Designing for Responsible Trust in AI Systems: A Communication Perspective

Q. Vera Liao
Microsoft Research
Montreal, Canada
veraliao@microsoft.com

S. Shyam Sundar
Penn State University
State College, USA
sss12@psu.edu

ABSTRACT

Current literature and public discourse on "trust in AI" are often focused on the principles underlying trustworthy AI, with insufficient attention paid to how people develop trust. Given that AI systems differ in their level of trustworthiness, two open questions come to the fore: how should AI trustworthiness be responsibly communicated to ensure appropriate and equitable trust judgments by different users, and how can we protect users from deceptive attempts to earn their trust? We draw from communication theories and literature on trust in technologies to develop a conceptual model called MATCH, which describes how trustworthiness is communicated in AI systems through *trustworthiness cues* and how those cues are processed by people to make trust judgments. Besides AI-generated content, we highlight *transparency* and *interaction* as AI systems' affordances that present a wide range of trustworthiness cues to users. By bringing to light the variety of users' cognitive processes to make trust judgments and their potential limitations, we urge technology creators to make conscious decisions in choosing reliable trustworthiness cues for target users and, as an industry, to regulate this space and prevent malicious use. Towards these goals, we define the concepts of *warranted trustworthiness cues* and *expensive trustworthiness cues*, and propose a checklist of requirements to help technology creators identify appropriate cues to use. We present a hypothetical use case to illustrate how practitioners can use MATCH to design AI systems responsibly, and discuss future directions for research and industry efforts aimed at promoting responsible trust in AI.

KEYWORDS

Trust in AI, human-AI interaction, human-centered AI, AI design

ACM Reference Format:
Q. Vera Liao and S. Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3531146.3533182

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533182

1 INTRODUCTION

With the popularity of complex AI systems used to augment or automate tasks that can affect many people's lives and have a long-lasting impact, trust is often cited as a key requirement for people to adopt AI technologies. The current academic and public discourses are predominantly structured around the guiding principles towards trustworthy AI [62, 64], often as a way to operationalize principles for responsible and ethical AI [44], such as ensuring effectiveness, fairness, transparency, robustness, privacy, security, and serving human values. These principles are inherently technocentric, focusing on what constitutes the trustworthiness of AI, when in fact trust is a human judgment or attitude, which can be formally defined as a judgment of dependability in situations characterized by vulnerability [34, 65]. The same AI technology can be judged differently by different people, with some forming inaccurate trust judgments. It is ultimately this psychological reality that determines how people would use and interact with the AI, and whether one could be harmed by inappropriate trust and consequent behavioral outcomes such as over-reliance and misuse.

We argue that the AI field's fixation on trustworthiness results in blind spots in how people make trust judgments as well as how to *communicate* the trustworthiness of AI appropriately and responsibly. Trustworthiness of a technology is not inherently established but communicated through *trustworthiness cues*, which are embedded in interface features, documentation, and other modes of information, such as speech acts for conversational AI. With this communication perspective, our focus is on *AI systems* rather than standalone models, where technology creators—including system developers and designers—need to make conscious decisions in choosing and designing these trustworthiness cues. The space of AI trustworthiness cues is becoming increasingly rich as researchers and practitioners build AI systems in numerous domains. The ethical imperative of transparency, in particular, calls for diverse types of information to be provided about the model's capabilities, limitations, decision processes, provenance, and so on. For example, the surging field of explainable AI (XAI) has produced a vast collection of techniques to generate model explanations [1, 8, 20, 37] with one goal, among others, being engendering trust in users.

By framing the design of AI systems as conveying trustworthiness cues, we foreground two issues that are of particular importance for holding AI technologies and their creators accountable. One is that malicious manipulation of trustworthiness cues can lead to undeserved trust with far-reaching harmful consequences. It is imperative to decouple the underlying model trustworthiness and the communication of it as a foundation to begin considering how to regulate AI system design. The other issue is that even well-intentioned technology creators may produce ill-designed

trustworthiness cues that harm users due to a lack of attention to users' cognitive basis for trust judgments. The field's fixation on AI's trustworthiness can foster a false assumption that there are only "ideal users" who can perfectly assess it from available information. In reality, people have varied abilities and motivations to make accurate trust judgments. For example, abundant empirical evidence suggests that even technically sound AI explanations can result in harmful over-trust and over-reliance [3, 11, 28, 60, 70]. Some user groups are more vulnerable to these harms than others, such as AI novices [60], people working in cognitively constrained settings [52], and even those with certain personality traits [17]. We urge the AI field to develop a deeper understanding of how people process information to make trust judgments in order to develop reliable trustworthiness cues, as well as accountable mechanisms to generate them, all geared toward ensuring appropriate user trust and equitable user experiences with AI systems.

To facilitate such an understanding, we present a conceptual model, named MATCH, focusing on the communication of trustworthiness in AI systems and the processes by which users make trust judgments (Section 3), by drawing from communication and human factors literatures (Section 2). MATCH decouples the trustworthiness attributes of the underlying AI model(s) and trustworthiness cues presented to the users, via three types of *affordances* of AI systems: AI-generated content (e.g. predictions), transparency, and interaction. With this conceptualization, we highlight transparency as an affordance to enable trust judgments rather than warranting trust in itself, and bring forward the role of interaction design in shaping user trust. MATCH also highlights the wide variety of people's cognitive processes to make trust judgments: instead of always being processed analytically to form rational trust judgments, trustworthiness cues often invoke *heuristics*—mental short-cuts or rules-of-thumb—for people to make speedy, but sometimes flawed, judgments. Communication theories further inform the varied tendencies to engage in heuristic trust judgment among different user groups.

On the basis of MATCH, we describe "good" trustworthiness cues that technology creators should use (Section 3.4). We define the concept of *warranted trustworthiness cues* with a checklist of requirements to urge technology creators to focus on the psychological reality of their target users rather than technological qualities alone. We further suggest the use of *expensive trustworthiness cues* as an industry practice that, by imposing a level of expense on technology creators, can help collectively guard against malicious means of deceiving user trust.

To illustrate the use of MATCH, we present a case study of designing a hypothetical AI system (Section 4). While it is still a nascent area, we survey related works in human-computer and human-AI interaction to suggest a list of trustworthiness cues and trust heuristics. In Section 5, we reflect on the implications of MATCH and propose three areas of call to action to build responsible trust in AI: to regulate technology creators' use of trustworthiness cues, to empower users to make accurate trust judgments; and to look beyond model intrinsic attributes, and leverage social, organizational, and industrial mechanisms to enable reliable trustworthiness cues in AI systems.

2 BACKGROUND AND RELATED WORK

2.1 Trustworthy AI and trust in AI

Ensuring the trustworthiness of AI, i.e., what's required for people to trust AI [64], has been considered an operational point to implement ethical AI principles [62]. Building on the classic ABI (Ability, Benevolence, and Integrity) framework from the social sciences [7, 39], which prescribes trustworthy characteristics of a trustee as the three ABI dimensions, Toreini et al. [62] propose four categories of trustworthiness technologies for AI, namely Fairness, Explainability, Auditability and Safety (FEAS). The authors further demonstrate that these trustworthiness technologies operationalize core dimensions in many existing ethical and principled AI policy frameworks from both industry and governments. In a similar vein, Varshney [64] maps out the trustworthy qualities of AI as predictive accuracy, robustness (to data shift and poisoning), fairness, interpretability, system-level provenance and transparency, and intention for social good.

Another relevant thread of work explores organizational and regulatory ecosystems for ensuring trustworthy AI. Shneiderman [54] proposes a three-layer governance structure: reliable systems, safety culture, and trustworthiness certification by independent oversight. Knowles and Richards [30] contend that the public distrust of AI originates from the underdevelopment of a regulatory ecosystem that would guarantee AI's trustworthiness. Drawing from the literature on institutional trust [19], the authors develop a model for public trust in AI-as-an-institution and highlight the pivotal role of auditable AI documentation in promoting public trust by constructing signals of trustworthy AI, establishing norms about what constitutes legal or ethical non-compliance, and allowing the exercise of control.

Despite outlining the complexity of trust [55, 62], these works are detached from the cognitive mechanism of how people make trust judgments. By examining the consequences of a collaborative system for data scientists, Thornton et al. [61] demonstrate the nuances in implementing trustworthiness principles and highlight the gaps between these principles and actually promoting user trust. The latter requires attending to the designed aspects of the system that "provide access to evidence of (dis)trustworthiness specific to a user, a technology and their context," or what they termed "trust affordances." More recently, inspired by the literature on interpersonal trust [42], Jacovi et al. [25] formalized human trust in AI as "contractual trust," such that trust between a user and an AI model is anticipating that some contract will hold. Under this formalization, AI principles such as fairness, accountability, robustness, intention for social good, and privacy, can be seen as contracts, each of which places different criteria for people to establish trust, working together to form overall trust. This formalization brings forward the concept of *warranted trust* (there exists a causal relationship between users' trust and the model's trustworthiness for a given contract). Accordingly, the authors suggest that the existence and level of warranted trust can be evaluated by manipulationist causality, i.e. whether and how much users vary their trust based on manipulated changes in the trustworthiness attribute of the model.

2.2 Trust in technologies: Lessons from communication and human factors literature

Aligning with Thornton et al. [61] and Jacovi et al. [25], we aim to disentangle the relation between human trust and model trustworthiness. We further delve into cognitive aspects of trust judgments, for which we draw inspiration from the literature on trust in automation and web technologies. The two areas share emphases on considering the basis of trust, users' cognitive process to make trust judgments, and the impact of contextual factors, but they have different foci. Research on automation often studies human trust in association with the outcome of machine reliance, and dedicates much effort on elucidating the basis of trustworthiness, which we can draw parallels with the current emphasis on trustworthy principles of AI. Web trust literature deals with how people judge information quality and dependability on web sites [63], with the bulk of research conducted under the term "web credibility" [51]. For simplicity, we use the term "web trust" throughout the paper.

Trust in automation. Our perspective is most directly informed by the seminal paper by Lee and See [34]. By synthesizing related literature, the paper proposes a conceptual model describing the process of trust formation in automation. Below we highlight a few key points of this model.

Trust is determined by people's perception of information about the trustworthiness attributes of the system and existing beliefs. There has been substantial work on conceptualizing trustworthiness attributes of automaton, which is often built on the ABI model [39] mentioned above. Lee and Moray [33] adapt the ABI model to three dimensions that more suitably characterize automated systems: performance (ability)—*what* automation does; process (integrity)—*how* automation operates; and purpose (benevolence)—*w*hy automation was developed. Lee and See [34] show that many necessary characteristics of trustees discussed in the trust literature can be mapped onto these dimensions.

People's perception of trustworthiness attributes is mediated by the display of automation information, which is assimilated by multiple cognitive processes: analytic, analogical (linking to known categories associated with trustworthiness), and affective (emotional response) processing. This perspective of multi-channel processing is key to understanding how people form trust judgments with rich displays that invoke trust-related heuristics and emotional responses.

Trust guides the behavior of reliance, but in a non-linear way, subject to the influence of individual, organizational and cultural contexts. Several important factors influence the behavior of reliance, such as workload, intention for exploratory behavior, efforts to engage, perceived risk, self-confidence, time constraints, and system configuration. The dynamic interplay between automation, trust, and reliance can generate substantial non-linear processes: e.g., information display shapes the formation of trust, but existing level of trust also affects the selection and interpretation of information. Contextual factors also impact the development of trust directly. Individual differences vary the propensity to trust as well as channels of cognitive processing. Organizational and cultural contexts (e.g., other people's comments) play significant roles in trust development, highlighting the often neglected ecological aspects of trust.

Trust and appropriateness of trust are multi-faceted. The scope of trust information display depends on the locus of trust, whether it is about trust in the system, function, or sub-functions. In the web trust literature, the locus of trust is differentiated for web, web sites, and messages [51]. However, users are not always able to disentangle information and beliefs about them. Technology design should strive for appropriate user trust, which also has multiple facets: calibration (trust matches system trustworthiness), resolution (changes in system trustworthiness match changes in user trust), and specificity (able to differentiate different functional components of system trustworthiness).

We also emphasize the mediating role of information display on trust judgments, and that appropriate trust relies on effective communication of system trustworthiness. To further elucidate the communication aspect, we now turn to the literature on trust in web technologies, which pays great attention to how interface cues shape trust judgments.

Effects of information cues on trust in web technologies: communication perspectives and heuristic approaches. Trust or credibility of information has long been studied in the fields of HCI, communication, and information science [51]. The early 2000s saw a rise in research on how people make trust judgments of web sites, including frameworks on what elements of web technologies influence people's trust [13, 14, 21, 40, 57, 67]. Practical means, including design guidelines [15, 16] and tooling [53, 68], have also come out of this line of work both to facilitate technology creators' design choices and empower web users to make better judgments.

Much of this literature is based on communication theories of **dual-processes models** for attitude formation, including Petty and Cacioppo [49]'s elaboration likelihood model (ELM) and Chaiken [5]'s heuristic-systematic model (HSM) (also related to Kahneman [27]'s System 1 and System 2 thinking). These theories postulate that web users engage in two cognitive processes to assess a website: one is "systematic" processing by paying attention to information content and performing a rigorous evaluation. The other is "heuristic" processing by attending to *cues* about the information quality provided by the interface. The cues then trigger *heuristics* that allow quick and cognitively easy judgments. A website can be seen as having two parts: its information content, and a repository of cues extrinsic to the content but contributing to trust judgments (e.g., article source, URL links, "likes"), also referred to as "content cues" and "contextual cues" [67] respectively. Modern web technologies provide an abundance of contextual cues. Furthermore, dual-process theories predict that when users lack an ideal level of *motivation* and *ability* (broadly defined) to engage in systematic processing, they are likely to resort to heuristic judgments, often based on contextual cues.

This cue-heuristic perspective raises an important question: *what cues are made available and what heuristics can be triggered by a given technology?* Web researchers have answered the question empirically [15, 56]. By surveying 2500 participants, Fogg [14] summarizes 18 types of cues people frequently notice on a website to base their trust judgments on, such as information structure, name recognition, and advertising, with the most frequently mentioned cue being the "design look." By conducting focus groups with 109 participants, Metzger et al. [41] show that people routinely invoke cognitive heuristics to assess the trustworthiness of web sites, such

as heuristics of reputation (e.g., website name recognition), endorsement (recommended by others or having good ratings), consistency (cross-validation in multiple websites), expectancy violation, and persuasive intent (e.g., advertising).

Researchers also developed theoretical frameworks to account for the types of cues in web technologies [21, 57, 67]. The MAIN model developed by Sundar [57] has had a long-lasting impact. Its central thesis is that a given technology has certain "affordances" capable of cueing trust related cognitive heuristics (**affordance-cue-heuristic** approach). Affordance is an important concept in psychology and HCI (human-computer interaction) literature, defined as displayed properties of a system suggesting ways in which it could be acted upon or used [18, 45]. The MAIN model earned its name by specifying four types of affordances that provide trust-related cues : Modality (e.g., visual modality cues realism heuristic), Agency (e.g., endorsement of "other users" cues bandwagon heuristic), Interactivity (e.g., the ability to customize cues control heuristic), and Navigability (e.g., the availability of many hyperlinks cues elaboration heuristic). Sundar summarizes a total of 29 heuristics that can be cued by these affordances [57]. These cues can risk invoking unwarranted trust, as not all of them are directly linked to the content's trustworthiness, especially in malicious websites. Yet these cues play a major role in shaping trust judgments, given the deluge of online information and the impossibility of close scrutiny given cognitive limitations. This is true for all users, but especially for users who lack the ability or motivation to engage in a careful reading of the contents [57].

Equipped with these theoretical bases, we develop a conceptual model to describe trust judgments of AI systems. Our model synthesizes the above perspectives on the basis of trustworthiness, mediating role of information cues on trust judgments, the dual-process models, and the affordance-cue-heuristic approach.

3 MATCH: A CONCEPTUAL MODEL OF USER TRUST IN AI

Our conceptual model aims to describe how AI trustworthiness is communicated in AI systems and processed by users to make trust judgments (Figure 1). The process is broken down into three parts: the underlying model (**M**) attributes, system affordances (**A**) to communicate AI trustworthiness (**T**) cues (**C**), and users' cognitive processing of these cues by invoking trust-related heuristics (**H**). We refer to this model as MATCH and discuss each part below.

As pointed out in Figure 1, our scope is concerned with *trust in the AI model(s)* that underlies a system, which we isolate from other loci of trust, such as trust in the institution (e.g., the company or brand producing the system) [51] and trust in AI-as-a-technology [30]. We focus on trust as an attitude rather than its effects on user behaviors such as reliance, and acknowledge that there are ecological factors beyond our focus on the internal cognitive processes that shape trust judgments and reliance, depending on individual, environmental, organizational, and cultural contexts.

3.1 Model trustworthiness attributes: what makes the basis of trustworthiness?

As reviewed, many define the trustworthiness of technologies [31, 33, 34, 62] based on the classic ABI model. Lee and Moray [33] adapted ABI to the context of automation using the dimensions of performance, process, and purpose. We adopt these dimensions and define the basis of AI trustworthiness as ability, intention benevolence, and process integrity. Note that the three core components of MATCH (trustworthiness attributes, trust affordances, and cognitive processes of trust judgments) should be agnostic to how the attributes are operationalized. We welcome future work to expand these dimensions or explore alternative operationalizations. In Figure 2, we suggest a non-exhaustive list of example attributes under each dimension, which will be discussed with an illustrative use case in Section 4.

For the sake of simplicity, our discussions in this paper assume that there is a single underlying model that serves as the basis of user trust in the AI system. For complex systems with multiple models or technological components, these dimensions would hold but one may need to define specific attributes differently (e.g., modular and joint abilities) and consider trust *specificity* for different components as one facet of appropriate trust [34].

Ability refers to the capabilities of the underlying AI model with regard to its output or the function it provides to the user (e.g., making predictions, generating answers). They cover *what* the AI can do. For example, *overall performance* is a key attribute of AI ability. Considering other trustworthy AI principles [62, 64], *performance fairness* (e.g., absence of performance differences between different protected demographic groups) and *performance robustness* (e.g., against data shift and poisoning [64]) are also ability attributes. We also consider *improvability* as an ability attribute. For example, an active learning model is set up to be improvable with user input. Note that this conceptualization distinguishes between objective performance as an underlying attribute of the model and performance metrics as trustworthiness cues to approximate (e.g., calculated using test data) and communicate the attribute.

Intention benevolence refers to the degree of benevolence behind the creation of the technology– *for what* is the AI developed? Besides *intended use* (e.g., social good, serving user values), we also consider *intended compliance* (e.g., privacy-preserving, security conscious) as an attribute of intention benevolence.

Process integrity is the degree to which the operational or decision process of the model is appropriate to achieve the users' goal, describing *how* the AI works. The standard of integrity should be context- and user-dependent, such as the absence of flawed logic, optimizing for the right goal, and aligning with known processes in the domain. While the level of process integrity could make a difference in the AI's ability attributes, the former creates a different basis of trust, one that is focused on the system's dispositional integrity rather than its outcomes.

These three dimensions of attributes determine the level of trust that users should have in an ideal world. However, in reality, these attributes are communicated through trustworthiness cues, and then the cues are judged through a plurality of cognitive processes, both of which introduce noise, as we discuss in the following.

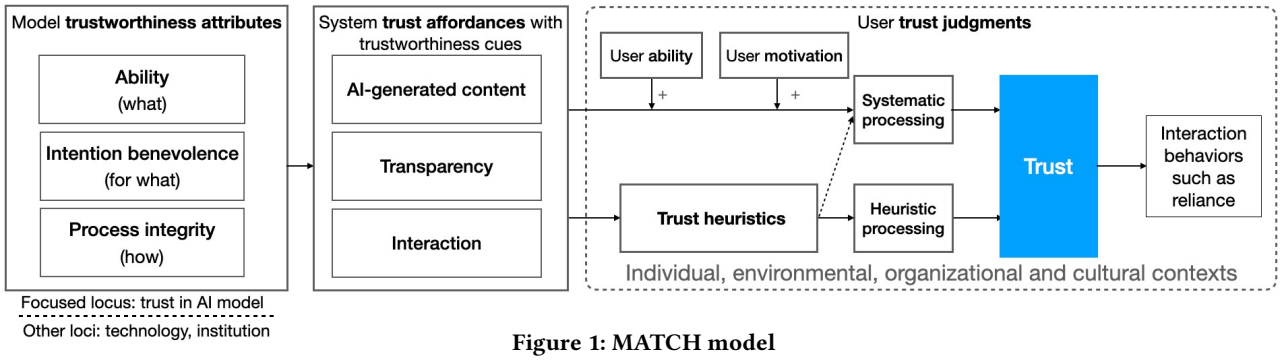

Figure 1: MATCH model

3.2 Affordances for trustworthiness cues: how is trustworthiness communicated in AI systems?

A *trustworthiness cue* is any information within a system that can cue, or contribute to, users' trust judgments. For individuals, trust is often conceived as a judgment of dependability, and not necessarily driven solely by cues that explicitly reflect the three bases of trustworthiness described above. An AI system can thus place its users in a rich environment of trustworthiness cues. According to the affordance-cue-heuristic approach [57], one may identify trustworthiness cues by conceptualizing the *affordances* of a given type of technology that engender such cues.

The question of "what are the trust affordances [61] of AI systems" can be challenging to answer since AI is far from a monolithic set of technologies. We base our proposal on currently popular AI systems (e.g., decision support, task assistance, recommender systems) and common system features in production and literature. A recent survey paper [32] maps out AI system elements that have been empirically studied for AI decision support in HCI and AI literature, including different types of prediction output, information about the prediction (e.g., local explanations, uncertainty information), information about the model (e.g., performance metrics, documentation, model-wide explanations, training data), and user control features (e.g., customization, feedback to improve the model). Accordingly, we suggest three types of common affordances of AI systems: AI-generated content, transparency, and interaction. We discuss these affordances below, and suggest a list of example trustworthiness cues provided by these affordances in Figure 2.

AI generated content refers to displays of the model output or the functional support provided by the AI system. Depending on the type of model, displays can take the form of a predicted class label, a score, a list of suggestions, generated texts or images, etc. These displays can serve as direct trustworthiness cues for users to assess the ability attributes of the AI model. In some cases the design, e.g., under what circumstances AI assistance is provided or not, can also cue users' judgment of the intention benevolence of the model.

Transparency affordance refers to displays allowing a better understanding of the model, broadly defined, including its behaviors, processes, development, and so on. We single out transparency as a unique affordance of AI systems given the increasing industry emphasis on providing transparency, exemplified by the prevalence of normative metrics (including performance, fairness, and robustness metrics), XAI features, and model documentation [2, 22, 43] (commonly including provenance information [30, 61] about how and for what it was developed). Recent literature also discusses governance structures to ensure trustworthy AI [50, 54], such as internal reviews, testing, independent and government oversight, and so on. Communicating the process and outcomes of such governance structures should also be considered a form of transparency. Transparency allows cueing for all three dimensions of trustworthiness attributes—ability (e.g., through metrics), intention benevolence (e.g., communicating intended use and compliance in the documentation), and process integrity (XAI features). This conceptualization highlights the role of transparency as an affordance for users to base their trust judgments on, rather than inherently warranting trustworthiness. This is related to Jacovi et al. [25]'s formalization of the goal of XAI as facilitating appropriate trust by increasing the trust of users in a trustworthy AI system and distrust in a non-trustworthy one. This goal can only be attained if trustworthiness cues in the transparency affordance are both truthfully communicated and appropriately assessed by the user.

Interaction affordance refers to displays that suggest how users can interact with the system, beyond the content of the model output, for which we consider both perceptual affordances (e.g., medium and design look) and action affordances (e.g., customization of the system, socialization possibilities with other people). The roles of interaction and interaction design are often overlooked in the current literature on trust in AI. We draw parallels with the web trust literature showing that people base their trust judgments not only on "content cues" but also on many "contextual cues" [49] on a web site, such as the design look, source reputation, or social information [14, 15, 41]. Some interaction affordance is directly enabled by the model ability, such as customization in guiding the model's behavior, and can directly cue the trustworthiness attributes of ability (improvability) and intended use (e.g., serving user preferences). Other interaction affordances may be extrinsic, even irrelevant, to the model (e.g., the choice of medium, such as using a visualization), but can still cue people's trust judgments. By bringing to light interaction as an affordance providing rich trustworthiness cues, we urge future research to better understand whether and how different interaction features of AI systems, even disassociated from the underlying model, can impact user trust.

3.3 Dual cognitive processes: how are trustworthiness cues processed by people?

MATCH conceptualizes this process based on dual-process models of attitude formation [5, 49]. The basic idea is that people process information to form a judgment or attitude through two routes: 1) **systematic processing** by rigorously assessing the information to make a rational judgment, and 2) **heuristic processing** by following known heuristics or rules-of-thumb to make a cognitively easier judgment. However, when the heuristics are applied inappropriately, the judgment is prone to errors. MATCH further highlights the roles of trust heuristics and individual differences.

Trust heuristics are any rules of thumb applied by a user to associate a given cue with a judgment of trustworthiness. There are many ways for individuals to acquire trust heuristics. Some are common cognitive heuristics applied to the context of AI. For example, online users tend to apply the *authority heuristic* by following the opinion of an authority on the subject matter [41]. This heuristic can be invoked by an interface cue showing certification from a regulatory body that audited the AI. Others are technology-specific heuristics learned from past experiences. For example, some groups of users may have a prominent *machine heuristic*, believing machines are more reliable than humans [58]. The phenomenon of XAI features leading to over-trust [3, 11, 28, 60, 70] can be attributed to an "*explainability heuristic*" [10, 36] that superficially associates being explainable with being capable, without deliberating on the actual content of the explanation. Heuristics can also be intentionally cultivated by technology creators. One example is to provide instruction and supporting evidence that a number above a certain threshold of a normative metric could be considered acceptable. The existence of heuristics varies between individuals. However, it is possible to enlist common heuristics based on psychology and communication theories [41, 57], or by empirically exploring the heuristics that are frequently invoked during target user groups' interaction with the AI, by using, for example, think-aloud methods to examine cognitive processes [23, 26]. It is worth pointing out that while heuristic processing necessarily involves heuristics, heuristics can also be used to aid systematic processing when they are applied in a conscious and deliberative fashion [5, 57]. In Figure 2, we provide example heuristics based on prior literature, to be discussed in Section 4.

Individual differences in systematic vs. heuristic processing. People have different tendencies to engage in systematic versus heuristic processing. Hence, the introduction of a trustworthiness cue can risk creating inequality in trust and user experience. For example, recent studies have repeatedly found that XAI features bring less benefit, or even harms (leading to over-trust), to certain user groups such as AI novices or users working in cognitively constrained settings [17, 52, 60]. Research based on dual-process models [49] has long established that when individuals lack either the *motivation* or *ability* to perform systematic processing and rationally assess trustworthiness, they are likely to resort to heuristics. Note that motivation and ability are umbrella terms that can encompass many user and contextual characteristics, which make these theoretical models powerful for understanding and predicting individual differences. For example, a user may lack ability due to a lack of AI knowledge, domain knowledge, or cognitive capacity; they may lack motivation due to perceived cost versus gain, personality traits, or competing motives [46, 48, 49]. By highlighting these individual differences, we encourage technology creators to carefully examine and mitigate the potential inequalities of experience among users who may not have an ideal profile of ability or motivation.

3.4 What are "good" trustworthiness cues?

Based on this conceptual model, we attempt to address an important question: *what are "good" trustworthiness cues that should be used by technology creators?* It is helpful to break down the consideration of "goodness" into two scenarios: 1) for a well-intended technology creator, a good trustworthiness cue is one that results in well-calibrated trust judgments by target users with regard to the true trustworthiness of the AI; 2) for the industry and society as a whole, a good trustworthiness cue is one that both has good calibration and is likely used truthfully to communicate the underlying trustworthiness, or in other words, not subject to malicious and deceptive use.

Warranted trustworthiness cue. To facilitate efforts around using and regulating good trustworthiness cues, we first introduce this concept. We consider a trustworthiness cue to be warranted if:

(1) It is truthfully used by the technology creator, without deceptive intentions (**truthfulness condition**).
(2) It is relevant to or reflective of the underlying model trustworthiness attributes, including ability, intention benevolence, and process integrity (**relevance condition**).
(3) It leads to well-calibrated trust judgment by the target users with regard to the trustworthiness attribute(s) it reflects (**calibration condition**).

This concept is relevant, but not identical to Jacovi et al. [25]'s formalization of warranted trust (if the incurred trust corresponds to the trustworthiness of the model), as our focus is on what kind of trustworthiness cues are likely to result in warranted trust in users. We now discuss the requirements for achieving each condition.

Relevance condition. This condition underscores the need to consider if a trustworthiness cue is indeed reflective of the underlying trustworthiness attributes of the AI. Technology creators should pay attention to *prominent irrelevant trustworthiness cues*—cues that shape users' trust judgments, but are not reflective of the trustworthiness attributes of the model, such as the surface design look or a link to an external web site. Often, some irrelevant cues are unavoidable because they support other user goals, but they can impact user trust in unintended ways. Such unintended effects can be mitigated by making these cues less prominent during users' trust-development stage, providing interventions to disrupt invocation of trust heuristics (e.g., a reminder that the design is inherited from a template), or guiding user attention to other relevant trustworthiness cues.

While we encourage technology creators to incorporate comprehensive trustworthiness cues that directly describe the model trustworthiness attributes of ability, intention, and process, the relevance condition should embrace any cues that provide supporting evidence for these attributes. We may differentiate between *model-intrinsic* and *model-extrinsic trust-relevant cues*. Intrinsic cues are generated directly from the model or its development process,

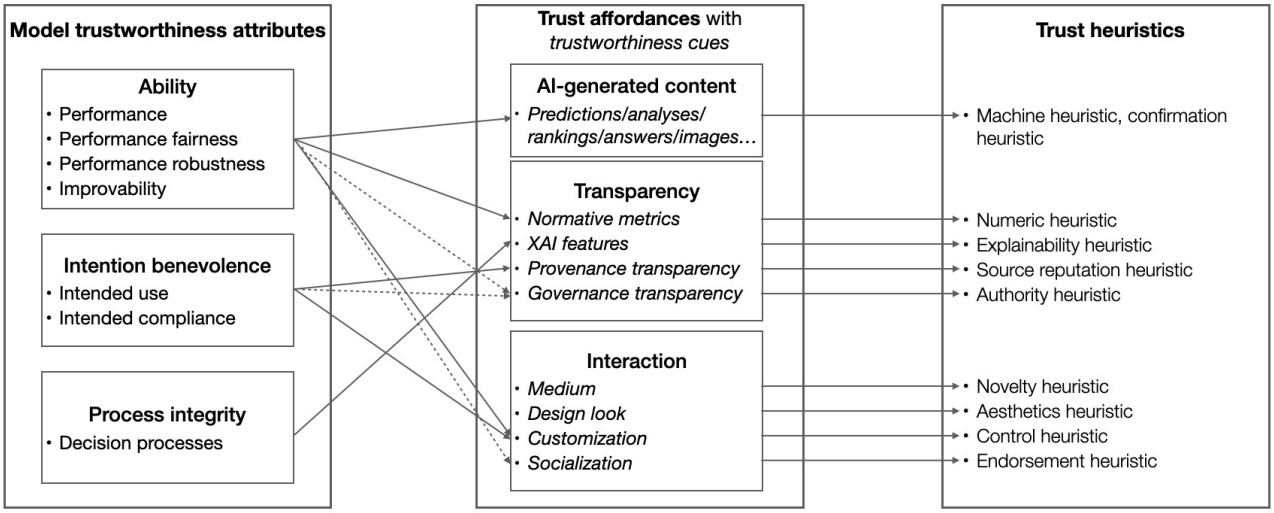

Figure 2: Example lists of trustworthiness attributes, trustworthiness cues and heuristics, used in the use case in Section 4

such as its output, performance metrics, and explanations. Extrinsic cues are generated from social, organizational, and industrial processes outside model development but can provide supporting evidence for its trustworthiness attributes. Examples may include other users' reviews, audit trails, and evidence from external or regulatory oversight.

Calibration condition. Calibration requires a match between a person's trust judgment based on a given trustworthiness cue and the true trustworthiness of the underlying model attribute(s) reflected by the cue. Calibration can be assessed with a **formal analysis**—measuring the change in people's trust judgment, by manipulating the quality of the corresponding model trustworthiness attribute(s) reflected by the trustworthiness cue [25]. A good trustworthiness cue should result in consistent changes between the two. Calibration should be considered probabilistic instead of a dichotomy—e.g., showing an accuracy metric may provide better calibration than showing predictions alone, but neither may be perfectly calibrated to the true changes in model ability.

However, it is not always feasible to perform costly formal analysis to quantify the calibration, which may also suffer from generalizability issues given individual and contextual differences. Based on MATCH, we suggest the following heuristics to help technology creators identify whether a given trustworthiness cue has a high or low probability of calibration. We postulate that a trustworthiness cue is more likely to satisfy the calibration condition if:

- *The target user group has the ability and motivation to perform systematic processing* (**systematic condition**).

Or

- *It does not invoke unfounded trust heuristics* (**no unfounded heuristic condition**).

Whether a heuristic is "unfounded"—with little evidence to support or low probability to hold—depends on the context and the user. For example, a recent study [38] shows that people follow the cognitive heuristic of *confirmation bias* when viewing AI predictions, whereby the agreement of AI predictions with their own judgment is seen as indicating high AI ability. This heuristic may be unfounded (low probability to hold) if users are novices to the decision task, but could be acceptable for domain-expert users. Meanwhile, some heuristics are generally unfounded and we should avoid invoking them—for example, recent studies of user interaction with XAI suggest the existence of unfounded *explainability heuristic* [36] (associating being explainable with being capable) and *numeric heuristic* (associating numerical explanations with algorithmic intelligence). According to the prominence-interpretation theory on web trust [14], the existence of unfounded heuristics should best be assessed jointly by the likelihood of noticing a feature and invoking a trust heuristic (prominence), and the likelihood of the heuristic being unfounded (interpretation).

Importantly, this condition acknowledges the benefit of *founded heuristics*. Heuristics are an indispensable part of people's cognitive process and it is unrealistic to expect all users to have the ability and motivation to perform systematic processing at all times. Individuals take advantage of heuristics because they are founded in past experiences or some conditions, and can generally help them make better judgments. Technology creators should strive to leverage common cognitive heuristics, reverse-engineer reliable mechanisms to make users' naturally invoked heuristics better founded, and cultivate founded heuristics by providing training, guidance, or reinforcement mechanisms. For example, it is known that people have a common *anchoring heuristic/bias* that hinges judgment on their first encounter with an object of trust. A design choice that makes this heuristic better founded in the context of AI systems is to present users with performance transparency information during the system on-boarding stage. This kind of effort is key to engendering equitable trust by enabling users with less-than-optimal motivation and ability to better assess AI systems.

Truthfulness condition. We consider this condition last as it is concerned with the intent of technology creators. Rather than asking what is required for this condition, the key question here is how to prevent deceptive use of untruthful trustworthiness cues. We bring in one perspective by drawing on "costly signaling theories" from evolutionary psychology [69]. Signaling theories are a body

of theoretical work [4, 6] concerned with how individuals (humans and animals) select signals (traits, actions, etc.) to present during communication to convey some desirable quality for achieving a social goal. Since individuals have motivations to deceive, collectively evolution would favor reliable signals that are "costly"—costing the signaller something that could not be afforded by those with less of a given quality.

With a similar motivation to collectively guard against deception, we argue that the industry as a whole should prioritize using **expensive trustworthiness cues** that would impose a level of expense on technology creators. We consider "expense" as any investment that a creator must make to present a trustworthiness cue to a believable extent to the users, including but not limited to development, time, and infrastructure expenses. For example, showing an accuracy metric is less expensive than a user-friendly XAI visualization; establishing positive audit trails and endorsement from others are generally costly in terms of time and effort. More expensive trustworthiness cues also include comprehensive documentation, certification from established review boards, and customization features. However, in practice, individual technology creators may need to weigh the expenses and limit their choices to cues within their affordable range. Like many responsible AI practices, costly implementation runs the risk of marginalizing smaller business entities and creating inequalities in the industry. While we suggest leveraging expense on technology creators to safeguard the truthfulness condition, a much more nuanced view on its relations with resources, gains, and other motivators and constraints needs to be developed to inform policy and industry practices.

4 USE CASE: USING MATCH TO DESIGN FOR APPROPRIATE TRUST IN AN AI SYMPTOM CHECKER

MATCH can help technology creators prospectively interrogate what trustworthiness cues should be presented in a system, or retrospectively understand the causes of users' inappropriate trust (e.g., whether due to salience of trust-irrelevant cues or miscalibration). We demonstrate the former with a hypothetical use case of designing an AI system. We also leverage this use case to present examples of model trustworthiness attributes, trustworthiness cues, and trust heuristics to help ground our conceptual model, as summarized in Figure 2. We start by describing the use case:

> *HealthChecker is an AI app that suggests diagnosis for common diseases based on a list of symptoms that a user provides. Its suggestions also take into consideration the patient's personal information, such as demographic and socioeconomic background, and health sensor data if the patient consents to their collection. HealthChecker also sends its suggestions to the patient's primary doctor for verification and suggestions.*
>
> *There are two groups of primary users. One is patients, to represent which the creators consider the persona "Eric"—an average mobile app user who has needs for diagnosis of common diseases a few times a year, is neither an AI expert nor health expert, but is keen on trying out new technologies. The other is the patients' primary doctors, for which the creators consider the persona "Jessie"—an average primary care doctor who is a medical expert but only moderately familiar, and usually cautious, with AI technologies.*

Step 1: Which model attributes determine its trustworthiness and should be communicated? The creators start with this question. Based on MATCH, they consider the model's ability attributes including performance, fairness, robustness and improvability, intention benevolence that governed the model development, including intended use and compliance (e.g., privacy-preserving is especially important here), and the model's process integrity to make diagnoses.

Step 2: What kinds of cues should be used to communicate trustworthiness? This analysis happens in parallel with other design decisions for the system, such as the kind of interface needed to support efficient use and a user-friendly experience. MATCH guides them to consider what trustworthiness cues should be designed into the affordances of AI-generated content, transparency, and interaction. For AI-generated content, the diagnosis suggestion itself can cue judgments of the AI's ability. They start with a simple design of presenting only the top suggestion.

For transparency, they consider multiple features that communicate the three dimensions of trustworthiness attributes: normative metrics for accuracy, fairness, and robustness to show ability; explanations of diagnoses to show process integrity; and documentation that highlights the intended use and compliance considerations of the AI. The solid arrows in Figure 2 show the mapping analysis between trustworthiness cues and trustworthiness attributes. They also consider what kind of model-extrinsic cue can provide supporting evidence for the model's trustworthiness attributes. An internal review board has been introduced in the company to oversee the development of AI technologies, reviewing regulatory and ethical compliance such as fairness and privacy. The creators choose to include information about this governance structure and a certification from the board in the documentation. The dashed arrow in Figure 2 shows the mapping between model-extrinsic cues and the trustworthiness attributes they support.

For interaction, the creators decide to invest in a customization function as an important trust-building feature, by allowing users to choose if they want to provide certain personal information. This feature lets users experience the improvability of the AI and the compliance intention to preserve privacy. The creators also need to examine other interaction features. The app has a built-in socialization feature that allows patients to see the doctors' ratings and feedback for diagnoses made by the app. The creators realize that patients will likely use such information to assess the AI's ability, as a type of model-extrinsic trustworthiness cue. By observing and talking to some users like Eric, the creators realize that the sleek design of the app and the use of a chatbot to gather information about their symptoms contributed significantly to their trust in the system. The design look and medium, however, cannot be mapped to the model trustworthiness attributes and should be considered irrelevant trustworthiness cues.

Step 3: Are these cues warranted trustworthiness cues for both user groups? If not, why? This question can help foresee inappropriate and inequality of trust, and identify causes for improvement. It should be asked iteratively as the design progresses

or whenever a new system feature is introduced. The creators leverage the conditions for warranted trustworthiness cues discussed in Section 3.4 to analyze the cues identified above. We assume that the truthfulness condition is satisfied for the creators of HealthChecker. For the relevance condition, the analysis above identifies that the medium (a chatbot interface) and the design look do not satisfy the condition, and can potentially trigger novelty and coolness heuristics [57] that lead to positive trust judgments. If over-reliance occurs, interventions should be introduced to either tone down these cues for new users or mitigate the prominence of triggered heuristics.

Next, they assess the calibration condition for the remaining trustworthiness cues, keeping in mind the two personas, Eric (patient) and Jessie (doctor). They consider the "systematic condition" *or* "no unfounded heuristic condition" to rate the *calibration likelihood* of each trustworthiness cue. The questions they ask are:

(1) Does the user group have the ability or motivation to perform systematic processing?
(2) If not, what kind of trust heuristic is likely to be invoked?
(3) How likely is this heuristic unfounded and how prominent is the unfounded heuristic?

We recommend answering these questions empirically with target users, e.g., recruiting participants with Eric and Jessica's profiles and conducting think-aloud studies as they interact with the system and make trust judgments. To complete this use case, we survey communication and HCI literature to enlist heuristics that have been identified for relevant trustworthiness cues, as shown in the last box of Figure 2. We enumerate the analysis for each trustworthiness cue below. The results are summarized in Figure 3, where the y-axis represents calibration likelihood.

Diagnosis suggestions: Jessie is able to make her own diagnosis and reason about the recommendation quality made by the AI. This cue satisfies the systematic condition and has a *high calibration likelihood for Jessie*. Eric lacks the ability to perform a systematic assessment. According to the literature, *machine heuristic* can be prominent for this group of users with a positive attitude towards AI [58], which can lead to over-trust. They may also resort to a positive *confirmation heuristic* if the AI's suggestions align with their own speculation [38], which is likely unfounded. Therefore, this cue has a *low calibration likelihood for Eric* and a high risk of leading to over-trust. The creators can improve the design to alleviate the unfounded heuristics, such as presenting uncertainty information and multiple candidate diagnoses.

Normative metrics: Jessie and Eric are not highly proficient with AI metrics, so it is unclear whether these cues can satisfy the systematic condition. Literature suggests that there exists a *numeric heuristic* whereby some people react positively to mathematical information about algorithms [9]. However, there is no reason to believe this heuristic is prominent for Jessie and Eric. The creators consider this cue to have *medium calibration likelihood for both Eric and Jessie*. To improve the calibration, they can provide evidence suggesting the acceptable range of these metrics.

Explanations: HealthChecker provides a feature-importance explanation to show how a diagnosis is made based on the most prominent symptoms. Jessie is able to understand and reason about these explanations analytically to assess the AI's process integrity. This cue has a *high calibration likelihood for Jessie*. Recent literature warned against presenting complex explanations to people who lack the ability to understand or assess them but commonly invoke an unfounded *explainability heuristic* that associates being explainable directly with superior capability [36]. Hence, explanations may have a *low calibration likelihood for Eric* and can lead to over-trust. The creators should only present explanation designs that are proven to be accessible for users like Eric.

Provenance and governance transparency in documentation: The creators invested in producing easy-to-read documentation that provides information about model provenance and governance structure. It is also expected that users like Jessie are often motivated to read the documentation for healthcare technologies. So these cues should satisfy the systematic condition and have a *high calibration likelihood for Jessie*. To cater to users unmotivated to spend time reading the documentation, it provides an overview that highlights the source of data used to train the model, the AI principles that the company follows, and certifications from an internal review board. These cues can invoke *source reputation heuristic* and *authority heuristic*, which are well founded in this context (likely to hold for improving trustworthiness). Hence they satisfy the no unfounded heuristic condition and have *high calibration likelihood for Eric*.

Customization: The customization feature mainly serves Eric to experience the improvability and compliance attributes of the AI. It is reasonable to expect that while using this feature, Eric would invest time to examine the effects although the improvement may not be easy to assess immediately. So the feature has a *medium-high calibration likelihood for Eric*. For Jessie, knowing that the system has a customization function may invoke the *control heuristic* that associates giving user control with a positive intention of the technology creators [57], which is reasonably founded, but not necessarily a prominent one. So the cue may have a *medium calibration likelihood for Jessie*.

Socialization: This feature allows Eric to view Jessie's feedback. Positive ratings from Jessie are likely to invoke an *endorsement heuristic* that leads to positive trust judgment. This heuristic is founded in this context and hence this feature provides cues that have a *high calibration likelihood for Eric*. It may not be applicable for Jessie's trust judgment.

Step 4: Which "expensive" trustworthiness cues should be prioritized? The creators ask this question to identify trustworthiness cues that would give their product an honest advantage to build user trust and also contribute to good industry practices. They solicit ratings from the team members on the expenses in developing and, if applicable, the processes and infrastructure to obtain, each trustworthiness cue. The X-axis of Figure 3 represents the results and ranks the more expensive cues to the right. Static presentations of performance metrics and diagnosis suggestions are low-expense because they require only plugging in outputs from the model development pipeline. The explanation visualization requires more effort for UI development. The carefully crafted documentation takes time to produce, and the governance transparency part requires company investment in the infrastructure. The socialization and customization features are expensive for the team and indeed provide competitive advantages over similar products.

Summary and guidance for using MATCH. As illustrated above, practitioners can use MATCH to design AI systems that

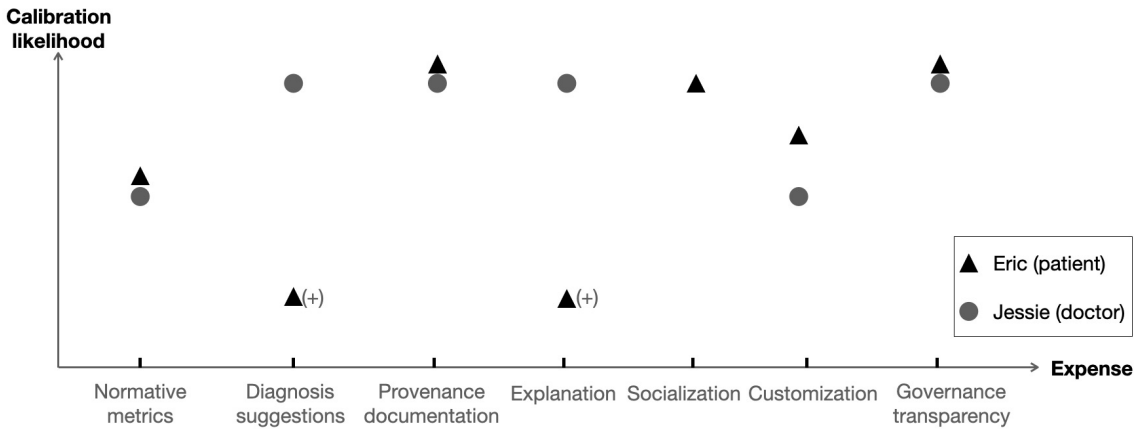

Figure 3: Calibration-expense analysis performed on the trustworthiness cues in HealthChecker. + indicates a risk of leading to over-trust.

responsibly communicate their true trustworthiness by following the four-step analysis, after prototypical user groups are identified and properly understood. This calibration-expense analysis (Figure 3) is best conducted during the planning stage to help the team identify trustworthiness cues to invest in, by focusing on those that are expensive within their affordable range, and can provide a good calibration for different user groups. It should also be done iteratively as the design and knowledge about target users progress. For example, the creator may attempt to mitigate users' machine heuristic by showing uncertainty information, but an empirical study could reveal that users like Eric have difficulties reasoning about quantitative uncertainty but invoke heuristics that lead to biased interpretation [24, 47]. In practice, it could be challenging to identify trustworthiness cues and trust heuristics for different user groups exhaustively. We view the MATCH model as a starting point to engage in careful considerations of the psychological reality of target users, and pinpoint detailed responsibilities for technology creators to ensure *appropriate and equitable* user trust. We discuss some future directions to advance these practices below.

5 DISCUSSION: TOWARDS RESPONSIBLE TRUST IN AI

We bring a communication perspective to the discourse on trust in AI. This conceptual work is intended to introduce and synthesize relevant theories on trust in technologies, elucidate the cognitive mechanisms of trust, and call out the requirements for using reliable trustworthiness cues. We invite future research to empirically investigate the topic and develop practical means for building responsible trust in AI, in the following directions.

Understanding and regulating the space of trustworthiness cues. Based on MATCH, technology creators' responsible use of trustworthiness cues has two essential sets of requirements: to truthfully and comprehensively communicate the model trustworthiness attributes, and to use cues based on which the target users are likely to make well-calibrated trust judgments. There are several challenges and complexities for future research to investigate. First, the mapping between cues and trustworthiness attributes is not always one-to-one, meaning that a system feature can cue multiple bases of trust [31]. It is important to recognize that trust does not reside solely in model ability. It provides an alternative explanation to the observations that adding transparency features often increase people's trust even if the model should not be relied upon [3, 59, 66, 70]: they may have enhanced people's intention and process based trust rather than ability based trust. Future research should further unpack the dimensions of trustworthy AI and their relations with conceptually relevant constructs, especially behavioral outcomes such as reliance and compliance [65].

The second challenge arises from our lack of understanding of what constitutes trustworthiness cues in AI systems. A conceptual analysis as we did in this paper is not enough. Future work should empirically study what people actually pay attention to and how they process them when making trust judgments, similar to what has been done in the web trust literature [14, 15, 40, 56]. To understand the effect of a trustworthiness cue, we echo the point made by Jacovi et al. [25] that it should be studied in relation to different levels of model trustworthiness. This aligns with the common practice in web trust literature where the effect of a web design feature is studied in contrast for web sites with high- versus low-credibility content [35]. Such an evaluation protocol allows identifying cues with low calibration (resulting in similar trust judgments for models with different trustworthiness) and also a naturalistic setting to avoid the response bias problem, i.e., users may recognize they should examine certain features yet rarely do so in actual practice [12, 40]. Through joint efforts of empirical analysis and theory development, we may outline a more complete design space of reliable trustworthiness cues to guide technology creators' choices [15, 51].

Empowering users to make accurate trust judgments. To guard against deceptive or flawed design of trustworthiness cues, a complementary area for responsible trust in AI is to explore means to empower end users to make more accurate trust judgments. Valuable lessons can again be drawn from what researchers have done for supporting web users, among which we highlight two areas of work. One is to provide training materials or guidance for users

to assess the system more critically (see review in [51]), such as a checklist to assess trustworthiness attributes, and to recognize irrelevant cues or unfounded trust heuristics. The other is to provide independent augmenting tooling to truthfully highlight an AI system's trustworthiness cues, which the creators may have downplayed or hidden [29, 53, 68]. Schwarz and Morris [53] developed visualization to augment web search results, displaying metrics that reflect the quality of content in a web site and making visible otherwise hidden information that provides supporting evidence for its level of trustworthiness, such as the web site's PageRank information and visiting patterns of other users. Yamamoto and Tanaka [68] built a system that shows scores of trustworthiness attributes of web sites and re-ranks the search results.

Leveraging model-extrinsic social, organizational, and industrial mechanisms to provide reliable trustworthiness cues. Communication literature points to many heuristics that people develop through social interactions and based on social structures [57], and these observations encourage looking into model-extrinsic mechanisms to generate cues that provide supporting evidence for the model trustworthiness attributes. As shown in the example of HealthChecker, a prominent *authority heuristic* can be invoked by communicating the model governance structure; a *source reputation heuristic* can be triggered by communicating the legitimacy of model provenance and track record of service. Also, people have a tendency to follow the opinion of many others (*bandwagon heuristic* [57]). A recent study explored features of "social transparency" in AI systems [9], by showing other users' interaction outcomes and feedback, and found them to help calibrate user trust by tapping into the bandwagon heuristic, among others. When these mechanisms are aligned with efforts needed to establish organizational and regulatory ecosystems for the assurance of trustworthy AI [9, 30, 54], they are likely to satisfy the calibration condition. While there exist non-trivial issues to ensure responsible implementation of these mechanisms and truthful communication, trustworthiness cues from these mechanisms are relatively expensive to obtain, which is another advantage to advocate for their use.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[2] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[4] Rebecca BliegeBird and EricAlden Smith. 2005. Signaling theory, strategic interaction, and symbolic capital. *Current anthropology* 46, 2 (2005), 221–248.

[5] Shelly Chaiken. 1980. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology* 39, 5 (1980), 752.

[6] Brian L Connelly, S Trevis Certo, R Duane Ireland, and Christopher R Reutzel. 2011. Signaling theory: A review and assessment. *Journal of management* 37, 1 (2011), 39–67.

[7] Graham Dietz and Deanne N Den Hartog. 2006. Measuring trust inside organisations. *Personnel review* (2006).

[8] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[9] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[10] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. 2021. The who in explainable ai: How ai background shapes perceptions of ai explanations. *arXiv preprint arXiv:2107.13509* (2021).

[11] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[12] Gunther Eysenbach, John Powell, Oliver Kuss, and Eun-Ryoung Sa. 2002. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. *Jama* 287, 20 (2002), 2691–2700.

[13] Andrew J Flanagin and Miriam J Metzger. 2007. The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New media & society* 9, 2 (2007), 319–342.

[14] Brian J Fogg. 2003. Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI'03 extended abstracts on human factors in computing systems*. 722–723.

[15] Brian J Fogg, Jonathan Marshall, Othman Laraki, Alex Osipovich, Chris Varma, Nicholas Fang, Jyoti Paul, Akshay Rangnekar, John Shon, Preeti Swani, et al. 2001. What makes web sites credible? A report on a large quantitative study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 61–68.

[16] Brian J Fogg, Cathy Soohoo, David R Danielson, Leslie Marable, Julianne Stanford, and Ellen R Tauber. 2003. How do users evaluate the credibility of Web sites? A study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences*. 1–15.

[17] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.

[18] James J Gibson. 1977. The theory of affordances. *Hilldale, USA* 1, 2 (1977), 67–82.

[19] Anthony Giddens. 1984. *The constitution of society: Outline of the theory of structuration*. Univ of California Press.

[20] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[21] Brian Hilligoss and Soo Young Rieh. 2008. Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management* 44, 4 (2008), 1467–1484.

[22] Michael Hind, Stephanie Houde, Jacquelyn Martino, Aleksandra Mojsilovic, David Piorkowski, John Richards, and Kush R Varshney. 2020. Experiences with improving the transparency of ai models and services. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[23] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[24] Jake M Hofman, Daniel G Goldstein, and Jessica Hullman. 2020. How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–12.

[25] 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*. 624–635.

[26] Monique WM Jaspers, Thiemo Steen, Cor Van Den Bos, and Maud Geenen. 2004. The think aloud method: a guide to user interface design. *International journal of medical informatics* 73, 11-12 (2004), 781–795.

[27] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

[28] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[29] Aniket Kittur, Bongwon Suh, and Ed H Chi. 2008. Can you ever trust a Wiki? Impacting perceived trustworthiness in Wikipedia. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 477–480.

[30] 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*. 262–271.

[31] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[32] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical

Studies. *arXiv preprint arXiv:2112.11471* (2021).
[33] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270.
[34] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
[35] Q Vera Liao and Wai-Tat Fu. 2014. Age differences in credibility judgments of online health information. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 1 (2014), 1–23.
[36] Q Vera Liao and Kush R Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790* (2021).
[37] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.
[38] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
[39] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.
[40] Miriam J Metzger. 2007. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American society for information science and technology* 58, 13 (2007), 2078–2091.
[41] Miriam J Metzger, Andrew J Flanagin, and Ryan B Medders. 2010. Social and heuristic approaches to credibility evaluation online. *Journal of communication* 60, 3 (2010), 413–439.
[42] Barbara Misztal. 2013. *Trust in modern societies: The search for the bases of social order*. John Wiley & Sons.
[43] 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*. 220–229.
[44] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1, 11 (2019), 501–507.
[45] Donald A Norman. 1988. *The psychology of everyday things*. Basic books.
[46] Daniel J O'Keefe. 2013. The elaboration likelihood model. *The Sage handbook of persuasion: Developments in theory and practice* (2013), 137–149.
[47] Lace MK Padilla, Maia Powell, Matthew Kay, and Jessica Hullman. 2021. Uncertain about uncertainty: How qualitative expressions of forecaster confidence impact decision-making with uncertainty visualizations. *Frontiers in Psychology* (2021), 3747.
[48] Richard E Petty and John T Cacioppo. 1984. Source factors and the elaboration likelihood model of persuasion. *ACR North American Advances* (1984).
[49] Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion*. Springer, 1–24.
[50] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.
[51] Soo Young Rieh and David R Danielson. 2007. Credibility: A multidisciplinary framework. *Annual review of information science and technology* 41, 1 (2007), 307–364.
[52] Justus Robertson, Athanasios Vasileios Kokkinakis, Jonathan Hook, Ben Kirman, Florian Block, Marian F Ursu, Sagarika Patra, Simon Demediuk, Anders Drachen, and Oluseyi Olarewaju. 2021. Wait, But Why?: Assessing Behavior Explanation Strategies for Real-Time Strategy Games. In *26th International Conference on Intelligent User Interfaces*. 32–42.
[53] Julia Schwarz and Meredith Morris. 2011. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1245–1254.
[54] Ben Shneiderman. 2020. Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy Human-Centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 4 (2020), 1–31.
[55] Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal* 31, 2 (2018), 47–53.
[56] Elizabeth Sillence, Pam Briggs, Lesley Fishwick, and Peter Harris. 2004. Trust and mistrust of online health sites. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 663–670.
[57] S Shyam Sundar. 2008. *The MAIN model: A heuristic approach to understanding technology effects on credibility*. MacArthur Foundation Digital Media and Learning Initiative.
[58] S Shyam Sundar and Jinyoung Kim. 2019. Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on human factors in computing systems*. 1–9.
[59] Harini Suresh, Natalie Lao, and Ilaria Liccardi. 2020. Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making. In *12th ACM Conference on Web Science*. 315–324.
[60] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*. 109–119.
[61] Lauren Thornton, Bran Knowles, and Gordon Blair. 2021. Fifty Shades of Grey: In Praise of a Nuanced Approach Towards Trustworthy Design. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 64–76.
[62] 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*. 272–283.
[63] Shawn Tseng and BJ Fogg. 1999. Credibility and computing technology. *Commun. ACM* 42, 5 (1999), 39–44.
[64] Kush R Varshney. 2019. Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 26–29.
[65] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–39.
[66] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. 318–328.
[67] C Nadine Wathen and Jacquelyn Burkell. 2002. Believe it or not: Factors influencing credibility on the Web. *Journal of the American society for information science and technology* 53, 2 (2002), 134–144.
[68] Yusuke Yamamoto and Katsumi Tanaka. 2011. Enhancing credibility judgment of web search results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1235–1244.
[69] Amotz Zahavi. 1975. Mate selection—a selection for a handicap. *Journal of theoretical Biology* 53, 1 (1975), 205–214.
[70] Yunfeng Zhang, Q Vera Liao, and Rachel KE 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*. 295–305.