MATCH: A Conceptual Model on Trust Judgments in AI Towards Designing for Responsible Trust

Q. VERA LIAO, Microsoft Research, Canada
S. SHYAM SUNDAR, Pennsylvania State University, USA

Current literature on "trust in AI" is often fixated on what constitutes trustworthy AI in principle. With individual AI systems differing in their level of trustworthiness, two open questions come to the fore: how should system trustworthiness be responsibly communicated to ensure appropriate and equitable trust judgments by different users, and how can we protect users from deceived trust? We draw from human factors and communication 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 such cues are processed by people to make trust judgments. Besides AI-generated content, we highlight *transparency* and *interaction* as AI systems' affordances for a variety of trustworthiness cues. By bringing to light the plurality 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 for the industry to regulate this space and prevent malicious use. Towards these goals, we define the concepts of and requirements for *warranted trustworthiness cues* and *expensive trustworthiness cues* to help technology creators identify appropriate cues to use. We also discuss future directions for research and industry efforts towards establishing responsible trust in AI.

ACM Reference Format:
Q. Vera Liao and S. Shyam Sundar. 2018. MATCH: A Conceptual Model on Trust Judgments in AI Towards Designing for Responsible Trust. In *CHI 2022 Workshop*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/1122445.1122456

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 dominantly structured around the guiding principles towards trustworthy AI [58, 60], often as a way to operationalize principles for responsible and ethical AI [41], such as ensuring effectiveness, fairness, transparency, robustness, privacy, security, and serving human values. These principles are inherently techno-centric, focusing on what constitutes the trustworthiness of AI, when in fact trust is a human perception and judgment, which can be formally defined as a judgment of dependability in situations characterized by vulnerability [31, 61]. The same AI technology can be perceived 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 defective behavioral outcomes such as over-reliance and misuse.

We argue that the AI field's fixation on algorithmic 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 to its users but communicated through *trustworthiness cues*, which are broadly presented in interface features, documentation, and other modes of information such as speech acts for conversational

Manuscript submitted to ACM

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 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, 18, 34] with one goal, among others, being engendering trust in users.

By framing the design of AI systems as communicating 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 deceived 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 to make 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, 25, 56, 66]. Some user groups are more vulnerable to these harms than others, such as AI novices [56], people working in cognitively constrained settings [48], and even those with certain personality traits [16]. 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, to ensure appropriate user trust and equitable user experiences with AI systems.

To facilitate such an understanding, we present a conceptual model, named MATCH, of how trustworthiness is communicated in AI systems and processed by users to make trust judgments (Section 3), based on communication and human factors literature (Section 2). MATCH decouples the trustworthiness attributes of the underlying AI model(s) and trustworthiness cues presented to the users, via three types of *affordance* of AI systems: AI-generated content (e.g., predictions), transparency, and interaction. MATCH also highlights the plurality of people's cognitive processes to make trust judgments: instead of always being processed analytically to form rational trust judgments, trustworthiness cues often invoke cognitive *heuristics* for people to make speedy, but sometimes flawed, judgments. Communication theories further inform the varied tendencies to engage in heuristic trust judgment between different user groups.

On the basis of MATCH, we consider what constitutes "good" trustworthiness cues that technology creators should use, for which we define the concept of *warranted trustworthiness cues* with a list of requirements (Section 3.4). 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 deception of user trust. We further reflect on MATCH and prior work on trust in technologies to propose areas of call to action to build responsible trust in AI (Section 4).

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 [60], has been considered as an operational point to implement ethical AI principles [58]. Building on the classic ABI framework from the social sciences [7, 36], which prescribes trustworthy characteristics of a trustee as Ability, Benevolence, and Integrity, Toreini et al. [58] propose four categories of trustworthiness technologies for AI, namely Fairness, Explainability, Auditability and Safety

(FEAS). In a similar vein, Varshney [60] maps out the trustworthy qualities of AI as predictive accuracy, robustness, 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 [50] proposes a three-layer governance structure:reliable systems, safety culture, and trustworthiness certification by independent oversight. Knowles and Richards [27] contend that the public distrust of AI originates from the underdevelopment of a regulatory ecosystem that would guarantee AI's trustworthiness, then 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 [51, 58], 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. [57] demonstrate the nuances in implementing trustworthiness principles such as transparency and provenance. They highlight the gaps between these principles and actually promoting user trust, which 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, Jacovi et al. [22], inspired by the literature of interpersonal trust [39], formalized human trust in AI as "contractual trust," such that trust between a user and an AI model is anticipating that some implicit or explicit 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. 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 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

To disentangle human trust and model trustworthiness, we further delve into people's cognitive process of trust judgments, by drawing inspiration from the literature on trust in automation and web technologies. Research on automation often studies human trust in association with the outcome of machine reliance, and dedicates 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 the information dependability on web sites [59], with the bulk of research conducted under the term "web credibility" [47]. For simplicity, we use the term "web trust".

Trust in automation. Our perspective is most directly informed by the seminal paper by Lee and See [31]. By synthesizing related literature, the paper proposes a conceptual model describing the process of trust formation in automation. Below we call out a few key points highlighted by 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 for inter-personal trust [36]. Lee and Moray [30] adapts the ABI model to three dimensions that more suitably characterize automated systems: performance (ability)—*what* the automation does; process (integrity)—*how* the automation operates; and purpose (benevolence)—*for what* the automation was developed. Lee and See [31] show that many necessary characteristics of trustee 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, including subject to the influence of individual, organizational and cultural contexts. Importantly, other factors can 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 current 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 aspect of trust.

Our proposed model also emphasizes the mediating role of information display between the underlying system trustworthiness and people's trust judgments, and the point 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 users' trust judgments.

Effects of information cues on trust in web technologies: communication perspectives and heuristic approaches. The early 2000s saw a rise in research on how people make trust judgments of web sites, including theoretical frameworks on what elements of web technologies influence people's trust [12, 13, 19, 37, 47, 53, 63]. Practical means, including design guidelines [14, 15] and tooling [49, 64], have come out of this line of work, both to facilitate technology creators' design of trustworthiness and empower web users to make better trust judgments.

Much of this literature is based on communication theories of **dual-processes models** for attitude formation, including Petty and Cacioppo [45]'s elaboration likelihood model (ELM) and Chaiken [5]'s heuristic-systematic model (HSM) (also related to Kahneman [24]'s System 1 and System 2 thinking). These theories postulate that web users engage in two cognitive processes to assess a website: "systematic" processing by paying attention to information content and performing a rigorous evaluation, and "heuristic" processing by attending to *cues* about the information quality provided by the interface, which 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" [63]. 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 allows asking 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 [14, 52]. By surveying 2500 participants, Fogg [13] summarizes 18 types of cues people frequently notice on a website to base their trust judgments on, such as information structure, name recognition, advertising, with the most frequently mentioned cue being the "design look." By conducting focus groups with 109 participants, Metzger et al. [38] showed 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 have also developed theoretical frameworks to account for the types of cues in web technologies [19, 53, 63]. The MAIN model developed by Sundar [53] has had a long-lasting impact. Its central thesis is that a given technology has certain "affordances" capable of cuing cognitive heuristics pertinent to trust judgments (**affordance-cue-heuristic** approach). Affordance is a concept in psychology and HCI (human-computer interaction) literature, defined as displayed

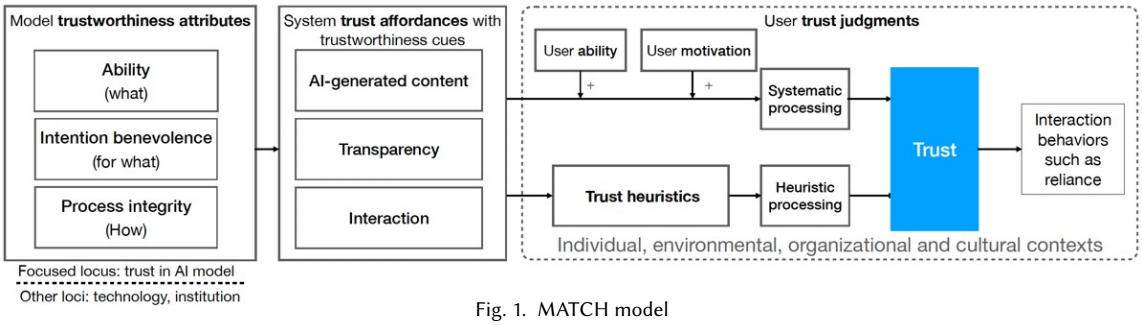

Fig. 1. MATCH model

properties of a system suggesting ways in which it could be interacted or used [17, 42]. The MAIN model earned its name by specifying four types of affordances common to Web technologies to provide trust-related cues : Modality (e.g., visual modality cues realism heuristic), Agency (e.g., the identity 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 [53]. These cues can risk invoking unwarranted trust, as not all of them are directly linked to the content trustworthiness, especially with malicious websites. However, with the deluge of online information, they are appealing for users to base their trust judgments on, especially for those lacking ability or motivation to engage in a careful reading of the contents [53].

Equipped with these theoretical bases, we develop a conceptual model to describe trust judgment of AI systems in the next section. 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 the trustworthiness of AI 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**) trustworthiness attributes as the basis of trustworthiness, system affordances (**A**) to communicate trustworthiness (**T**) cues (**C**) of the model, and users' dual cognitive processes to process these cues through systematic and heuristic processing by invoking trust-related heuristics (**H**) (MATCH model). 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 brand or institution [47] and trust in AI-as-a-technology [27]. We focus on trust as an attitude rather than its impact 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 the individual, environmental, organizational and cultural contexts.

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

As reviewed, many define the trustworthiness of technologies [28, 30, 31, 58] based on the classic ABI model. Lee and Moray [30] adapted ABI to the context of automation using the terms 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 should be agnostic to how the trustworthiness attributes are operationalized, and we welcome future work to expand these dimensions or explore alternative operationalizations.

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 core attribute of AI ability. Considering other trustworthy AI principles [58, 60], *performance fairness* (e.g., absence of

performance differences between different demographics) and *performance robustness* (against data shift and poisoning, for example) 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, reflecting *for what* the AI is developed. Besides *intended use* (e.g., for social good, serving user values), we also consider *intended compliance* (e.g., privacy-preserving, security compliance) 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 is context- and user-dependent, such as the absence of flawed logic, optimizing for the right goal, and aligning with domain process. While process integrity could impact the AI's ability, the former creates a different basis of trust on dispositional integrity rather than the 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 noises, as we discuss in the following.

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, and contribute to, users' trust judgments. For individuals, trust is often conceived as a judgment of dependability and people do not necessarily base it solely on 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 [53], one may identify trustworthiness cues by conceptualizing what are the *affordances* of a given type of technology to allow such cues.

The question of "what are the trust affordances [57] 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 [29] 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 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.

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, a prediction 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. Depending on the design, for example, when these cues are provided or not, in some cases they can also cue people's judgment regarding 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), explainable AI (XAI) features, and model documentation [2, 20, 40] (commonly including model provenance information [27, 57] about how and why it was developed). Recent literature discusses governance structures to ensure trustworthy AI [46, 50], such as internal reviews, testing, independent and

government oversight, and so on. Communicating the process and outcomes of such governance structures can also be considered a form of transparency. Transparency allows cues 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 warrant trustworthiness in itself. This is related to Jacovi et al. [22]'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.

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 using the system). The roles of interaction and interaction design are often overlooked in the current "trust in AI" literature. 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" [45] on a web site, such as the design look, source reputation, or social information. Some interaction affordances are enabled by the model ability, such as customization in guiding the model's behavior, and can directly cue trustworthiness attributes such as ability 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 how different interaction features of AI systems 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, 24, 45]. The basic idea is that people process information to form a judgment 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 speedy judgment. MATCH highlights the roles of trust heuristics and individual differences.

Trust heuristics are any rules-of-thumb a user has associating a given cue to a judgment of trustworthiness. There are many ways for people to acquire trust heuristics. Some are common cognitive heuristics applied to the context of AI. For example, people tend to have an *authority heuristic* by following the opinion of an authority on the subject matter [38]. This heuristic can be invoked when seeing a certification from a regulatory body that audited the AI. Others are technology-specific heuristics learned about a type of technology. For example, some groups of users have a prominent *machine heuristic*, believing machines are more reliable than humans, which can be triggered by simply noticing that predictions are made by an AI system [54]. The phenomenon of XAI features leading to over-trust [3, 11, 25, 56, 66] can be attributed to an "explainability heuristic" [10, 33] that superficially associates being explainable with being capable. 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. It is possible to enlist common heuristics based on psychology and communication theories [38, 53], or by empirically studying what heuristics are frequently invoked in target user groups' interaction with the AI, for example by using think-aloud method [21, 23].

Individual differences in systematic vs. heuristic processing. Different users 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 repeatedly found that XAI features bring less benefit, or even harms (leading to over-trust and over-reliance), to certain user groups such as AI novices or users working in

cognitively constraining settings [16, 48, 56]. Theories on dual-process models [45] have established that when people 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 the theory 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 [43–45]. By highlighting these individual differences, we encourage technology creators to carefully examine and mitigate the potential inequalities of experience for users without an ideal profile of ability or motivation.

3.4 What are "good" trustworthiness cues?

Equipped with 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 likely be 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 (**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**).

Relevance condition. Technology creators should pay attention to *prominent irrelevant trustworthiness cues*—what users pay attention to when making trust judgments but are disassociated with the three trustworthiness attributes of the model, such as the surface design look or an irrelevant URL link. Often, some irrelevant cues are unavoidable because they support other user goals, but they can impact user trust in unintended ways. The unintended impact can be mitigated by making these cues less prominent during users' trust-development stage, or providing interventions to disrupt invoking of trust heuristics (e.g., a reminder that the design is inherited from a template).

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, 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 oversights.

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) the cue reflects. The existence and level of calibration can be assessed with a **formal analysis**— measuring the change in people's trust judgment based on the trustworthiness cue by manipulating the quality of the corresponding model trustworthiness attribute(s) [22].

However it is not always feasible to perform costly formal analysis to quantify the calibration of trustworthiness cues, 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 trustworthiness cues with 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**).

To categorize a heuristic as "unfounded"—with little evidence to support or low probability to hold—depends on the context and the user, while some heuristics should be generally avoided invoking. For example, a recent study [35] shows that people follow the cognitive heuristic of *confirmation bias*, whereby the agreement of AI predictions with their own judgment is seen as indicating high AI ability. This heuristic may be unfounded if users are novices to the decision task, but could be acceptable for domain-expert users. The explainability heuristic [33] (associating being explainable with being capable) and numeric heuristic found in an XAI study [10] (associating numerical explanations with algorithmic intelligence) are generally unfounded. According to the prominence-interpretation theory on web trust [13], 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. Technology creators should strive to leverage common cognitive heuristics and reverse-engineer mechanisms to make them better founded, or cultivate founded heuristics in users by providing training, guidance, or reinforcement mechanisms. For example, it is known that people have an *anchoring heuristic* 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 [65]. 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 the social goal. Since individuals have motivation 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 auditing trails and endorsement from others are generally costly in time and efforts. 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 that are within their affordable range. Like many RAI practices, the costly implementation may also risk marginalizing smaller business entities and creating inequalities in the AI industry. While we suggest the role of expense to safeguard the truthfulness

condition, a much more nuanced view on its relations with resources, gains, other motivators and constraints need to be developed to inform policy and industry practices.

4 DISCUSSION: TOWARDS RESPONSIBLE TRUST IN AI

This conceptual work is intended to synthesize relevant theories on trust in technologies, elucidate the cognitive mechanisms of trust, call out the requirements for using reliable trustworthiness cues, and with these, invite future research to develop practical means for building responsible trust in AI. We discuss a few directions below.

Understanding and regulating the space of trustworthiness cues. Based on MATCH, technology creators' responsible use of trustworthiness cues has two essential 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 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 [28]. It provides an alternative explanation to the observations that transparency features can increase people's trust [3, 55, 62, 66]: they may have enhanced people's intention and process based trust rather than ability based trust. The second challenge arises from our lack of understanding about what constitutes trustworthiness cues in AI systems. 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 web trust literature [13, 14, 37, 52]. To investigate the effect of a trustworthiness cue, we echo the point in [22, 32] that it should be studied in relation with different levels of model trustworthiness. 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 [14, 47].

Empowering users to make accurate trust judgments. To guard against defective and deceptive use of trustworthiness cues, a complementary area for responsible trust in AI is to explore practical 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 [47], such as a checklist or reminder 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 hidden [26, 49, 64]. For example, Schwarz and Morris [49] developed a visualization to augment web search results, which presents metrics that reflect the quality of content in a web site and makes visible information that provides supporting evidence for its level of trustworthiness, such as the web site's PageRank information and other users' visiting patterns.

Leveraging model-extrinsic social, organizational and industrial mechanisms to provide reliable trustworthiness cues. Communication literature points to many heuristics that people have based on social structures and interactions [53]. They encourage looking into model-extrinsic mechanisms to generate cues that provide supporting evidence for the model trustworthiness attributes. For example, an *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. 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[53], among others. Trustworthiness cues from these mechanisms are relatively expensive to obtain, which is another advantage to advocate for their use. When these mechanisms are aligned with efforts needed to establish social, organizational and regulatory ecosystems for the assurance of trustworthy AI [9, 27, 50], they are likely to satisfy the calibration condition.

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] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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).

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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).

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

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

[32] 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.

[33] Q Vera Liao and Kush R Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790* (2021).

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] Barbara Misztal. 2013. *Trust in modern societies: The search for the bases of social order*. John Wiley & Sons.

[40] 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.

[41] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1, 11 (2019), 501–507.

[42] Donald A Norman. 1988. *The psychology of everyday things*. Basic books.

[43] Daniel J O'Keefe. 2013. The elaboration likelihood model. *The Sage handbook of persuasion: Developments in theory and practice* (2013), 137–149.

[44] Richard E Petty and John T Cacioppo. 1984. Source factors and the elaboration likelihood model of persuasion. *ACR North American Advances* (1984).

[45] Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion*. Springer, 1–24.

[46] 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.

[47] Soo Young Rieh and David R Danielson. 2007. Credibility: A multidisciplinary framework. *Annual review of information science and technology* 41, 1 (2007), 307–364.

[48] 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.

[49] 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.

[50] 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.

[51] Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal* 31, 2 (2018), 47–53.

[52] 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.

[53] S Shyam Sundar. 2008. *The MAIN model: A heuristic approach to understanding technology effects on credibility*. MacArthur Foundation Digital Media and Learning Initiative.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] Shawn Tseng and BJ Fogg. 1999. Credibility and computing technology. *Commun. ACM* 42, 5 (1999), 39–44.

[60] Kush R Varshney. 2019. Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 26–29.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] Amotz Zahavi. 1975. Mate selection—a selection for a handicap. *Journal of theoretical Biology* 53, 1 (1975), 205–214.

[66] 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.