

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>

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28

29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54

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



55
56
57
58
59
60
61
62
63
64
65
66

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*

67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

**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

96 **Abstract**

97

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

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

118

119

120 **Key words**

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

123

124 **Table of Contents**

125 **1. Introduction** 1

126 **2. Trust is a Human Trait** 1

127 2.1. Purpose of Trust 1

128 2.2. Distrust & Cognition 2

129 2.3. Trust, Distrust, and Cooperation: The Role They Play 3

130 2.3.1. Factors that lead to Trusting and Distrusting 4

131 **3. Trust in Automation** 5

132 3.1. Computers as Social Actors 5

133 3.2. Human Factors, Trust and Automation 5

134 **4. Trust in Artificial Intelligence** 6

135 4.1. AI Trustworthiness 6

136 4.2. User Trust in AI 7

137 4.3. User Trust Potential 8

138 4.4. Perceived System Trustworthiness 8

139 4.4.1. User Experience 9

140 4.4.2. Perceived Technical Trustworthiness 10

141 4.5. Examples of AI User Trust 12

142 13

143 4.5.1. AI Medical Diagnosis 13

144 4.5.2. AI Musical Selection Scenario 16

145 **5. Summary** 19

146 **6. Works Cited** 20

147 **Appendix A** 23

148

149 **List of Tables**

150 Table 1 User Trust Potential Research Question 8

151 Table 2 User Experience Research Question 10

152 Table 3 Pertinence Research Question 12

153 Table 4 Sufficiency Research Questions 12

154 Table 5 Risk Research Question 12

155 Table 6 Medical AI System Scenario User Trust Potential 13

156 Table 7 Perceived Pertinence of Medical AI Trustworthy Characteristics 14

157 Table 8 Perceived Sufficiency of Medical AI Trustworthy Characteristics' values 15

158 Table 9 Musical Selection AI System Scenario User Trust Potential 16

159 Table 10 Perceived Pertinence of the Musical Selection AI System Trustworthiness Characteristics 17

160 Table 11 Perceived Sufficiency of Medical AI Trustworthy Characteristics' values 18

161 Table 12 Perceived Accuracy Trustworthiness 19

162 Table 13 AI User Trust Research Questions20
 163 Table 14 Definitions **Error! Bookmark not defined.**
 164

165 **List of Figures**

166 Figure 1 AI User Trust Scenario.....7
 167 Figure 2 the User Experience Front End and the AI System Trustworthy Characteristics Backend.....9
 168 Figure 3 Medical AI User Trust Scenario.....13
 169 Figure 4 Music Selection AI User Trust Scenario13
 170

171 **List of Equations**

172 Equation 2 Perceived System Technical Trustworthiness10
 173 Equation 3 The Relationship of Perceived Pertinence and Perceived Sufficiency of the Trustworthy
 174 Characteristic.....10
 175 Equation 4 Normalization of the Perceived Pertinence Value of a Trustworthy Characteristic11
 176 Equation 5 The Perceived Sufficiency of an AI Trustworthy Characteristic.....12
 177 Equation 6 Perceived Pertinence of Accuracy for the Medical AI Scenario14
 178 Equation 7 Perceived Pertinence of Accuracy for the Music Selection Scenario.....17
 179

180

181

182 **1. Introduction**

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

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

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

207

208

209 **2. Trust is a Human Trait**

210 **2.1. Purpose of Trust**

211

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

216

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

220

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

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

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

235 **2.2. Distrust & Cognition**

236

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

246 For instance, this has been demonstrated in the Wason Rule Discovery Task, where
247 participants complete the following two steps after being shown the number sequence “2,
248 4, 6”: 1) generate a hypothesized rule characterizing the number sequence and 2) generate
249 several number sequences to test their hypothesized rule. In general, most individuals
250 hypothesize the rule “+2” and generate only sequences that follow their rule for the second
251 step (positive hypothesis tests). This underscores our tendency toward congruent
252 processing, which, in this case, often leads to a failure to discover the true rule (i.e., “any
253 series of increasing numbers”). Experiments showed that individuals low in dispositional
254 trust and those primed with distrust were found to be significantly more likely to generate
255 sequences that did not follow their rule (negative hypothesis tests) [9]. Distrust improved
256 performance on the task by invoking a consideration of alternatives. Similarly, a state of
257 distrust has been found to lead to faster responses to incongruent concepts and a greater
258 number of incongruent free associations [10].

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

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

264 *searches for dissimilarities in an attempt not to be influenced by an*
265 *untrustworthy environment.”*

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

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

275 **2.3. Trust, Distrust, and Cooperation: The Role They Play**

276
277 Trust and distrust are so fundamental that they are often concealed within the most
278 mundane decisions in our daily lives. Without some trust we would not leave our homes
279 due to overwhelming fear of others. Meanwhile, distrust permits us to navigate a world of
280 potentially deceitful actors and misinformation.

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

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

291 *“If we were blessed with an unlimited computational ability to map out*
292 *all possible contingencies in enforceable contracts, trust would not be a*
293 *problem”.*

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

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

306

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?

352 **3. Trust in Automation**

353 **3.1. Computers as Social Actors**

354
355 The Computers as Social Actors (CASA) paradigm lends support to the viability of human-
356 machine trust as a construct. CASA has been used by communication researchers to
357 demonstrate that humans respond socially to computers [20]. In a CASA experiment, a
358 computer replaces one of the humans in the social phenomenon under investigation to see
359 if the social response by the human holds [21]. This method has revealed that people use
360 politeness [21], gender stereotypes [22], and principles of reciprocal disclosure [23] with
361 computers. Notably, the original CASA experiments were conducted with experienced
362 computer users interacting with simple, text-based interfaces [24].

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

367
368 **3.2. Human Factors, Trust and Automation**

369
370 Human factors researchers began studying trust in response to the increasing prevalence of
371 automation in work systems. Muir [25] was one of the first to challenge the notion that
372 behavior toward automation was based solely on its technical properties. Her view evokes
373 a theme of our preceding discussion of trust between people—an operator simply cannot
374 have complete knowledge of an automated system. The trustor’s (operator’s) perceptions
375 become important because of the trustee’s (automation’s) freedom to act, and the trustor’s
376 inability to account for all possibilities of the trustee’s action.

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

380 *“The source of this disparity must lie in the individuals themselves, in*
381 *something they bring to the situation.”*

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

386 Since this early work, researchers have contributed a significant amount of
387 understanding of relevant factors in trust in technology. Lee and See’s [28] review
388 emphasizes how the increasing complexity of automated systems necessitates an
389 understanding of trust. Hoff and Bashir [27] reviewed the empirical work that followed
390 Lee and See’s [28] and defined three sources of variability in trust in automation:
391 dispositional, situational, and learned. Dispositional factors include the age, culture, and
392 personality of the trustor (i.e., the automation operator or user) among other characteristics.
393 Situational factors concern the context of the human-automation interaction and various
394 aspects of the task, such as workload and risk. Learned trust is a result of system
395 performance characteristics as well as design features that color how performance is

396 interpreted. This three-layer model is compatible with Mayer et al.'s [16] human-human
397 model, which considers trustor characteristics (dispositional), perceived risk (situational),
398 and perceived trustworthiness that is dynamically updated by observing trustee behavior
399 (learned). As previously discussed with respect to Mayer et al.'s model, these human-
400 automation trust factors inform our later discussion of AI user trust.

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

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

418

419 **4. Trust in Artificial Intelligence**

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

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

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

431

432 **4.1. AI Trustworthiness**

433

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

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

438 *their information. Trustworthy Computing is computing that is as*
439 *available, reliable and secure...". [32] [33] [34]*

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

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

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

449

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

458

459 4.2. User Trust in AI

460

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

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

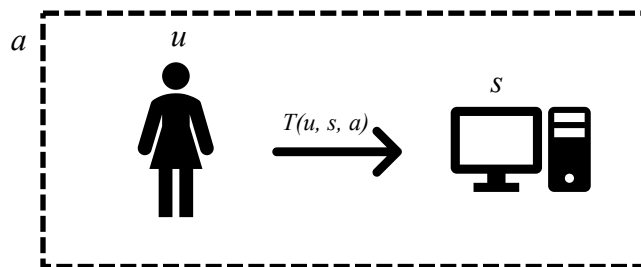


Figure 1 AI User Trust Scenario

468 The research on human-human and human-automation trust suggest two main
469 sources of variability in trust in an AI system: the user and the system. Therefore, we
470 conceptualize user trust in AI in terms of two main components: *User Trust Potential*,

471 UTP(u), and *Perceived System Trustworthiness*, $PST(u, s, a)$ ¹. User trust can be expressed
472 as a function f of these two components:

473

$$474 \quad T(u, s, a) = f(UTP(u), PST(u, s, a))$$

475

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

481

$$482 \quad T(u, s, a) = UTP(u) * PST(u, s, a)$$

483

484 We carry this illustrative probabilistic assumption through the remainder of our
485 discussion and examples but emphasize the contextual nature of perceived trustworthiness
486 and trust. Trust is based on the trustee's (system's) expected behavior and should not be
487 interpreted literally as a 'chance' decision. The probabilistic representation allows us to
488 quantitatively express differences in trust due to various factors².

489

490 **4.3. User Trust Potential**

491

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

498

499 Table 1 User Trust Potential Research Question

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

500

501 **4.4. Perceived System Trustworthiness**

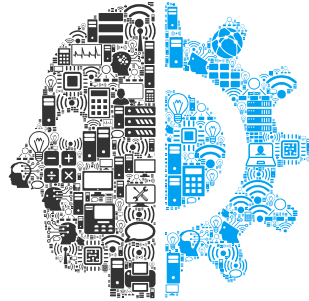
502

503 What we refer to as *Perceived System Trustworthiness*, $PST(u, s, a)$, is made up of a
504 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.

505 (PTT) of the AI system. These two components can be thought of as front end-related
506 (UX) and back end-related (PTT) factors in the user u 's trust of the AI system s in context
507 a .
508
509



510
511 Figure 2 the User Experience Front End and the AI System Trustworthy Characteristics
512 Backend

513 We first represent Perceived System Trustworthiness as a generalized function g of
514 UX and PTT:
515

$$516 \quad PST(u, s, a) = g(UX, PTT)$$

517
518 For illustrative purposes, this may be thought of as a multiplicative function of
519 independent probabilities:

520 Perceived AI System Trustworthiness

$$522 \quad PST(u, s, a) = UX * PTT$$

521
523 Thus, as with overall trust T , PST will lie in the range $[0, 1]$ and represent the degree
524 to which the system is perceived as trustworthy. Further research is needed to identify the
525 relationship between UX and PTT.

526 527 **4.4.1. User Experience** 528

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

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

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

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

547

548 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?

549

550

551 4.4.2. Perceived Technical Trustworthiness

552

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

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

564

565

Equation 1 Perceived System Technical Trustworthiness

566

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

567

568

569

570

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*:

571

572

573

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

574

$$ptt_c = p_c * s_c$$

575

576

577

578

579

580

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.”

581

582

583

584

585

586

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.

587

588

589

4.4.2.1. Pertinence

590

591

592

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.

593

594

595

596

597

598

599

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.

600

601

602

603

604

605

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.

606

607

608

609

610

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:

611

612

613

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

614

615

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

616

617 Table 3 Pertinence Research Question

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

618

619 4.4.2.2. Sufficiency

620

621 *Sufficiency* is the answer to the question, “How good is the value of this characteristic for
622 this context?” Sufficiency involves the user’s consideration of each characteristic’s
623 measured value and a judgement of how suitable that value is with respect to contextual
624 risk.

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

632

633 Equation 4 The Perceived Sufficiency of an AI Trustworthy Characteristic

634

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

635

636 Table 4 Sufficiency Research Questions

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

637

638 Table 5 Risk Research Question

Research Question
1. How do you rate Risk?

639

640 4.5. Examples of AI User Trust

641

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

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



Figure 3 Medical AI User Trust Scenario

647
648

649

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



Figure 5 Music Selection AI User Trust Scenario

652
653
654
655

4.5.1. AI Medical Diagnosis

657

4.5.1.1. Medical AI User Trust Potential

659

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

665

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

666
667

668 **4.5.1.2. Perceived Pertinence of the Medical AI System Trustworthiness**
 669 **Characteristics**

670
 671

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

672

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

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

682

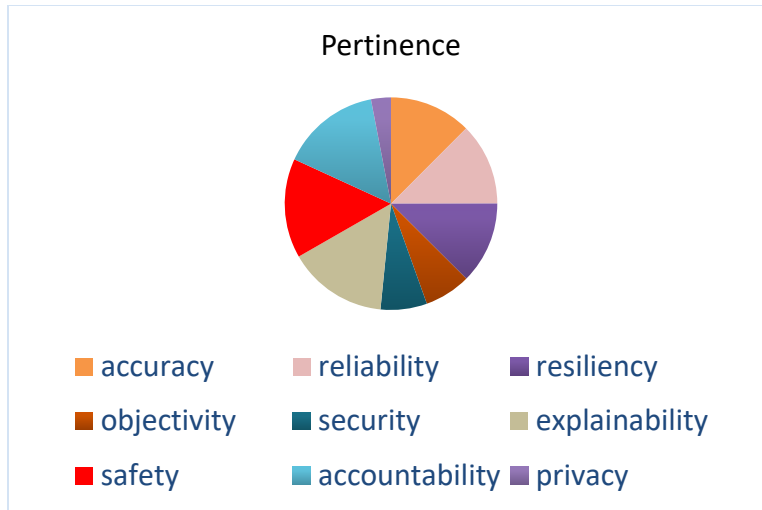
683 Equation 5 Perceived Pertinence of Accuracy for the Medical AI Scenario

684
$$0.1238 = \frac{9}{65}$$

685

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

689



690
691

Chart 1 Perceived Pertinence for the Medical AI System Trustworthy Characteristics

692

4.5.1.3. Perceived Sufficiency of a Medical AI System Trustworthiness Characteristics

693
694
695

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.

696

697

698

699

700

701

702

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

703

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

704

705

706

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

707

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

708

709 **4.5.2. AI Musical Selection Scenario**

710

711 **4.5.2.1. Music Selection AI User Trust**

712

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

717

718

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

719

720

721
722
723
724
725
726

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

727

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

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

737

738 Equation 6 Perceived Pertinence of Accuracy for the Music Selection Scenario

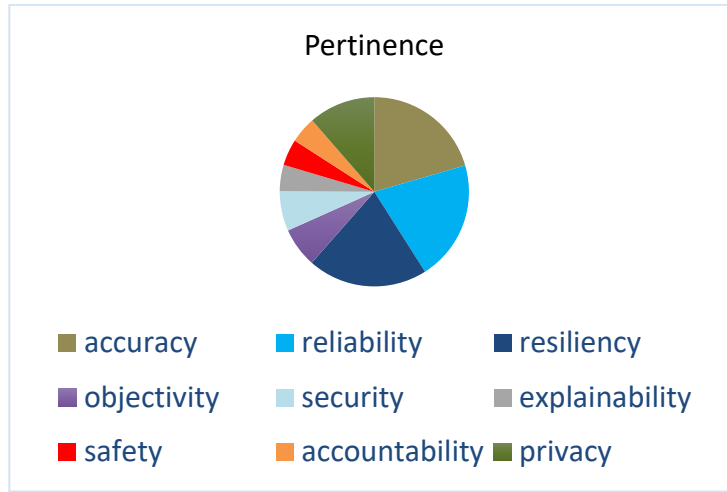
739

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

740

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

744



745

746

Chart 2 Perceived Pertinence of Music Selection AI Trustworthy Characteristics

747

748

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

749

750

751

752

753

754

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.

755

756

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

757

758

759

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

760

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

761

762

763

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

764
765

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

766
767
768
769
770
771
772

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*.

773
774

5. Summary

775
776
777
778

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.

779
780
781
782
783

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.

784
785
786
787
788
789

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.

790
791
792
793
794
795

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.

796
797
798

799 Table 13 AI User Trust Research Questions
 800

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?

801

802 **6. Works Cited**

803

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

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

[3] S. E. Taylor, *The Tending Instinct: How nurturing 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 Studies*, 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," *Washington University Law Review*, pp. 1717-1796, 2006.

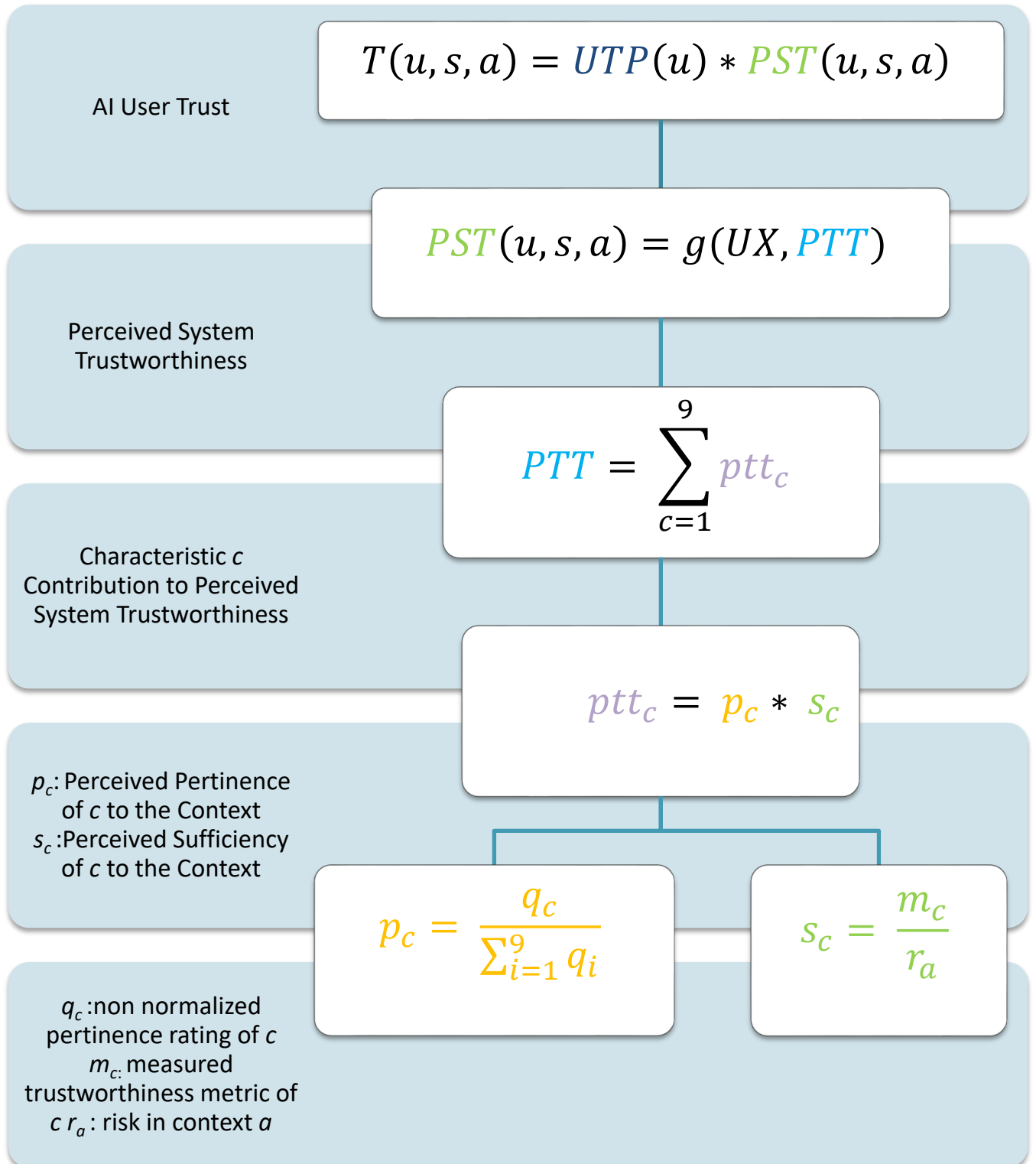
[12] M. G. Haselton, D. Nettle and D. R. Murrar, "The evolution of cognitive bias," *Than Handbook of Evolutionary Psychology*, pp. 1-20, 2015.

- [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," *Academy 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, "Similarities 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 January 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 Standards and Related Tools, Prepared in response to Executive 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, "A method 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 Theoretical 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/IEEE *15206-1:2019*, 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.

804
805
806

Appendix A AI User Trust Equations



807