Towards a Framework for Evaluating Trust Based on User Confidence in Outcomes in the Human-AI Collaboration Setting

JOSHUA BOLEY, KATELYN KOZINSKI, and MAOYUAN SUN, Northern Illinois University, USA

Over the past decades, a great deal of effort has been spent to establish the role of trust in human-AI interactions, more recently with regard to sociotechnical systems where ethical considerations such as fairness and safety have attracted intense and often unfavorable scrutiny to the far-ranging impact of AI on public welfare. A variety of theories and frameworks have emerged, each offering valuable insights, either approaching from a general, high-level perspective, attempting to quantitatively model factors influencing trust relationships in specific types of scenarios, or empirically investigating characteristics of trust in real users. In this work, we briefly survey a few prominent works on trust and trustworthiness and consider how they contribute to a reformulation of trust-building as a process that accommodates many kinds of factors. We then begin laying the groundwork for a framework to guide empirical assessment of user objectives and reinforcing trust through confidence in a way that extends to the context of human-AI collaboration.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Machine learning**; • **Human-centered computing** → **HCI theory, concepts and models**; • **Applied computing** → **Sociology**; **Psychology**.

Additional Key Words and Phrases: Human-AI collaboration, HAI, trust, human factors, trust calibration, confidence

ACM Reference Format:
Joshua Boley, Katelyn Kozinski, and Maoyuan Sun. 2022. Towards a Framework for Evaluating Trust Based on User Confidence in Outcomes in the Human-AI Collaboration Setting. 1, 1 (April 2022), 8 pages.

1 INTRODUCTION

The power of black-box algorithms such as deep neural networks (DNNs) have transformed the sociotechnical domain. These powerful algorithms bring enormous parametric spaces to bear that capture huge ranges of subtle relationships between inputs and allow realistically robust, complex decision-making processes in a production setting [1]. For some time, artificial intelligence (AI) has been a driving force in the design of flexible, adaptive business systems and has been a crucial component of mainstream business strategies [13]. Modernized high-traffic and high-demand computational ecosystems have increasingly relied on intelligent agents to aid human decision makers, including those in domains that have high impact and significant consequences on society [30]. Indeed, intelligent systems have become prevalent in many critical applications—including law enforcement, finance, and government assistance programs—with far-reaching consequences to both individuals and groups [6, 8, 26].

Despite the astonishing pace of technical advances and proliferation of AI-driven sociotechnical systems in the modern world, the inscrutability of black-box algorithms not only remains a barrier to full acceptance, but also, in the worst case, may pose unacceptable levels of risk in critical applications. An unfortunate litany of examples such as

Authors' address: Joshua Boley, jboley2@niu.edu; Katelyn Kozinski, z167824@students.niu.edu; Maoyuan Sun, smaoyuan@niu.edu, Northern Illinois University, 1425 W Lincoln Hwy, DeKalb, Illinois, USA, 60115.

Manuscript submitted to ACM

early release of dangerous criminals back into the public by bail recommender systems stand as evidence of the dangers of misplaced trust when the stakes are high [27]. In recognition of the liabilities inherent to automated black-box decision-making and generally poor domain- and instance-specific understanding of how outcomes are generated, researchers have invested considerable time and effort into developing Explainable AI (XAI) techniques to place a battery of tools into the hands of developers and practitioners that help elucidate the mechanics and behaviors of AI models [9]. These solutions have the potential to provide invaluable insights that, among many benefits, may allow us to infer if, when, and how decisions are impartial, fair, or even *safe*.

However, while the tools supporting clearer understanding of how and why a model generates an outcome or generally behaves as it does are certainly available, our understanding of how to most effectively employ them in building appropriate trust relationships is still in its relative infancy. On the one hand is the human; each individual presents their own unique perspective, has a given level of familiarity with the application, and is likely to color their interpretations of both the model and any explanation given according to both their expertise and their personal and social contexts [10]. On the other is the nature and application domain of the system itself, constraining both the targeted problem domain and the means and opportunities for a user to interact with it (the machine) over the duration of its use. Further complicating the matter is the fact that many problems require the reasoning and input of the human to find adequate (not necessarily optimal) solutions, a collaborative process that can be referred to as a "human-in-the-loop" arrangement [4, 18].

As of the time of writing, and to the best of our knowledge, current trust definitions and frameworks are either too general to make direct recommendations as to which tools should be applied to a given application or do not generalize well to a human-AI collaboration setting. In this work, we set out to build a more grounded conceptualization of the relationship between trust-building and the specific outcomes that elicit trust. We begin in section 2 with an overview of some influential definitions of and frameworks for trust in the literature, briefly discussing what prevents them from generalizing well to cases of trust-building in human-AI collaboration. We then in section 3 propose a new conceptualization of trust, integrating multiple existing viewpoints, that we believe steers the discussion in a direction that benefits those cases. In section 4, we consider how our conceptualization may be further extended and conclude with a few final remarks.

2 BACKGROUND

2.1 Definitions and Conceptions of Trust

Several definitions of trust and related concepts have been proposed in the fields of HCI, sociology, and beyond. In publications concerning AI systems, human-human trust definitions are often adapted from the social sciences to support human-machine trust more directly [29]. Although trust is an important component of decision-making, trust as a psychological state, or *attitude*, is generally distinguished from trust-motivated *actions* [29], which can be affected by a multitude of factors beyond trust itself. A summary of concepts frequently encountered in the literature and their relationships can be found in Table 1.

According to Troshani et al. in [28], in the broadest sense, human trust in AI systems is elicited by two major factors: the "humanness" of the AI and its "intelligence". Troshani et al. further contend that human trust is an artifact of the user's perception of the AI's beneficence, effectiveness, and reliability. Others, such as Hoffman [16], Crockett et al. [7], and Jacovi et al. [19] additionally posit *risk* or user *vulnerability* to a potentially unfavorable outcome produced by an

Table 1. Trust-Related Concepts

Concept	Related	Meaning/Relevance
Trustworthiness		Related to characteristics of the AI that support user trust [19, 21]
	Perceived trustworthiness	User beliefs about the AI's trustworthiness [29]
Trust, distrust		User attitudes towards an AI; trust and distrust are situated either alongside or in opposition to each other [29]
	Risk/vulnerability	A component of trust implying potential negative consequences for the user [7, 16, 19, 29]
	Appropriate trust	A balance of trust and distrust suitable for some application [15, 21]
Contractual (trust)		A specific conceptualization of trust and distrust based on the trustee's adherence to one or more expected functionalities or contracts [15, 19]
	Warranted (trust)	Trust that is supported, or confirmed, by the outcomes generated by a trustworthy system [19]
	Unwarranted (trust)	Trust that is not supported by the trustworthiness of a system [19]
	Intrinsic trust	Warranted trust arising from the model's characteristics [19]
	Extrinsic trust	Warranted trust arising from the behavior of the model [19]
Reliance		User decision to rely on an AI; often contrasted with trust along anthropomorphic lines involving intent [15, 19, 29]

AI system as an essential component of trust. While these are extremely useful characterizations, they leave open the question of how risk, as a factor in a trust relationship, may be dimensionalized.

Related to trust is the notion of *distrust*. In [29], Vereschak et al. identify two main conceptions of distrust among researchers. In the first, distrust is situated in direct opposition to trust; high distrust, for example, indicates low trust. In the second, distrust exists alongside trust; high distrust, for instance, can coexist with high trust. Distrust should not be mistaken for low or absent trust; it is closer to trust in a negative outcome than a lack of belief in a positive outcome [19, 29].

In some definitions, trust is framed as *contractual* in nature [15, 19]. From Jacovi et al.'s perspective in [19], in the context of human interactions with AI systems, trust can be regarded as the user's expectation that the system will function in some anticipated manner (in other words, fulfill some contract) in the user's interests. One such contract might pertain to the fairness of a system's decisions. Another contract might relate to the accuracy or precision of a model. Conceptualizing trust in this way enables a multidimensional view of trust, in which an overall sense of trust arises from the combination of a user's trust expectations regarding multiple implicit or explicit contracts.

As an attitude, trust must be distinguished from *trustworthiness*, which concerns the characteristics of the trustee that provide a basis for user trust [19, 21]. In the context of contracts, trustworthiness can be measured by the system's ability to fulfill one or more of its associated contracts [19]. Since human trust is a complex phenomenon with both rational and affective (emotional) components [22], trust in—and distrust of—an AI system can be said to be *warranted* or *unwarranted* based on whether or not the user's trust expectations are based on the system's trustworthiness [19]. These notions of warranted and unwarranted trust are particularly relevant in discussions of *appropriate trust* in AI systems, where users' expectations match the relative trustworthiness of the system [15, 21]. In particular, warranted

trust can be considered *intrinsic* when the user understands what makes a system (or aspect of a system) trustworthy and *extrinsic* when the user's trust is based on a trustworthy evaluation of the system's outcomes [19].

Another concept related to trust is *reliance*, which Vereschak et al. describe as a decision to act on a recommendation made by a trustee such as an AI [29]. Hawley summarizes a common philosophical differentiation between the notions of reliance and trust, where reliance lacks an anthropomorphic sense of intent placed upon the AI by the user (wherein failure to behave as expected, or trusted, may result in a sense of betrayal) [15]. Because AI is often anthropomorphized in the real world, Jacovi et al. posit trust as the more relevant concept for this context [19].

While the contractual formalization of trust discussed by Jacovi et al. and others ([15, 19]) is appealing in a general sense, there are some difficulties in a purely contractual conceptualization of trust, particularly the causal nature of warranted trust, where Ferrario et al. observe in [11] that it is difficult to justify a specific level of trust if trust and trustworthiness have degrees. Additionally, the contractual formalization requires a rigid codification of specific points on which trust may be established (supported by documentation such as fairness checklists, factsheets, etc.), and trust in the system is only warranted (that is, ethical and useful) if the contract is maintained [19]. However, since documentation is key to the development and maintenance of trust, a purely contractual notion of trust cannot accommodate an evolving, dynamic, and experiential relationship that adjusts to the *actual* trustworthiness of the system [11, 21, 24]. We also contend that the contractual formalization misses a key, *longitudinal* aspect of what is essentially a long-term dynamic relationship between the user and the system, particularly in the human-AI collaborative context, which is bound to change as new weaknesses (or strengths) are discovered.

2.2 Trust as a Dynamic Process

Human trust is dynamic and evolving; simple intuition and observation readily inform us that this is so. User trust in automated—and more recently, intelligent—systems has, unsurprisingly, long been regarded as a dynamic process in the computer science and engineering literatures. Related to trust, an earlier account of human trust in computer systems by Fogg and Tseng in [12], introduces the concept of computer credibility and describes processes through which credibility is gained and lost. Besides noting a disproportionate tendency for users to penalize even relatively minor mistakes while grudgingly rewarding good performance in their personal assessment of credibility, they make the slightly counterintuitive observation that errors that are highly consistent in frequency, type and context may actually *increase* a system's credibility with its users. Later works, such as [3], touch upon the dynamic nature of user trust in the context of how it may be gained through developer initiative. More recently, Holliday et al. noted that user trust is especially dynamic in the presence of intelligent systems which modify their own behavior as new information is absorbed by its model, i.e., in online learning scenarios [17]. Beauxis-Aussalet et al. in [2] note not only the dynamic, changing trust relationship between a user and intelligent system but also that systems used in "high-stakes" scenarios additionally exacerbate the trust building process, as users tend to gain trust more slowly and lose it much more quickly than in "low-stakes" scenarios.

The concept of *calibration*, or experiential manipulation (presumably to the user's benefit) of the human-machine trust relationship is well-established. To our knowledge, the first use of the term arises in the work of Lee and See [21] in the context of automated systems, where well-calibrated trust yields an appropriate level of reliance on the system. Lee and See propose that poorly calibrated trust manifests primary as two opposing conditions: 1) *Overtrust*, where the user's expectations of the system exceed its actual capabilities, and 2) *Distrust*, where the user's expectations underestimate its actual capabilities. In [14], Han and Schultz present a more nuanced model of trust and its calibration.

Han and Schultz similarly characterize calibration (in the context of visual analytics systems) as refinement of the trust relationship, stating that "Trust calibration aligns the trust put into a VA system by the user with the system's actual trustworthiness". User trust in the system is conceptualized as falling within a continuum, increasing from complete distrust to complete trust in scalar fashion. Drawing upon Cho et al.'s survey in [5], Han and Schultz discuss four discrete levels, corresponding to contiguous ranges on the continuum from far left, where the user completely lacks trust, to far right, where the user has complete trust: Distrust, undistrust, untrust and trust. They additionally incorporate Marsh and Brigg's concepts of a *cooperation threshold* and a *limit of forgivability* [25] into their trust continuum, where the *cooperation threshold*, towards the trust end of the continuum, is the threshold at which (positive) trust is established well enough such that the user is willing to cooperatively engage with the system, and the *limit of forgivability*, towards the distrust end, is the point beyond which decreasing trust compels the user to regard the system antagonistically.

3 A CONFIDENCE-BASED TRUST CALIBRATION PROCESS

In [22], Lewis et al. make the observation that "trust is a process containing complex feedback loops". From this perspective, *we conceptualize trust as an artifact, or result of an ongoing, internal process that functions much like a stateful cost/benefit assessment performed by the user.* From the standpoint of the purposes for which the user engages a system, trust can furthermore be thought to align to the user's objectives and goals. If trust is reinforced by the AI's ability to meet the user's expectations, and if those expectations reflect the user's goals, then trust may be considered indirectly as a function of the user's objectives. A user may have multiple objectives, leading to both distinct and overlapping sets of expectations regarding each, depending on the complexity of the problem space and the sophistication of collaborative interactions. Our conceptualization roughly parallels Jacovi et al.'s discussion of multiple trust contracts in [19].

We also introduce the use of the term *confidence* to separate the long-term impacts of the outcomes of the use of a system, or the instances feeding back into a trust-generating process, from the long-term process that reinforces or degrades the overall trust relationship. This is similar to Luhmann's high-level notion of confidence as an internal sense of agency of the user that develops into the trust relationship when overt expectations—that the user's actions have meaningful impact on outcomes—are met [23]. We believe that positioning confidence as an intermediate factor—or the immediate temporal artifact generated by the process—is particularly salient in human-AI collaborative systems. Similar to Jacovi et al.'s notion of intrinsic and extrinsic trust [19], in our context of use, confidence can be defined as the user's assessment or expectation that the system 1) uses a model that has characteristics that are useful within the context of the immediate application and user objectives—to the extent of the user's expertise with and understanding of the AI; and 2) has behaved as anticipated and produced an acceptable outcome. However, recognizing that there is a distinction between contractual trustworthiness and *actual* trustworthiness [11], it is not restricted by the implied assertion that trust arising from sources external to some contract are unwarranted. Relying on the notion of a trust continuum as described in [14], we additionally postulate that changes in confidence through positively- and negatively-reinforcing outcomes *calibrates* the trust relationship, similar to Lee and See's usage in [21], along multiple dimensions respective to the user's goals and objectives. The term also carries a statistical connotation that we feel is convenient towards the eventual development of a useful formalism, and we also suggest that it might be possible to meaningfully model or approximate, perhaps quantitatively, a process that calibrates trust.

Most important to the long-term trust calibration process is change in the user's confidence over time. Any change in confidence as a result of an instance or outcome of use can be parameterized by the kind of high-level, contractual-based expectations discussed by Jacovi et al. [19]. In addition to the more legislatively-focused factors discussed in that

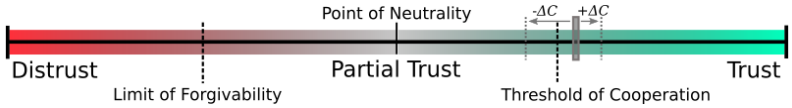

Fig. 1. Trust continuum, adapted from Han and Schultz [14] and Cho et al. [5]. User trust is calibrated based upon changes in confidence, adjusting either towards distrust or trust, based upon user perception that outcomes support expectations.

work, practical *in-situ* performance metrics relating to the reliability of the system also apply, such as overall accuracy, precision, and response; however, they impact the support of the user's objectives. It is also useful to assume that a user's confidence may not necessarily "start from zero". A variety of factors, including organizational or peer influence, and other historical factors whether resulting from prior experience or reputation of the developer, are likely to impart an initial sense of confidence, conditional as it may be. Assuming a rational user, these factors would tend to establish a starting point in the long-term trust relationship between the user and the system, particularly in the professional human-AI teaming context with which we are most interested.

In a more formal sense, we can suppose that the magnitude of a change in confidence is somehow related to the distance of the outcome from an *ideal outcome*, and furthermore introduce a bit of flexibility by assuming that there could be a range of acceptable outcomes "surrounding" the ideal; the farther outside of that range an outcome falls, the greater the *decrease* in confidence, while the closer to the ideal within that range, the greater the *increase*. Specifying a single "ideal" outcome is an extremely difficult problem, however, and as a user may not be fully cognizant or able to express what constitutes an "acceptable range" of outcomes, we propose that the problem might be better (and equivalently) considered in terms of *the strength of a user's belief that the outcomes positively support their objectives*. We further propose that the difference between such belief in a current interaction response and the prior response (or responses) calibrates the trust relationship: Decrease in belief between consecutive uses calibrates trust towards the left end of the trust continuum (distrust), while increase calibrates trust towards the right (trust), as illustrated in Figure 1.

Posing calibration of trust as an adjustment based upon strength of a user's belief that outcomes support objectives over time provides ample room to work in the notion of uncertainty; a user's belief in an outcome is naturally counterbalanced by their uncertainty in the same. In addition, the notion of strength of belief naturally lends itself to expression in *probabilistic* terms, i.e., 0 corresponding to the total absence of belief and 1 to the total belief in the proposition that an outcome supports a user objective. We can also suppose that outcomes map to a strength of belief for *each* user objective. (On a tangential note, we can also begin to relate our notion of belief with the uncertainty—and more importantly, communication of it—discussed by, e.g., Joslyn and LeClerk in [20], though we leave that discussion to further work).

4 SUMMARY AND OPEN QUESTIONS

Over the course of this brief overview and introduction, we have examined other works in the human-AI trust literature and used the best concepts to help springboard our own formulation of the trust process, which we hope will not only assist developers in framing their choice of XAI tools in a rigorous theoretical framework, but also extend it to accommodate the complex interactions and relationships in human-AI teaming scenarios. We incorporate insights from works such as Jacovi et al. [19] while focusing the discussion on the human factors, and lay the foundations for a framework that connects a set of parameters—potentially including "contractual" factors—to a process that, through the manipulation of confidence, calibrates the trust relationship along several, objective-related dimensions.

Our work is still in a very early state, and a great deal of thought, investigation, and (undoubtedly) reexamination remains. One among many questions that needs to be investigated is how useful the somewhat mechanistic formulation of calibration occurring along multiple objective-based dimensions would be in practice. In many respects, our ideas also assume an ideal user driven solely by rational decision-making objectives, and so do not suggest any obvious or simple means to account for organizational or cultural influences, at least not in a direct fashion. This is perhaps not necessarily a problem; after all, a developer would likely find a rationality-driven baseline profile for their intended users to be fairly helpful. However, user irrationality would ultimately result in the *miscalibration* of trust, which we do not currently address.

The purpose of building this framework is the support of empirical investigation, which naturally raises the question of whether or not the foundation we are laying will lead to a framework that either incorporates or suggests guidelines for identifying novel quantifiable measures for analysis. Of particular interest in this regard are the proposed confidence-manipulation parameters. It must also be asked to what degree any related metric would influence the long-term trust-calibration process, and for what kinds of users.

We additionally suggest a probabilistic expression of user confidence but do not explore a formalism that would benefit from it. Our further work investigates a useful formalism based on Dempster-Shafer theory, which we feel lends itself well to an evidence-based notion of user belief in propositions regarding the system's ability to meet objective-driven expectations.

ACKNOWLEDGMENTS

This research is supported in part by the NSF Grant IIS-2002082 and the Research and Artistry Opportunity Grant from Northern Illinois University.

REFERENCES

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

[2] Emma Beauxis-Aussalet, Michael Behrisch, Rita Borgo, Duen Horng Chau, Christopher Collins, David Ebert, Mennatallah El-Assady, Alex Endert, Daniel A Keim, Jörn Kohlhammer, et al. 2021. The Role of Interactive Visualization in Fostering Trust in AI. *IEEE Computer Graphics and Applications* 41, 6 (2021), 7–12.

[3] Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)* 12, 2 (2005), 293–327.

[4] André Calero Valdez and Martina Ziefle. 2018. Human factors in the age of algorithms. understanding the human-in-the-loop using agent-based modeling. In *International Conference on Social Computing and Social Media*. Springer, 357–371.

[5] Jin-Hee Cho, Kevin Chan, and Sibel Adali. 2015. A survey on trust modeling. *ACM Computing Surveys (CSUR)* 48, 2 (2015), 1–40.

[6] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[7] Keeley Crockett, Matt Garratt, Annabel Latham, Edwin Colyer, and Sean Goltz. 2020. Risk and Trust Perceptions of the Public of Artifical Intelligence Applications. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[8] Jessica Dai, Sina Fazelpour, and Zachary Lipton. 2021. Fair machine learning under partial compliance. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 55–65.

[9] Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371* (2020).

[10] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Springer, 449–466.

[11] Andrea Ferrario and Michele Loi. 2022. How Explainability Contributes to Trust in AI. *Available at SSRN 4020557* (2022).

[12] Brian J Fogg and Hsiang Tseng. 1999. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 80–87.

[13] Gartner. 2021. Top Strategic Technology Trends for 2021. https://www.gartner.com/smarterwithgartner/gartner-top-strategic-technology-trends-for-2021

[14] Wenkai Han and Hans-Jörg Schulz. 2020. Beyond trust building—Calibrating trust in visual analytics. In *2020 IEEE workshop on trust and expertise in visual analytics (TREX)*. IEEE, 9–15.

[15] Katherine Hawley. 2014. Trust, distrust and commitment. *Noûs* 48, 1 (2014), 1–20.

[16] Robert R Hoffman. 2017. A taxonomy of emergent trusting in the human–machine relationship. *Cognitive systems engineering: The future for a changing world* (2017), 137–164.

[17] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st international conference on intelligent user interfaces*. 164–168.

[18] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.

[19] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 624–635.

[20] Susan Joslyn and Jared LeClerc. 2013. Decisions with uncertainty: The glass half full. *Current Directions in Psychological Science* 22, 4 (2013), 308–315.

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

[22] J David Lewis and Andrew J Weigert. 2012. The social dynamics of trust: Theoretical and empirical research, 1985-2012. *Social forces* 91, 1 (2012), 25–31.

[23] Niklas Luhmann. 2000. Familiarity, confidence, trust: Problems and alternatives. *Trust: Making and breaking cooperative relations* 6, 1 (2000), 94–107.

[24] Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. 2021. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics* 113 (2021), 103655.

[25] Stephen Marsh and Pamela Briggs. 2009. Examining trust, forgiveness and regret as computational concepts. In *Computing with social trust*. Springer, 9–43.

[26] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[27] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[28] Indrit Troshani, Sally Rao Hill, Claire Sherman, and Damien Arthur. 2021. Do we trust in AI? Role of anthropomorphism and intelligence. *Journal of Computer Information Systems* 61, 5 (2021), 481–491.

[29] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–39.

[30] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.