FAT* '19), Association for Computing Machinery, New York, NY, USA, 220–229. DOI:https://doi.org/10.1145/3287560.3287596

[60] Guido Möllering. 2006. *Trust: Reason, Routine, Reflexivity*. Emerald Group Publishing.

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

[62] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *PLOS ONE* 15, 2 (February 2020), e0229132. DOI:https://doi.org/10.1371/journal.pone.0229132

[63] Andrea Papenmeier, Dagmar Kern, Daniel Hienert, Yvonne Kammerer, and Christin Seifert. 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), Association for Computing Machinery, New York, NY, USA, 1–13. DOI:https://doi.org/10.1145/3491102.3501915

[64] Raja Parasuraman and Dietrich H. Manzey. 2010. Complacency and bias in human use of automation: An attentional integration. *Hum Factors* 52, 3 (June 2010), 381–410. DOI:https://doi.org/10.1177/0018720810376055

[65] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39, 2 (June 1997), 230–253. DOI:https://doi.org/10.1518/001872097778543886

[66] Peter N. Prewett. 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 (March 1995), 69–72. DOI:https://doi.org/10.1037/1040-3590.7.1.69

[67] Amy Rechkemmer and Ming Yin. 2022. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *CHI Conference on Human Factors in Computing Systems*, ACM, New Orleans LA USA, 1–14. DOI:https://doi.org/10.1145/3491102.3501967

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

[69] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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* (KDD '16), Association for Computing Machinery, New York, NY, USA, 1135–1144. DOI:https://doi.org/10.1145/2939672.2939778

[70] Tobias Rieger and Dietrich Manzey. 2022. Human Performance Consequences of Automated Decision Aids: The Impact of Time Pressure. *Hum Factors* 64, 4 (June 2022), 617–634. DOI:https://doi.org/10.1177/0018720820965019

[71] Tobias Rieger, Eileen Roesler, and Dietrich Manzey. 2022. Challenging presumed technological superiority when working with (artificial) colleagues. *Sci Rep* 12, 1 (March 2022), 3768. DOI:https://doi.org/10.1038/s41598-022-07808-x

[72] Robert Rosenthal and Lenore Jacobson. 1968. Pygmalion in the classroom. *Urban Rev* 3, 1 (September 1968), 16–20. DOI:https://doi.org/10.1007/BF02322211

[73] Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. 1998. Introduction to special topic forum: Not so different after all: A cross-discipline view of trust. *The Academy of Management Review* 23, 3 (1998), 393–404. DOI:https://doi.org/10.5465/amr.1998.926617

[74] Marie Roy Christine, Olivier Dewit, and Benoit A. Aubert. 2001. The impact of interface usability on trust in Web retailers. *Internet Research* 11, 5 (January 2001), 388–398. DOI:https://doi.org/10.1108/10662240110410165

[75] Kristin E. Schaefer, Jessie Y. C. Chen, James L. Szalma, and P. A. Hancock. 2016. A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Hum Factors* 58, 3 (May 2016), 377–400. DOI:https://doi.org/10.1177/0018720816634228

[76] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. Should I Follow AI-based Advice? Measuring Appropriate Reliance in Human-AI Decision-Making. *arXiv preprint arXiv:2204.06916}* (2022), 10. DOI:https://doi.org/DOI: 10.5445/IR/1000145647

[77] Nadine Schlicker and Markus Langer. 2021. Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. In *Mensch und Computer 2021*, ACM, Ingolstadt Germany, 325–329. DOI:https://doi.org/10.1145/3473856.3474018

[78] Nadine Schlicker, Alarith Uhde, Kevin Baum, Martin C. Hirsch, and Markus Langer. 2022. Calibrated Trust as a Result of Accurate Trustworthiness Assessment – Introducing the Trustworthiness Assessment Model. DOI:https://doi.org/10.31234/osf.io/qhwvx

[79] Ignacio Serna, Aythami Morales, Julian Fierrez, and Nick Obradovich. 2022. Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *Artificial Intelligence* 305, (April 2022), 103682. DOI:https://doi.org/10.1016/j.artint.2022.103682

[80] Mattia Setzu, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2021. GLocalX - From Local to Global Explanations of Black Box AI Models. *Artificial Intelligence* 294, (May 2021), 103457. DOI:https://doi.org/10.1016/j.artint.2021.103457

[81] C Spearman. 1961. *General Intelligence" Objectively Determined and Measured"*. Appleton-Century-Crofts, East Norwalk, CT, US. DOI:https://doi.org/10.1037/11491-006

[82] Claire Textor, Rui Zhang, Jeremy Lopez, Beau G. Schelble, Nathan J. McNeese, Guo Freeman, Richard Pak, Chad Tossell, and Ewart J. de Visser. 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 (December 2022), 252–281. DOI:https://doi.org/10.1177/15553434221113964

[83] Isabel Thielmann and Benjamin E. Hilbig. 2015. Trust: An Integrative Review from a Person–Situation Perspective. *Review of General Psychology* 19, 3 (September 2015), 249–277. DOI:https://doi.org/10.1037/gpr0000046

[84] Carl Thompson, Len Dalgleish, Tracey Bucknall, Carole Estabrooks, Alison M. Hutchinson, Kim Fraser, Rien de Vos, Jan Binnekade, Gez Barrett, and Jane Saunders. 2008. The Effects of Time Pressure and Experience on Nurses' Risk Assessment Decisions: A Signal Detection Analysis. *Nursing Research* 57, 5 (October 2008), 302–311. DOI:https://doi.org/10.1097/01.NNR.0000313504.37970.f9

[85] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ACM, Barcelona Spain, 272–283. DOI:https://doi.org/10.1145/3351095.3372834

[86] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (October 2021), 1–39. DOI:https://doi.org/10.1145/3476068

[87] Ewart de Visser, Marieke M.M. Peeters, Malte Jung, Spencer Kohn, Tyler Shaw, Richard Pak, and Mark Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics* 12, (May 2020). DOI:https://doi.org/10.1007/s12369-019-00596-x

[88] Ewart J. de Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. 2014. A design methodology for trust cue calibration in cognitive agents. In *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments*, Randall Shumaker and Stephanie Lackey (eds.). Springer International Publishing, Cham, 251–262. DOI:https://doi.org/10.1007/978-3-319-07458-0_24

[89] Qi Wei, Shu-E Zeng, Li-Ping Wang, Yu-Jing Yan, Ting Wang, Jian-Wei Xu, Meng-Yi Zhang, Wen-Zhi Lv, Christoph F. Dietrich, and Xin-Wu Cui. 2022. The added value of a computer-aided diagnosis system in differential diagnosis of breast lesions by radiologists with different experience. *J of Ultrasound Medicine* 41, 6 (June 2022), 1355–1363. DOI:https://doi.org/10.1002/jum.15816

[90] Lisa van der Werff, Kirsimarja Blomqvist, and Sirpa Koskinen. 2021. Trust Cues in Artificial Intelligence: A Multilevel Case Study in a Service Organization. In *Understanding Trust in Organizations*. Routledge.

[91] Magdalena Wischnewski and Nicole Krämer. 2023. *Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions*. PsyArXiv. DOI:https://doi.org/10.31234/osf.io/qt86s

[92] Cihang Xie and Yuxin Wu. 2019. Feature denoising for improving adversarial robustness}. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 501–509. DOI:https://doi.org/10.1109/CVPR.2019.00059

[93] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ACM, Glasgow Scotland Uk, 1–12. DOI:https://doi.org/10.1145/3290605.3300509

[94] John Zerilli, Umang Bhatt, and Adrian Weller. 2022. How transparency modulates trust in artificial intelligence. *Patterns* (February 2022), 100455. DOI:https://doi.org/10.1016/j.patter.2022.100455

[95] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (FAT* '20), Association for Computing Machinery, New York, NY, USA, 295–305. DOI:https://doi.org/10.1145/3351095.3372852