Draft NISTIR 8332

Trust and Artificial Intelligence

Brian Stanton
Theodore Jensen

This draft publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8332-draft

National Institute of
Standards and Technology
U.S. Department of Commerce

Draft NISTIR 8332

Trust and Artificial Intelligence

Brian Stanton
Information Technology Laboratory

Theodore Jensen
University of Connecticut

This draft publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8332-draft

March 2021

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
James K. Olthoff, Performing the Non-Exclusive Functions and Duties of the Under Secretary of Commerce for Standards and Technology & Director, National Institute of Standards and Technology

National Institute of Standards and Technology NISTIR 8332-draft
Natl. Inst. Stand. Technol. NISTIR 8332-draft
29 pages (March 2021)

This draft publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8332-draft

Please send comments on this document to: AIUserTrustComments@nist.gov

Abstract

The artificial intelligence (AI) revolution is upon us, with the promise of advances such as driverless cars, smart buildings, automated health diagnostics and improved security monitoring. In fact, many people already have AI in their lives as "personal" assistants that allow them to search the internet, make phone calls, and create reminder lists through voice commands. Whether consumers know that those systems are AI is unclear. However, reliance on those systems implies that they are deemed trustworthy to some degree. Many current efforts are aimed to assess AI system trustworthiness through measurements of Accuracy, Reliability, and Explainability, among other system characteristics. While these characteristics are necessary, determining that the AI system is trustworthy because it meets its system requirements won't ensure widespread adoption of AI. It is the user, the human affected by the AI, who ultimately places their trust in the system.

The study of trust in automated systems has been a topic of psychological study previously. However, artificial intelligence systems pose unique challenges for user trust. AI systems operate using patterns in massive amounts of data. No longer are we asking automation to do human tasks, we are asking it to do tasks that we can't. Moreover, AI has been built to dynamically update its set of beliefs (i.e. "learn"), a process that is not easily understood even by its designers. Because of this complexity and unpredictability, the AI user has to trust the AI, changing the dynamic between user and system into a relationship. Alongside research toward building trustworthy systems, understanding user trust in AI will be necessary in order to achieve the benefits and minimize the risks of this new technology.

Key words

Artificial Intelligence; Automation; Cognition; Collaboration; Perception; System Characteristics; Trust; Trustworthiness; User; User Experience,

Table of Contents

List of Tables

List of Figures

List of Equations

1. Introduction

Although the study of user trust in automated systems has been a topic of psychological study previously, Artificial Intelligence (AI) changes previous User Interface paradigms dramatically. AI systems can be trained to “notice” patterns in large amounts of data that are impossible for the human brain to comprehend. No longer are we asking automation to do our tasks—we are asking it to do tasks that we can’t. Asking the AI to perform the same task on two different occasions may result in two different answers as the AI has “learned” in the time between the two requests. AI has the ability to alter its own programming in ways that even those who build AI systems can’t always predict. Given this significant degree of unpredictability, the AI user must ultimately decide whether or not to trust the AI. The dynamic between AI user and AI system is a relationship, a partnership where user trust is an essential part.

To achieve the improved productivity and quality of life that are hoped for with AI, an understanding of user trust is critical. We outline the importance of user trust for the development of AI systems by first establishing the integral role of trust in our own evolutionary history, and how this has shaped our current cognitive processes. We then briefly discuss research on factors in trust between humans and summarize the substantial body of research that has extended the notion of trust to operators of automated systems.

Next, we deal specifically with the unique trust challenges associated with AI. We distinguish between the notion of AI’s technical trustworthiness and user’s trust. Then we propose an illustrative equation representing a user’s level of trust in an AI system, which involves a judgement of its technical trustworthiness characteristics with respect to the operational context. This document is also intended to highlight important areas of future research toward understanding how users trust AI systems. These areas of future research are placed in tables within the sections.

2. Trust is a Human Trait

2.1. Purpose of Trust

Trust serves as a mechanism for reducing complexity [1]. When we make a decision to trust, we are managing the inherent uncertainty of an interaction partner’s future actions by limiting the number of potential outcomes. Distrust serves the same purpose. As Kaya [2] states,

> *“In ancestral environments, distrust was key for survival, given that it led humans to be cautious against their most deadly enemies: other humans. Individuals who considered other humans to be potentially dangerous and exploitative were more likely to stay alive and pass on their genes to future generations”*

The development of trust alleviates the individual of having the sole responsibility for survival. Trust allows one to harness cooperative advantages. Taylor [3] states in her book, *The Tending Instinct*:

> *As the insistence of day to day survival needs has subsided, the deeper significance of group life has assumed clarity. The cooperative tasks of hunting and warfare represent the least of what the social group can accomplish.*

Overall, in the evolutionary landscape, trust and distrust are used to manage the benefits and risks of social interaction. Reliance on another individual can offer advantages, but it simultaneously makes one vulnerable to exploitation and deceit. If you trust too little, you will be left wanting; trust too much and you will be taken advantage of. Game theory research has confirmed that conditional trust, a strategy for discerning between the trustworthy and untrustworthy, is evolutionarily advantageous [4] [5] [6]. As such, trust was fundamental to our survival and continues to drive our interactions.

2.2. Distrust & Cognition

The role of trust and distrust in our thinking align with their central place in our evolutionary struggle. In particular, human cognition is largely characterized by congruency—we tend to process incoming information in ways that align with a prior referent. This is explained in Kahneman's book "Thinking Fast and Slow," as Confirmation Bias [7]. Accessibility effects, likewise, are characterized by exposure to an initial stimuli which alters subsequent processing—a positive prime (the initial referent) invokes a congruently more positive evaluation of an unrelated target than does a negative prime [8]. Distrust, however, has been found to reduce such effects of congruent processing. Instead, distrust appears to invoke the consideration of incongruent alternatives [8].

For instance, this has been demonstrated in the Wason Rule Discovery Task, where participants complete the following two steps after being shown the number sequence "2, 4, 6": 1) generate a hypothesized rule characterizing the number sequence and 2) generate several number sequences to test their hypothesized rule. In general, most individuals hypothesize the rule "+2" and generate only sequences that follow their rule for the second step (positive hypothesis tests). This underscores our tendency toward congruent processing, which, in this case, often leads to a failure to discover the true rule (i.e., "any series of increasing numbers"). Experiments showed that individuals low in dispositional trust and those primed with distrust were found to be significantly more likely to generate sequences that did not follow their rule (negative hypothesis tests) [9]. Distrust improved performance on the task by invoking a consideration of alternatives. Similarly, a state of distrust has been found to lead to faster responses to incongruent concepts and a greater number of incongruent free associations [10].

This effect of distrust in disrupting our congruent processing is understandable given its function to protect ourselves from deceit. Mayo [8] aptly summarizes this:

> *"...when the possibility is entertained that things are not as they seem, the mental system's pattern of activation involves incongruence; that is, it spontaneously considers the alternatives to the given stimuli and*

> *searches for dissimilarities in an attempt not to be influenced by an untrustworthy environment."*

Highlighted again in this cognitive consideration of distrust is the role of risk. The distrust mindset makes more salient one's vulnerability to the actions of other actors. This reminds us that trust is inescapably linked to perception of risk in a given context. Following from game theory, conditional trust and distrust protect the individual from deceptive others, while still reaping the potential benefits of cooperation.

The cognitive mechanisms that drive our everyday willingness to rely on peers were ultimately borne out in our environment of evolutionary adaptation [11] [12]. In other words, our evolutionary history is informative of how we manage risk and uncertainty with our trust today.

2.3. Trust, Distrust, and Cooperation: The Role They Play

Trust and distrust are so fundamental that they are often concealed within the most mundane decisions in our daily lives. Without some trust we would not leave our homes due to overwhelming fear of others. Meanwhile, distrust permits us to navigate a world of potentially deceitful actors and misinformation.

As Luhmann [13] noted, trust and distrust are not opposites, but functional equivalents. We use both to reconcile the uncertainty of the future with our present—deciding only that someone is not to be trusted does not reduce complexity, but considering the reasons to distrust them does [13]. Lewicki, McAllister, and Bies [14] proposed that many organizational relationships, and often the healthiest, are characterized by simultaneously high levels of trust and of distrust (e.g., "trust but verify"). We constantly use both trust and distrust to manage the risk in our interactions with others and achieve favorable outcomes.

Gambetta [15] illustrates how the modern trust environment consists of an interplay between trust among individuals and rules and regulations that govern our behavior:

> *"If we were blessed with an unlimited computational ability to map out all possible contingencies in enforceable contracts, trust would not be a problem".*

Gambetta refers to such contracts or agreements as "economizing on trust," noting that these do not adequately replace trust, but instead serve to reduce the extent to which individuals worry about trust.

This is mirrored by Hill and O'Hara's [11] discussion of legal regulations that enforce "trust that" a party will do something, without necessarily building "trust in" that party. Such regulations can even contribute to distrust, since the trustor may infer that the trustee would not act favorably without rules in place. This stresses that trust remains fundamental to our interactions, even while our species is largely removed from the conditions in which trust evolved, and lives in a society that largely focuses on doing away with trust via regulatory mechanisms. Its "complexity-reducing" function [1] remains important. As a result, many researchers have identified characteristics that inform a person's trust in another.

2.3.1. Factors that lead to Trusting and Distrusting

Mayer, Davis, and Schoorman's model [16] of trust in organizational relationships gives a parsimonious view of the factors that contribute to a trustor's "willingness to be vulnerable" to a trustee. It is undoubtedly the mostly widely referenced work on trust. The model includes trustor-related, trustee-related, and contextual factors. Each of these factors will be considered in our later discussion of AI user trust.

The central trustor factor is dispositional trust, defined as the trustor's general willingness or tendency to rely on other people [17]. It is viewed as a stable trait across interactions. For AI user trust, we define *User Trust Potential* (UTP) to account for each users' unique predisposition to trust AI. Two users may perceive a system to be equally trustworthy, but UTP accounts for differences in how perceived trustworthiness impacts overall trust.

Trustee factors consist of their ability, benevolence, and integrity or, more specifically, the trustor's perception of these characteristics. Ability is a domain- or context-specific set of skills that the trustee possesses. Benevolence is a sense of goodwill that the trustee has with respect to the trustor. Integrity involves the maintenance of a set of acceptable principles to which the trustee adheres. Mayer et al.'s [16] perceived trustworthiness characteristics are reflective of characteristics proposed in several other researchers' formulations of the construct. For instance, Rempel, Holmes, and Zanna [18] , focusing on trust between romantic partners, identify predictability, dependability, and faith as components of trust. Becker [19] refers to credulity, reliance, and security of the trustee. In each case, the trustee's (perceived) skills, character and intentions understandably relate to a trustor's willingness to be vulnerable. For AI user trust, we define *Perceived System Trustworthiness* (PST) as the user's contextual perceptions of an AI system's characteristics that are relevant for trust. As we shall discuss, this involves perception of a system's various technical characteristics as well as user experience factors. Importantly, we argue that, as in human-human trust, trustworthiness is perceived by the trustor, rather than a direct reflection of trustee characteristics.

Situational factors are unrelated to characteristics of the trustor or trustee. As with the aforementioned characteristics, situational factors relevant to trust relate to the degree of vulnerability that the trustor is exposed to. These may include mechanisms and rules that aim to coerce cooperation or "economize on trust" [15]. Importantly, Mayer et al. [16] distinguish trust from perceived risk. The latter consists of an evaluation of negative and positive outcomes "outside of considerations that involve the relationship with the particular trustee." They suggest that "risk-taking in relationship" or trusting behavior results if the trustor's level of trust exceeds their level of perceived risk. While trust is inherently linked to risk, they are distinct constructs. To account for situational factors in AI user trust, PST is evaluated with respect to the specific deployment context or action that the AI system is performing. Two different tasks or levels of risk will lead to two distinct perceptions of trustworthiness.

The vulnerability in our interactions with technology creates conditions for a similar trust-based interaction. The question of human-technology interaction becomes the following: how does our evolutionarily ingrained and socially conditioned trust mechanism respond to machines?

3. Trust in Automation

3.1. Computers as Social Actors

The Computers as Social Actors (CASA) paradigm lends support to the viability of human-machine trust as a construct. CASA has been used by communication researchers to demonstrate that humans respond socially to computers [20]. In a CASA experiment, a computer replaces one of the humans in the social phenomenon under investigation to see if the social response by the human holds [21]. This method has revealed that people use politeness [21], gender stereotypes [22], and principles of reciprocal disclosure [23] with computers. Notably, the original CASA experiments were conducted with experienced computer users interacting with simple, text-based interfaces [24].

Although CASA does not rule out the unique learned aspects of our interactions with machines, it emphasizes our predisposition to interactions with people. Trust and distrust developed to predict the uncertain behavior of our human peers. It is natural that our use of trust extends to automation.

3.2. Human Factors, Trust and Automation

Human factors researchers began studying trust in response to the increasing prevalence of automation in work systems. Muir [25] was one of the first to challenge the notion that behavior toward automation was based solely on its technical properties. Her view evokes a theme of our preceding discussion of trust between people—an operator simply cannot have complete knowledge of an automated system. The trustor's (operator's) perceptions become important because of the trustee's (automation's) freedom to act, and the trustor's inability to account for all possibilities of the trustee's action.

Muir's [25] gives an example of some people using automated banking machines while others do not, with the properties of the banking machines remaining constant, introducing user trust in technology:

> *"The source of this disparity must lie in the individuals themselves, in something they bring to the situation."*

Experiments subsequently confirmed that operators were able to report on their subjective level of trust in an automated system, that this trust was influenced in sensible ways by system properties, and that trust was correlated with reliance on (use of) automation [26] [27].

Since this early work, researchers have contributed a significant amount of understanding of relevant factors in trust in technology. Lee and See's [28] review emphasizes how the increasing complexity of automated systems necessitates an understanding of trust. Hoff and Bashir [27] reviewed the empirical work that followed Lee and See's [28] and defined three sources of variability in trust in automation: dispositional, situational, and learned. Dispositional factors include the age, culture, and personality of the trustor (i.e., the automation operator or user) among other characteristics. Situational factors concern the context of the human-automation interaction and various aspects of the task, such as workload and risk. Learned trust is a result of system performance characteristics as well as design features that color how performance is interpreted. This three-layer model is compatible with Mayer et al.'s [16] human-human model, which considers trustor characteristics (dispositional), perceived risk (situational), and perceived trustworthiness that is dynamically updated by observing trustee behavior (learned). As previously discussed with respect to Mayer et al.'s model, these human-automation trust factors inform our later discussion of AI user trust.

Even with establishment of human-machine trust as a viable construct, the question of how it relates to human-human trust remains. Indeed, the aforementioned human-automation trust researchers drew from sociological and psychological theories on trust to formulate their own [25] [28]. CASA supports this theoretical extension [20]. But how relevant is our trust mechanism, evolved for interaction with other people, to our interactions with machines? Do we do something different when trusting an automated system?

Madhavan and Wiegmann [29] reviewed several studies comparing perceptions of automated and human aids. They suggest that perceptions of machines as invariant and humans as flexible lead to fundamental differences in trust toward these two different kinds of aids. For instance, the Perfect Automation Schema holds that people expect automation to perform flawlessly. As a result, errors made by automation are more damaging to trust than errors made by automated aids. Studies finding that more anthropomorphic (i.e., humanlike) automation elicits greater "trust resilience" support this notion that more humanlike technology is more readily forgiven [30]. One must question the extent to which perceptions of machine invariance associated with automation will persist with the advent of AI.

4. Trust in Artificial Intelligence

Again, Luhmann's [1] sociological viewpoint stresses the role of trust in the face of uncertainty:

> *"So it is not to be expected that scientific and technological development of civilization will bring events under control, substituting mastery over things for trust as a social mechanism and thus making it unnecessary. Instead, one should expect trust to be increasingly in demand as a means of enduring the complexity of the future which technology will generate."*

Although not specifically referring to technological trustees, Luhmann sets the stage for the specific challenges associated with AI user trust, based in complexity and uncertainty.

4.1. AI Trustworthiness

The use of trustworthy as it applies to computing can be traced back to an email that Bill Gates sent out to all Microsoft employees in 2002 [31]. In this email he states,

> *"…Trustworthy Computing. What I mean by this is that customers will always be able to rely on these systems to be available and to secure*

their information. Trustworthy Computing is computing that is as available, reliable and secure…”. [32] [33] [34]

This practice of Trustworthy Computing continues to be adopted by some in the computer science and system engineering fields. There are: The Institute of Electrical and Electronics Engineers (IEEE) and The International Electrotechnical Commission (IEC)/ The International Organization for Standardization (ISO)/IEEE standard definitions of trustworthiness built around the concept and Gates’ system trustworthiness attributes:

(1) trustworthiness of a computer system such that reliance can be justifiably placed on the service it delivers [33]

(2) of an item, ability to <u>perform as and when required</u> [34] (emphasis added).

It is this second definition that encourages the creation of characteristics an AI must have in order to be trustworthy. The development of characteristics, how to measure them, and what the measurements should be, based on a given AI use case, are all critical to the development of an AI system. Yet, as good as the characteristic definition process is, it doesn’t guarantee that the user will trust the AI. As stated above, dispositional factors of the trustor also influence trust [27], and so not all users will trust an AI system the same. Asserting that an AI system is “worthy of trust” doesn’t mean that it will be automatically trusted.

4.2. User Trust in AI

Much like our trust in other people and in automation is based on perceptions of trustworthiness, user trust in AI is based on perceptions of its trustworthiness. The actual trustworthiness of the AI system is influential insofar as it is perceived by the user. Trust is a function of user perceptions of technical trustworthiness characteristics.

Given a scenario where a user u interacts with an AI system s within a context a, the user’s trust in the system can be represented as $T(u, s, a)$, Figure 1 AI User Trust Scenario

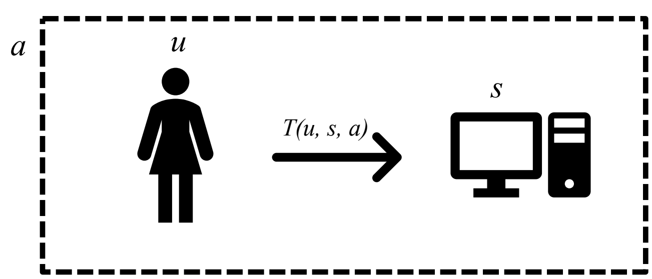

Figure 1 AI User Trust Scenario

The research on human-human and human-automation trust suggest two main sources of variability in trust in an AI system: the user and the system. Therefore, we conceptualize user trust in AI in terms of two main components: *User Trust Potential*, UTP(u), and *Perceived System Trustworthiness*, PST(u, s, a)[1]. User trust can be expressed as a function *f* of these two components:

$$T(u, s, a) = f(UTP(u), PST(u, s, a))$$

Research is needed into the nature of the relationship between UTP and PST. In this document, for illustrative purposes, we consider the two components to be independent and to multiply toward overall trust. Moreover, we consider each as a probability value, such that the product of the two will lie in the range [0, 1], representing the likelihood that user *u* will trust the system *s* to perform the specified action:

$$T(u, s, a) = UTP(u) * PST(u, s, a)$$

We carry this illustrative probabilistic assumption through the remainder of our discussion and examples but emphasize the contextual nature of perceived trustworthiness and trust. Trust is based on the trustee's (system's) expected behavior and should not be interpreted literally as a 'chance' decision. The probabilistic representation allows us to quantitatively express differences in trust due to various factors[2].

4.3. User Trust Potential

What we refer to as *User Trust Potential,* UTP(*u*), consists of the intrinsic personal attributes of the user *u* that affect their trust in AI systems. Characteristics of the user have been suggested as influential in trust in technology [35] [27]. These include attributes such as personality, cultural beliefs, age, gender, experience with other AI systems, and technical competence. More research is needed to establish the role of these and other user variables in trust in AI systems.

Table 1 User Trust Potential Research Question

Research Question
1. What are the set of attributes that define User Trust Potential?

4.4. Perceived System Trustworthiness

What we refer to as *Perceived System Trustworthiness*, PST(*u, s, a*), is made up of a relationship between *User Experience* (UX) and the *Perceived Technical Trustworthiness*

[1] Hoff and Bashir [27] and Mayer et al. [16] refer to situational factors in trust in addition to those related to the trustor and trustee. We account for these within Perceived System Trustworthiness, which consists of the context-based perception of an AI system's trustworthiness.

[2] For instance, a user *u* for whom UTP(*u*) is 0 is indiscriminately distrusting of any AI system with which they interact. A user *u* for whom UTP(*u*) is 1 will not necessarily rely on the system but will trust based on PST. It is likely that most users fall somewhere in the middle of the UTP spectrum, opting to trust based on PST to some extent. It is also possible that users with greater UTP will consistently report greater PST of the particular system. The independence assumption here merely allows us to point out these distinct relevant factors in user trust.

(PTT) of the AI system. These two components can be thought of as front end-related (UX) and back end-related (PTT) factors in the user u's trust of the AI system s in context a.

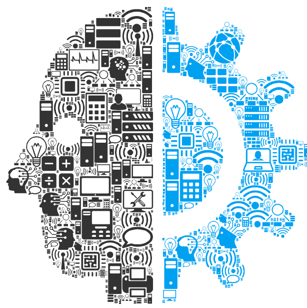

Figure 2 the User Experience Front End and the AI System Trustworthy Characteristics Backend

We first represent Perceived System Trustworthiness as a generalized function g of UX and PTT:

$$PST(u, s, a) = g(UX, PTT)$$

For illustrative purposes, this may be thought of as a multiplicative function of independent probabilities:

Perceived AI System Trustworthiness

$$PST(u, s, a) = UX * PTT$$

Thus, as with overall trust T, PST will lie in the range [0, 1] and represent the degree to which the system is perceived as trustworthy. Further research is needed to identify the relationship between UX and PTT.

4.4.1. User Experience

User Experience represents contributions to *Perceived System Trustworthiness* from user experience design factors external to technical trustworthiness characteristics that make up PTT. These external factors are also associated with user perception.

Usability, the main component of *User Experience*, is made up of three metrics according to an international standard [20]: efficiency, effectiveness, and user satisfaction. These metrics can be measured in different manners. Efficiency can be both task completion rate (the time it took to complete all tasks) and task time (the time that was spent on a single task). Effectiveness can be the number of errors made or the quality of the task output, and User Satisfaction can be amount of frustration, amount of engagement, or enjoyment.

Given all the variations of how to measure usability, for perceived AI system trustworthiness, one usability score is used. There are many different methods of combining usability measures into one score [21] [23] [22], with the most well-known method being “The Single Usability Metric” (SUM) [22]. This method takes as input task time, errors, satisfaction, and task completion and will calculate a SUM score with confidence intervals.

The challenge with the *UX* variable is discovering those usability methods that most influence system trust.

Table 2 User Experience Research Question

Research Question
1. What User Experience Metrics Influence User Trust?
2. How do User Experience Metrics Influence User Trust?

4.4.2. Perceived Technical Trustworthiness

AI system designers and engineers have identified several technical characteristics that are necessary for system trustworthiness. There are, at the time of this writing, nine identified characteristics that define AI system trustworthiness: *Accuracy*, *Reliability*, *Resiliency*, *Objectivity*, *Security*, *Explainability*, *Safety*, *Accountability*, and *Privacy* (*Privacy* added after [36]). From an engineering perspective, an AI system needs these characteristics if it is to be trusted.

From the perspective of user trust, these characteristics are necessary but not sufficient for trust. Ultimately, the user’s perception of available technical information is what contributes to their trust. *Perceived Technical Trustworthiness* can be expressed by the following formula, where c is one of the nine characteristics, and ptt_c is the user’s judgement of characteristic c:

Equation 1 Perceived System Technical Trustworthiness

$$PTT = \sum_{c=1}^{9} ptt_c$$

The variable ptt_c indicates the contribution of each characteristic to overall PTT, and consists of its pertinence to the context, p_c, and the sufficiency of that characteristic’s measured value to the context, s_c:

Equation 2 The Relationship of Perceived Pertinence and Perceived Sufficiency of the Trustworthy Characteristic

$$ptt_c = p_c * s_c$$

This formulation is reminiscent of utility functions used to represent human decision-making quantitatively. The utility of a decision outcome therein is the product of that outcome's probability and its value. High utility of an outcome can be due to either high probability, high value, or both. The sum of the utilities of all possible outcomes represents the expected "payoff."

Perceived Technical Trustworthiness is the sum of each characteristic's perceived sufficiency weighted by its pertinence. Here, high "utility" of a characteristic can occur due to high pertinence, high sufficiency, or both. While not necessarily the same as a "payoff," the sum of these utilities represents the degree of perceived trustworthiness of the system based on contributions from each characteristic. We describe the two components in more detail below.

4.4.2.1. Pertinence

Pertinence is the answer to the question, "How much does this characteristic matter for this context?" Pertinence involves the user's consideration of which technical trustworthiness characteristics are the most consequential based on the unique nature of the use case.

In her model of human-automation trust, Muir [25] proposed that the relative importance of different components of perceived trustworthiness (persistence, technical competence, fiduciary responsibility) is not equal, nor the same across contexts. Likewise, Mayer, Davis, and Schoorman [16] note how context influences the relative importance of each of their perceived trustworthiness characteristics (ability, integrity, and benevolence) to trust. Thus, pertinence is the "weight" of each characteristic's contribution to overall perceived trustworthiness.

If only one characteristic is perceived as contextually important, its perceived pertinence would be 1. If only two characteristics are perceived as important, and equally so, the perceived pertinence for each would be 0.5. It does not imply that a relevant characteristic is less important for trust when it shares pertinence with another. If two characteristics are both deemed critical for contextual performance, they make an equal contribution to PTT.

Pertinence is a perceptual weighting of the importance of c relative to the other characteristics. Thus, all p_c values sum to 1, and each represents a percentage of importance to the overall trustworthiness evaluation. If the measured pertinence of each characteristic, q_c, is rated on a scale where the sum is not 1, this normalized perceived pertinence, p_c, can be obtained by dividing q_c by the sum of all characteristics' ratings on that scale:

Equation 3 Normalization of the Perceived Pertinence Value of a Trustworthy Characteristic

$$p_c = \frac{q_c}{\sum_{i=1}^{9} q_i}$$

Table 3 Pertinence Research Question

Research Question
1. What should the measurement be for Pertinence?

4.4.2.2. Sufficiency

Sufficiency is the answer to the question, "How good is the value of this characteristic for this context?" Sufficiency involves the user's consideration of each characteristic's measured value and a judgement of how suitable that value is with respect to contextual risk.

While pertinence perceptions certainly involve consideration of contextual risk (since completely non-pertinent characteristics are not expected to contribute to negative outcomes), the perception of sufficiency is characterized by a more explicit evaluation of trustworthiness metrics with respect to risk. A higher metric m_c for a given characteristic will be needed to increase perceived trustworthiness under greater perceived risk, r_a. High sufficiency can be the result of a large metric, m_c, or low perceived contextual risk, r_a. Perceived sufficiency may thus be calculated for each characteristic as follows:

Equation 4 The Perceived Sufficiency of an AI Trustworthy Characteristic

$$s_c = \frac{m_c}{r_a}$$

Table 4 Sufficiency Research Questions

Research Questions
1. What is the criterion for Sufficiency?
2. What scale does Sufficiency use?

Table 5 Risk Research Question

Research Question
1. How do you rate Risk?

4.5. Examples of AI User Trust

As seen in Figure 1 AI User Trust Scenario, where a user *u* interacts with an AI system *s* within context *a*, the user's trust in the system can be represented as *T(u, s, a)*. Consider two AI scenarios.

First, a medical doctor (*u*), a medical diagnostic system (*s*), in a critical care facility (*a*) (in Figure 3 Medical AI User Trust Scenario)

a

Figure 3 Medical AI User Trust Scenario

Second, a college student (*u*), a music suggestion system (*s*), on a college campus. (*a*) (in Figure 4 Music Selection AI User Trust Scenario).

a

Figure 5 Music Selection AI User Trust Scenario

4.5.1. AI Medical Diagnosis

4.5.1.1. Medical AI User Trust Potential

The AI Medical User Trust Scenario is a high risk context (*a*) as the AI system (*s*) is making a medical diagnosis in a critical care unit. A medical doctor is the recipient of this diagnosis and is in a highly specialized field (*u*). The doctor would like to have a highly accurate diagnosis given the high-risk setting. Factors in the *User Trust Potential* for the medical doctor can summarized as follows:

Table 6 Medical AI System Scenario User Trust Potential

Attribute	Value
Personality	Caring (Risk Averse)
Cultural	Western
Age	56
Gender	Female
Technical Competence	Low
AI Experience	High

4.5.1.2. Perceived Pertinence of the Medical AI System Trustworthiness Characteristics

Table 7 Perceived Pertinence of Medical AI Trustworthy Characteristics

Trustworthy Characteristic	Perceived Pertinence (1-10)	Normalized Value
Accuracy	9	0.12
Reliability	9	0.12
Resiliency	9	0.12
Objectivity	3	0.07
Security	3	0.07
Explainability	10	0.15
Safety	10	0.15
Accountability	10	0.15
Privacy	2	0.03

As Table 6 Perceived Pertinence of Medical AI Trustworthy Characteristics indicates, the medical doctor considers *Explainability*, *Safety*, and *Accountability* as having the highest pertinence. These ratings are contextually appropriate given that the doctor will have to explain the AI's decision to the patient, in a high-risk environment, with the doctor having to take on full responsibility, respectively.

The "Normalized Value" column shows how the characteristics measured on different scales are transformed to a percentage of importance. This is demonstrated below using *Accuracy* as an example, based on Equation 4 Normalization of the Perceived Pertinence Value of a Trustworthy Characteristic:

Equation 5 Perceived Pertinence of Accuracy for the Medical AI Scenario

$$0.1238 = \frac{9}{65}$$

Accuracy accounts for roughly 12% of *Perceived Technical Trustworthiness*. The chart below further illustrates how the doctor has weighted each characteristic's pertinence to the scenario:

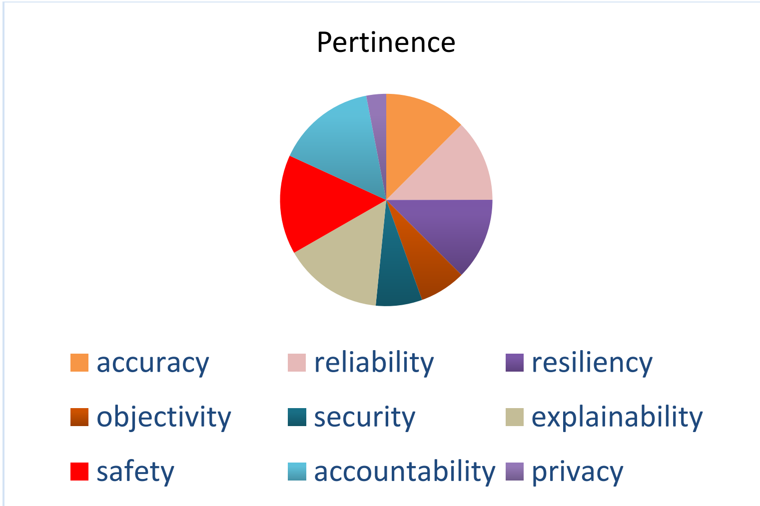

Chart 1 Perceived Pertinence for the Medical AI System Trustworthy Characteristics

4.5.1.3. Perceived Sufficiency of a Medical AI System Trustworthiness Characteristics

Each trustworthiness characteristic has a sufficiency value indicating the extent to which its measured value is good enough based on context and risk. These values will be measured with standards and guidelines that are being developed by AI System Trustworthiness groups at NIST.

Here, the risk in the context, r_a, rated on a scale of 1 (low risk) to 10 (high risk), is 10:

$$0.090 = \frac{90\%}{10}$$

Based on Equation 5 The Perceived Sufficiency of an AI Trustworthy Characteristic, the sufficiency value for *Accuracy* is 0.090.

Table 8 Perceived Sufficiency of Medical AI Trustworthy Characteristics' values

Trustworthy Characteristic	Characteristic Value (m_c)	Sufficiency Value (s_c)
Accuracy	90%	0.090
Reliability	95%	0.095
Resiliency	85%	0.085
Objectivity	100%	0.100
Security	99%	0.099
Explainability	75%	0.075
Safety	85%	0.085
Accountability	0%	0.000
Privacy	80%	0.080

4.5.2. AI Musical Selection Scenario

4.5.2.1. Music Selection AI User Trust

The AI Music Selection User Trust Scenario is a low risk context (*a*) as the AI system (*s*) is deciding what music the college student may like in a campus setting. The student is the recipient of the music and may have specific musical tastes (*u*). Factors in the *User Trust Potential* for the student can be summarized as follows:

Table 9 Musical Selection AI System Scenario User Trust Potential

Attribute	Value
Personality	Adventurous
Cultural	Western
Age	26
Gender	Male
Technical Competence	High
AI Experience	Low

4.5.2.2. Perceived Pertinence of the Musical Selection AI System Trustworthiness Characteristics

Table 10 Perceived Pertinence of the Musical Selection AI System Trustworthiness Characteristics

Trustworthy Characteristic	Perceived Pertinence (1-10)	Normalized Value
Accuracy	9	0.205
Reliability	9	0.205
Resiliency	9	0.205
Objectivity	3	0.068
Security	3	0.068
Explainability	2	0.045
Safety	2	0.045
Accountability	2	0.045
Privacy	5	0.114

As Table 9 Perceived Pertinence of the Musical Selection AI System Trustworthiness Characteristics shows, the student considers *Accuracy*, *Reliability*, and *Resiliency* as having the highest pertinence. These ratings are contextually appropriate given that the student would like to listen only to music he likes, whenever he wants to, and to have the system adapt when a selection is rejected.

The "Normalized Value" column shows how the characteristics measured on different scales are transformed to a percentage of importance. This is demonstrated below using *Accuracy* as an example, based on Equation 4 Normalization of the Perceived Pertinence Value of a Trustworthy Characteristic:

Equation 6 Perceived Pertinence of Accuracy for the Music Selection Scenario

$$0.205 = \frac{9}{44}$$

Accuracy accounts for roughly 21% of Perceived Technical Trustworthiness. The chart below indicates how the student has weighted each characteristic's pertinence to the scenario:

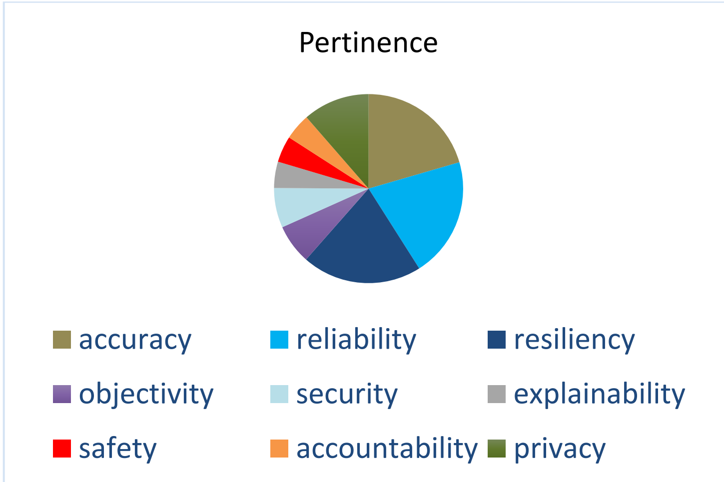

Chart 2 Perceived Pertinence of Music Selection AI Trustworthy Characteristics

4.5.2.3. Perceived Sufficiency of a Musical Selection AI System Trustworthiness Characteristics

Each trustworthiness characteristic has a sufficiency value indicating the extent to which its measured value is good enough based on context and risk. These values will be measured with standards and guidelines that are being developed by AI System Trustworthiness groups at NIST.

Table 11 Perceived Sufficiency of Medical AI Trustworthy Characteristics' values

Trustworthy Characteristic	Characteristic Value (m_c)	Sufficiency Value (s_c)
Accuracy	90%	0.450
Reliability	95%	0.475
Resiliency	85%	0.425
Objectivity	0%	0.000
Security	30%	0.150
Explainability	2%	0.010
Safety	5%	0.025
Accountability	0%	0.000
Privacy	0%	0.000

Here, the risk in the context, r_a, rated on a scale of 1 (low risk) to 10 (high risk), is 2:

$$0.450 = \frac{90\%}{2}$$

Based on Equation 5 The Perceived Sufficiency of an AI Trustworthy Characteristic, the sufficiency value for *Accuracy* is 0.450.

Table 12 Perceived Accuracy Trustworthiness

	Perceived Accuracy Pertinence (p_c)	Accuracy Value	Perceived Sufficiency (s_c)	$p_c * s_c$
Medical Scenario	0.120	90%	0.090	0.011
Musical Selection Scenario	0.205	90%	0.450	0.092

As Table 11 Perceived Accuracy Trustworthiness indicates, although *Accuracy* has the same value in both scenarios, the effect of risk is much higher in the medical scenario. Giving an incorrect diagnosis is more consequential than recommending the wrong song. Lower risk lends to greater perceived sufficiency of the 90% *Accuracy* value in the music scenario. Greater pertinence in the music scenario means that this perceived sufficiency will contribute more to *Perceived Technical Trustworthiness*.

5. Summary

Trust is one of the defining attributes of being human. It allows us to make decisions based on the information our limited senses can perceive. Should I give that person my phone number? Should I let that car drive me to my destination? It is trust that allows us to live our lives.

Technology continues to pervade many aspects of our professional and personal lives. Moreover, systems are becoming more complex. Trust, a complexity-reduction mechanism, will become even more important the less we know about our technology. It is because of this increasing technological complexity that we must look to the user's perspective if we are to understand trust in AI.

Trust in AI will depend on how the human user perceives the system. This paper is meant to complement the work being done on AI system trustworthiness. If the AI system has a high level of technical trustworthiness, and the values of the trustworthiness characteristics are perceived to be good enough for the context of use, and especially the risk inherent in that context, then the likelihood of AI user trust increases. It is this trust, based on user perceptions, that will be necessary of any human-AI collaboration.

There are many challenges to be faced with the approach in this paper. Starting with those in Table 12 AI User Trust Research Questions, more challenges will arise as we delve deeper into what enables a person to trust AI. Like any other human cognitive process, trust is complex and highly contextual, but by researching these trust factors we stand to enable use and acceptance of this promising technology by large parts of the population.

Table 13 AI User Trust Research Questions

Research Questions
User Trust Potential
1. What are the set of attributes that define User Trust Potential?
UX Influences on User Trust
2. What User Experience Metrics Influence User Trust? 3. How do User Experience Metrics Influence User Trust?
Pertinence
4. What should the measurement be for Pertinence
Sufficiency
5. What is the criterion for Sufficiency?
6. What scale does Sufficiency use?
Risk
7. How do you rate Risk?

6. Works Cited

[1] N. Luhmann, "Defining the Problem: Social Complexity," in *Trust and Power*, John Wiley & Sons, 1979, pp. 5 - 11.

[2] S. Kaya, "Outgroup Predjudice from an Evolutionary Perspective: Survey Evidence from Europe," *Journal of International and Global Studies,* 2015.

[3] S. E. Taylor, The Tending Instinct: How nuturing is essential to who we are and how we live, NY: Harry Holt & Company LLC, 2002.

[4] Macy and Skvoretz, 1998.

[5] K. Sedar, 2015.

[6] Axelrod and Hamilton, *Journal of International and Global Studfies,* 1981.

[7] D. Kahneman, "A Bias to Believe and Confirm," in *Think Fast and Slow*, New York, Farrar, Straus and Giroux, 2011, pp. 80-85.

[8] R. Mayo, "Cognition is a Matter Of Trust: Distrust Tunes Cognitive Processes," *European Review of Social Psychology,* pp. 283-327, 2015.

[9] R. Mayo, D. Alfasi and N. Schwarz, "Distrust and the positive test heuristic: Dispositional and situated social distrust improves performance on the Wason Rule Discovery Task," *Journal of Experimental Psychology: General,* vol. 143, no. 3, pp. 985 - 990, 2014.

[10] Y. Schul, R. Mayo and E. Burnstein, "Encoding Under Trust and Distrust: The Spontaneous Activation of Incongruent Cognition," *Journal of Personality and Social Psychology,* 2004.

[11] C. A. Hill and E. A. O'Hara, "A Cognitive Theory of Trust," *Washinngton University Law Review,* pp. 1717-1796, 2006.

[12] M. G. Haselton, D. Nettle and D. R. Murrary, "The evolution of cognitive bias," *Than Handbook of Evolutionary Psychology,* pp. 1-20, 2015.

[13] N. Luhmann, "Trust and Distrust," in *Trust and Power*, John Wiley & Sons, 1979, pp. 79-85.

[14] R. J. Lewicki, D. J. McAllister and R. J. Bies, "Trust and Distrust: New Relationships and Realities," *Academy of Management Review,* pp. 438-458, 1998.

[15] D. Gambetta, "Can we Trust Trust," in *Trust: Making and Breaking Cooperative Relations*, 2000, pp. 213-237.

[16] R. C. Mayer, J. H. Davis and F. D. Schoorman, "An Integrative Model of Organizational Trust," *Acamdemy of Management Review,* pp. 709-734, 1995.

[17] J. B. Rotter, "Interpersonal trust, trustworthiness, and gullibility," *American Psychologist,* vol. 35, no. 1, pp. 1 - 7, 1980.

[18] J. K. Rempel, J. G. Holmes and M. P. Zanna, "Trust in close relationships.," *Journal of Personality and Social Psychology,* vol. 49, no. 1, pp. 95-112, 1985.

[19] L. Becker, "Trust as noncognitive security about motives," *Ethics,* vol. 107, no. 1, pp. 43-61, 1996.

[20] B. Reeves and C. I. Nass, The media equation: How people treat computers, television, and new media like real people and places, Cambridge University Press, 1996.

[21] C. Nass, J. Steuer and E. R. Tauber, "Computers are social actors," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1994.

[22] C. Nass, Y. Moon and N. Green, "Are machines gender neutral? Gender-stereotypic responses to computers with voices.," *Journal of Applied Social Psychology,* vol. 27, no. 10, pp. 864 - 876, 1997.

[23] Y. Moon, "Intimate exchanges: Using computers to elicit self-disclosure from consumers," *Journal of Consumer Research,* vol. 26, no. 4, pp. 323 - 339, 2000.

[24] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers," *Journal of Social Issues,* vol. 56, no. 1, pp. 81 - 103, 2000.

[25] B. Muir, "Trust in Automation: Part 1. Theoretical Issues in the study of trust and human intervention in automated systems," *Ergonomics,* vol. 37, no. 11, pp. 1905-1922, 1994.

[26] B. M. Muir and N. Moray, "Trust in Automation Part II. Experimental studies of trust and human intervention in a process control simulation," *Ergonomics,* vol. 39, no. 3, pp. 429-460, 1996.

[27] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors,* vol. 57, no. 3, pp. 407-434, 2006.

[28] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors,* vol. 46, no. 1, pp. 50-80, 2004.

[29] P. Madhavan and D. A. Wiegmann, "Simularities and differences between human-human and human-automation trust: an integrative review.," *Theoretical Issues in Ergonomics Science,* vol. 8, no. 4, pp. 277-301, 2007.

[30] E. J. De Visser, S. S. Monfort, R. McKendrick, M. A. Smith, P. E. McKnight, F. Krueger and R. Parasuraman, "Almost human: Anthropomorphism increases trust resilience in cognitive agents," *Journal of Experimental Psychology: Applied,* vol. 22, no. 3, pp. 331 - 349, 2016.

[31] "Bill Gates: Trustworthy Computing," 17 Janurary 2002. [Online]. Available: https://www.wired.com/2002/01/bill-gates-trustworthy-computing/.

[32] WIRED, 2002. [Online]. Available: Https://www.wired,com/2002/bill-gates-trustworthy-computing/. [Accessed August 2019].

[33] IEEE, *982.1-2005 Standard Dictionary of Measures of the Software Aspects of Dependability,* IEEE, 2005.

[34] ISO/IEC/IEEE, *15206-1:2019 Systems and software engineering - systems and software assurance, Part 1: Concepts and vocabulary,* ISO/IEC/IEEE, 2019.

[35] B. M. Muir, "Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems," *Ergonomics,* vol. 37, no. 11, pp. 1905 - 1922, 1994.

[36] National Institute of Standards and Technology, "US LEADERSHIP IN AI: A Plan for Federal Engagement in Developing Technical Syandards and Related Tools, Prepared in response to Excutive Order 13859," 2019.

[37] P. L. McDermott and R. N. ten Brink, "Practical Guidance for Evaluating Calibrated Trust," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Los Angeles, CA, 2019.

[38] T. Tullis and B. Albert, Measuring the User Experience 2nd Edition, Waltham: Morgan Kaufmann, 2013.

[39] B. Skyrms, "Trust, Risk, and the Social Contract," *Synthese,* pp. 21-25, 2008.

[40] Y. Schul and E. Burnstein, "Encoding under trust and distrust: the spontaneous activation of incongruent cognitions," *Journal of Personality and Social Psychology,* 2004.

[41] J. Sauro and J. Lewis, Quantifying the User Experience, Cambridge: Morgan Kaufmann, 2016.

[42] J. Sauro and E. Kindlund, "Amethod to standardize usability metrics into a single score," in *Proceedings of CHI 2005*, Portland, 2005.

[43] M. Mohtashemi and L. Mui, "Evolution of indirect reciprocity by social information: the role of trust and reputation in evolution of altruism," *Journal of Theorical Biology,* pp. 523-531, 2003.

[44] M. W. Macy and J. Skvoretz, "The evolution of Trust and Cooperation Between Strangers: A Computational Model," *American Sociological Review,* pp. 638-660, 1998.

[45] B. Barber, The Logic and Limits of Trust, Rutgers University Press, 1983.

[46] R. Axelrod and W. D. Hamilton, "The Evolution of Cooperation," *Science,* pp. 1390-1396, 1981.

[47] *ISO/IEC/IEEEE 15026-1:2019,* 2019.

[48] International Organization for Standardization TC/ 159/ SC 4, *ISO 9241-11:2018 Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts,* Geneva: International Organization for Standardization, 2018.

[49] S. Kaya, "Outgroup Prejudice from an Evolutionary Perspective: Survey Evidence from Europe," *Journal of International and Global Studies,* 2015.

Appendix A AI User Trust Equations