Distrust in (X)AI - Measurement Artifact or Distinct Construct?

NICOLAS SCHAROWSKI* and SEBASTIAN A. C. PERRIG,

Center for General Psychology and Methodology, University of Basel, Switzerland

Trust is a key motivation in developing explainable artificial intelligence (XAI). However, researchers attempting to measure trust in AI face numerous challenges, such as different trust conceptualizations, simplified experimental tasks that may not induce uncertainty as a prerequisite for trust, and the lack of validated trust questionnaires in the context of AI. While acknowledging these issues, we have identified a further challenge that currently seems underappreciated - the potential distinction between *trust* as one construct and *distrust* as a second construct independent of trust. While there has been long-standing academic discourse for this distinction and arguments for both the one-dimensional and two-dimensional conceptualization of trust, distrust seems relatively understudied in XAI. In this position paper, we not only highlight the theoretical arguments for distrust as a distinct construct from trust but also contextualize psychometric evidence that likewise favors a distinction between trust and distrust. It remains to be investigated whether the available psychometric evidence is sufficient for the existence of distrust or whether distrust is merely a measurement artifact. Nevertheless, the XAI community should remain receptive to considering trust *and* distrust for a more comprehensive understanding of these two relevant constructs in XAI.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**.

Additional Key Words and Phrases: AI, XAI, Trust, Distrust, Attitude, Measures, Measurement, Operationalization, Psychometrics

ACM Reference Format:
Nicolas Scharowski and Sebastian A. C. Perrig. 2023. Distrust in (X)AI - Measurement Artifact or Distinct Construct?. In *CHI 2023: Workshop on Trust and Reliance in AI-Human Teams, April 23 – April 28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 8 pages.

1 INTRODUCTION

Trust has been studied for decades under different disciplinary lenses, such as philosophy [10], social sciences [12], and economics [3]. This has led to a multi-layered perspective on trust and sometimes divergent conceptions of trust in different disciplines. In the social sciences, trust has been defined as the expectation of non-hostile behavior; in the context of economics, trust is often conceptualized through game theory; in psychological terms, trust represents cognitive learning from experiences, and philosophically speaking, trust is based on moral relationships between individuals [1]. Researchers have introduced accounts of trust that are appropriate in interactions between humans and machines [23], including AI systems [16]. More recently, research on explainable AI (XAI) has regarded trust as a key motivation when creating more transparent and interpretable AI [28].

While there seems to be a consensus in the XAI community that trust is a critical factor in human-AI interaction, researchers have identified challenges in measuring trust in the context of AI. For example, different conceptualizations of trust exist that are not clearly distinguished from one another (e.g., appropriate trust [15], calibrated trust [22], warranted trust [16], or reliance [33]). These various conceptualizations may lead to differences in the operationalization of trust. For example, *trust as an attitude* should be viewed as a psychological construct and therefore be measured with

*Corresponding Author.

Manuscript submitted to ACM

questionnaires, whereas *reliance as a behavior* can be measured with observational methods such as such as behavioral changes [38]. Moreover, researchers mostly measure trust at a single point in time that does not meaningfully capture the dynamics of trust development over time [25], and empirical studies often use proxy-tasks rather than actual decision-making tasks [5] that do not necessarily impose uncertainty or risk that has been proposed as a prerequisite of trust [16]. Even if these challenges are addressed in the research design, AI-researchers who attempt to measure trust have to resort to questionnaires or survey scales from other disciplines because a validated questionnaire for trust in AI does not exist. One frequently used scale in XAI research is the *Trust between People and Automation* (TPA) scale developed by Jian et al. [17]. However, as the TPA was initially designed to examine trust in automation rather than trust in AI, researchers must rephrase the TPA items, such as "the system is dependable" to "the artificial intelligence is dependable." Terminological differences (e.g., using the word "AI" rather than "algorithm") impact people's perceptions and evaluations of technology [21], which raises concerns about whether an adapted scale still measures what it was initially intended to measure. As a consequence, the psychometric quality of a scale (i.e., reliability and validity) must be reevaluated after such adaptations [11, 18].

While it has been recognized that these issues complicate the measurement of trust in AI, we identified a different, noteworthy, and often underappreciated challenge - the lack of a theoretical consideration and potential distinction between *trust* on the one hand and *distrust* on the other. A first indication of this distinction was provided by Spain et al. [44], who validated the TPA in the context of automation and crucially revealed a two-factor model in a factor analysis that distinguished trust and distrust. These results suggest that trust is not uni-dimensional as initially proposed by Jian et al. [17], but a two-dimensional construct consisting of trust and distrust as distinct factors [44]. Preliminary results from our ongoing work support this conceptualization and also indicate that a two-factor solution is preferable for the TPA in the context of AI [31]. These findings have implications for using the TPA. Researchers have to consider the revealed two-factor structure and potential distinction between trust and distrust to obtain a better model fit and improved values for the scale in terms of validity and reliability compared to a single-factor conceptualization of trust. However, this raises the question of whether evidence of a two-factor structure is sufficient for two distinct and independent dimensions of trust and distrust or merely a research artifact caused by reverse-coded items.

The XAI community has provided important insights into how trust in AI can be developed and maintained, but distrust has been relatively understudied.[1] This ignores decades of research, which has been in a critical discourse on whether trust and distrust constitute the same construct at opposite ends of a continuum or should be treated as separate constructs on two distinct dimensions. A notable exception is the work from Schelble et al. [39], which appeared at the TRAIT 2022 workshop and partially motivated this position paper.

Our contribution to the TRAIT Workshop 2023 is threefold: First, we provide a concise but informative overview of the theoretical discourse and the main arguments for distrust as a distinct construct. Second, we critically discuss from a psychometric perspective if the existence of a two-factor structure of the TPA is sufficient evidence for distrust or whether a two-factor solution might be a measurement artifact caused by reverse-coded items. Third, we critically discuss the possible implications of distrust as a distinct construct for XAI and suggest opportunities for future research examining whether distrust genuinely exists independently from trust. If this is the case, this contributes to a more comprehensive understanding of trust *and* distrust in the context of AI that could inform the XAI community.

[1] Indicated in a search on the ACM Digital Library on the 15. February 2023, searching for the words trust/distrust in combination with AI or XAI in the keywords and abstracts. For instance, at CHI 22, there were 12 publications on trust and AI/XAI, while there were no publications on distrust and AI/XAI. Overall, there were 280 hits in the ACM Digital Library for trust, in contrast to just six for distrust.

2 TRUST AND DISTRUST: POLAR OPPOSITES, OR INDEPENDENT CONSTRUCTS?

According to Lewicki et al. [25] there generally are two conceptualizations of trust.

I The uni-dimensional model, which treats trust and distrust as bipolar opposites, ranging from distrust to trust [e.g., 17, 35, 41].

II The two-dimensional model, which argues that trust and distrust are two distinctly differentiable dimensions that can vary independently, each ranging from low to high [e.g., 14, 25, 29, 30, 36, 43].

Over the last 40 years, there have been advocates for both the uni-dimensional and two-dimensional models. In the following, we will introduce and mainly focus on the arguments for the two-dimensional trust model as we view this as the less established model.

The underlying question raised by these two models is whether it is conceivable that trust and distrust can exist simultaneously and independently or whether trust and distrust are two sides of the same coin [7]. The uni-dimensional model suggests that high trust equals low distrust, and low trust equals high distrust, implying that the manifestation of trust is always dependent on the manifestation of distrust [7]. However, from the perspective of the two-dimensional model, distrust is more than the absence of trust, and vice versa [20]. Thus, high trust does not automatically imply low distrust, and the two constructs can coexist simultaneously and independently.

Luhmann [29] is considered one of the main contributors to the theoretical foundation for a distinction between trust and distrust, and many later proponents of the two-dimensional model draw on his reasoning [14, 20, 24]. Luhmann argued that distrust is associated with stronger emotional reactions than trust and reflects the negatively charged human survival instinct, while trust is more calm and composed, rendering the two constructs distinct. Lewicki et al. [24] further developed this idea and claimed that trust is based on more positive emotional responses (e.g., hope, faith, confidence) and distrust on more negative emotions (e.g., fear, skepticism, cynicism), so they may not just be at different ends of the same continuum, but orthogonal [14]. Based on these emotional differences, Lewicki et al. proposed a 2x2 framework with trust on one y-axis and distrust on the x-axis to capture not only the positive and negative emotions but also the potential for high and low levels of each construct to coexist simultaneously. Within this 2x2 framework, each quadrant represents one possible combination of the two constructs: low trust/low distrust (quadrant 1), high trust/low distrust (quadrant 2), low trust/high distrust (quadrant 3), and high trust/high distrust (quadrant 4), with each quadrant characterizing a distinct relationship between the two constructs and different challenges that go with it [24]. More recently, neuroimaging studies have supported this proposed difference in the emotional makeup of trust and distrust by showing that trust is more associated with the brain's reward, prediction, and uncertainty area. In contrast, distrust is associated with the brain's intense emotions and fear of loss areas, suggesting different neurological processes [9].

The two-dimensional model (i.e., the potential coexistence of trust and distrust) is also supported by findings that attitudes can be ambivalent and possess both positive and negative components [34]. For example, smokers trying to quit smoking may have both positive and negative feelings towards cigarettes, suggesting that positive and negative attitudes can coexist simultaneously [6]. This coexistence of apparently contradictory emotions allows for a more complex view of the trust relationship [25], acknowledging that there may be reasons to both trust and distrust simultaneously within the same relationship [24]. For example, A trusts B to do Y, yet distrusts B to do Z. Harrison McKnight and Chervany [14] exemplified this with the cooperation between Stalin and Roosevelt in the second world war, where both parties trusted and distrusted each other at the same time. For Luhmann, trust and distrust are "functional equivalents," meaning that rational actors use both trust and distrust to contain and manage uncertainty and complexity, but they do so by different means [24]. Trust reduces complexity by compelling a person to take action that exposes them to risk

(i.e., undesirable outcomes are removed from consideration to form positive expectations [20]), while distrust reduces complexity by compelling a person to take protective action to reduce risk (i.e., undesirable outcomes are accentuated in consideration to form negative expectations [20]) [2]. In summary, an argument can be made that both the antecedents (e.g., the associated emotions) and the consequences (e.g., the resulting function) of trust and distrust may be distinct [6, 7, 13, 24]

This extensive work on the two-dimensional model, which for the purpose of this position paper can only be covered briefly, has led some authors to note that "most trust theorists now agree that trust and distrust are separate constructs that are the opposites of each other" [14, p. 42]. However, there are also influential contributions and compelling arguments for the one-dimensional model. For example, Schoorman et al. [41] replied to the statement above that "if they [trust and distrust] are opposites of each other, there is little added value to treating them as separate constructs" [41, p. 8]. The authors further noted that there is no theoretically or empirically viable evidence that trust and distrust are conceptually different and that researchers who studied distrust have merely reverse-coded measures of trust to represent their measures of distrust [41]. This remark points to a psychometric question, namely, whether the two-factor solution found by Spain et al. [44] for the TPA questionnaire by Jian et al. [17], which our preliminary findings confirmed in the context of AI [31], is sufficient evidence for the independent existence of distrust or whether it is merely a research artifact.

3 DISTRUST: MERELY A RESEARCH ARTIFACT?

In psychometrics (i.e., a branch of psychology concerned with the theory and technique of measurement), there is evidence that uni-dimensional models are, at times, mistakenly considered multidimensional due to errors or artifacts of measurement [40]. In the following, we will illustrate how these errors and artifacts can be introduced into questionnaires.

A questionnaire or survey scale typically consists of a list of questions, called items, that reflect on different aspects of the underlying construct(s). When measuring a construct indirectly through the items of a scale, researchers make a crucial assumption: The response to the items is caused by the strength or level of the underlying construct [8]. Thus, the construct that is not directly observable and its magnitude influences people's responses to the scale items. During scale development and refinement, researchers can use exploratory and confirmatory factor analyses to identify and confirm theoretical models that best fit the data that was gathered using the scale's items [4]. As part of this process, researchers form a theoretical model for their scale by defining and refining how many constructs are measured through the items and how these constructs relate to each other. Results from these processes thus shape how researchers understand their constructs of interest (e.g., trust and distrust) and how to use their scale for measurement in research.

However, past research has shown that so-called *reverse-coded items* can distort the factor structures of scales [32, 40, 42, 48], thus leading to false conclusions regarding the dimensionality and theoretical structure of a questionnaire. Reverse-coded items, also called negatively worded items, are items worded opposite to the regular scale items (i.e., positively formulated), which need to be re-coded prior to data analysis (e.g., a value of two on a Likert-type scale from 1 - 7 will be coded into a value of six). Two common strategies used to create reverse-coded items are negating the target expression (e.g., adding "not") or working with antonyms (e.g., "bad" instead of "good") [45]. For example, an XAI researcher developing a scale to measure trust could thus decide to create a reversed version of the item "The AI-system is reliable" [17] through negation ("The AI-system is not reliable") or an antonym ("The AI-system is unreliable").

Theoretically, respondents who agree with a regular item should similarly disagree with a reverse-coded item. As a result, the negatively worded items should yield comparable results to the regular items after re-coding. Nevertheless,

past research has cast doubt on this theoretical assumption. Respondents are likely to misinterpret or misrespond to reverse-coded items [37], either by responding carelessly or because of so-called *reversal ambiguity* [46]. Participants sometimes develop patterns of answering questionnaires based on the first few items they read (e.g., continuously responding with the value four), which reduces their attention [40]. The resulting carelessness leads respondents to overlook the negative wording of the reverse-coded items when filling out the questionnaire. Schmitt and Stuits [40] found that if just 10% of respondents are careless while filling out a questionnaire, the factor structure can be distorted, highlighting the gravity of careless responding.

In addition, the wording used to create reverse-coded items can be ambiguous to the respondents if it leaves room for interpretation (i.e., display "reversal ambiguity"), causing even those who are careful while filling out the questionnaire to misrespond [46]. In that case, even cautious respondents do not understand the antonyms used for item reversal in line with what the researchers intended, which happens especially in cross-cultural research [47]. Finally, Kam et al. [19] have highlighted that respondents with a neutral opinion on an issue are likely to choose answers towards a scale's midpoint and tend to agree with both the regular and the reverse-coded items. Given that such respondents have no strong opinion towards either extreme of the scale, this behavior is perfectly reasonable, even if it goes against what the researchers might have expected when constructing the scale [19]. As Priester and Petty [34] pointed out, the literary figure Hamlet both longs for and, at the same time, fears his death. As a result, he would provide primarily "neutral" or "slightly positive" responses toward suicide on a traditional bipolar attitude scale, which would cause him to agree with both regular and reversed items.

In summary, using reverse-coded items in questionnaires can result in a two-dimensional scale structure due to the agreement with both regular and reversed items for two reasons. On the one hand, this contradictory response behavior can happen due to mistakes by the respondents (i.e., careless responding), where the wording of the reverse-coded items is intentionally or unintentionally ignored. On the other hand, misunderstandings between respondents and researchers (i.e., reversal ambiguity, neutral responses) can also cause response patterns which the researchers did not expect when designing the scale because they expect opposite responses to the reverse and regular items, while respondents agree or disagree with both types. As a result of these mistakes and misunderstandings, the regular and reverse-coded items will load on two distinct constructs in factor analysis, not because there are two distinct constructs to be measured but due to methodological issues related to the item wording.

Returning to trust and distrust, it is possible that the two-dimensional structure of the TPA identified by past research [44] and in our preliminary findings [31] in the context of AI is not necessarily evidence that trust and distrust are two distinct constructs. Instead, it might indicate the presence of methodological artifacts influencing the measurement of subjective trust in an AI system. Such cases have been reported on in past HCI research and beyond (e.g., psychology [40]), for example, concerning the System Usability Scale [37] and the Usability Metric for User Experience [27], both scales which were initially assumed to measure a single uni-dimensional construct (i.e., usability). Regarding the System Usability Scale, Lewis and Sauro [26] recommended still treating the scale as uni-dimensional, measuring just one construct because a distinction due to item tone (i.e., positive or negative) would not make sense based on the underlying theory. A comparable assumption could be made concerning trust measured by the TPA, but in this case, a theoretical distinction between trust and distrust seems more reasonable, as we have illustrated.

4 DISCUSSION

Based on the arguments presented here, we conclude that while a distinction between trust and distrust on a theoretical level appears to be sensible, it is still to be determined if trust and distrust genuinely are two distinct constructs that can be measured independently or if they are the same construct, artificially separated due to methodological issues.

A discussion regarding the role of trust and distrust is relevant because members of the XAI community have emphasized that a key motivation of XAI is to *increase* trust of the user in a trustworthy AI (i.e., warranted trust) but also to *increase* the distrust of the user in a non-trustworthy AI (i.e., warranted distrust) [16]. While we generally agree with this position, the question arises as to how a one-dimensional model of trust can do justice to a simultaneous increase in trust and distrust and the difference between having low trust and distrust. A simplistic understanding of trust may not capture the complexity of peoples' attitudes toward AI. The argument that trust and distrust can coexist simultaneously seems particularly important in today's world, where AI is becoming more and more generalized and can perform multiple tasks (e.g., foundation models). This increased generalizability raises the question of which tasks to trust AI with and which to distrust it with. For example, a large language model (e.g., ChatGPT) might be trusted to write an email but distrusted to generate code or play chess. Consequently, there might be factors that contribute to the increase and decrease of trust, but also factors that contribute to the increase and decrease of distrust [24]. A two-dimensional model of trust seems more appropriate to account for this circumstance and arguably is more sensitive to such changes, as trust and distrust are not mapped on the same dimension.

If trust and distrust are distinct, efforts to eliminate distrust do not necessarily establish trust, and in that case, it is necessary to examine whether the two constructs have different antecedents and consequences [7]. The XAI community might want to give more attention to these differences, and future research could investigate how and if the two constructs can be measured independently in the AI context and what role methodological factors play in this regard. Research outside of an AI context [e.g., 2, 7] attempted to investigate if trust and distrust are distinct constructs in experimental settings. XAI researchers could attempt the same and examine how trust and distrust relate to one another, as well as if the two constructs can predict other subjective and objective measures relevant to the human-AI interaction (e.g., reliance).

There still seems to be no conclusive answer as to whether trust should be understood uni-dimensionally (with trust and distrust at the two ends of a continuum) or two-dimensionally (with distrust as a distinct and independent construct). However, it is precisely for this reason that this discourse should find its way into the XAI community. In this position paper, we outlined the theoretical arguments in support of a possible distinction between trust and distrust while at the same time showing that this question is not simply answered by psychometrics but by theoretical considerations that feed into experiments. The XAI community should regard these open questions as an opportunity for a more nuanced understanding of the factors influencing the human-AI interaction.

REFERENCES

[1] Peter Andras, Lukas Esterle, Michael Guckert, The Anh Han, Peter R. Lewis, Kristina Milanovic, Terry Payne, Cedric Perret, Jeremy Pitt, Simon T. Powers, Neil Urquhart, and Simon Wells. 2018. Trusting intelligent machines: Deepening trust within socio-technical systems. *IEEE Technology and Society Magazine* 37, 4 (dec 2018), 76–83. https://doi.org/10.1109/MTS.2018.2876107

[2] John Benamati, Mark A Serva, and Mark A Fuller. 2006. Are trust and distrust distinct constructs? An empirical study of the effects of trust and distrust among online banking users. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, Vol. 6. IEEE, 121b–121b. https://doi.org/10.1109/HICSS.2006.63

[3] Joyce Berg, John Dickhaut, and Kevin McCabe. 1995. Trust, reciprocity, and social history. *Games and economic behavior* 10, 1 (1995), 122–142. https://doi.org/10.1006/game.1995.1027

[4] Timothy A Brown. 2015. *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press, New York, NY, USA. 462 pages.

[5] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. ACM, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[6] John T Cacioppo and Gary G Berntson. 1994. Relationship Between Attitudes and Evaluative Space: A Critical Review, with Emphasis on the Separability of Positive and Negative Substrates. *Psychological Bulletin* 115, 3 (1994), 401–423. https://doi.org/10.1037/0033-2909.115.3.401

[7] Yong-Sheng Chang and Shyh-Rong Fang. 2013. Antecedents and distinctions between online trust and distrust: Predicting high-and low-risk internet behaviors. *Journal of Electronic Commerce Research* 14, 2 (2013), 149.

[8] Robert F. DeVellis. 2017. *Scale development: Theory and applications* (4 ed.). SAGE publications, Inc., Thousand Oaks, CA, USA.

[9] Angelika Dimoka. 2010. What does the brain tell us about trust and distrust? Evidence from a functional neuroimaging study. *MIS Quarterly* 34, 2 (2010), 373–396. https://doi.org/10.2307/20721433

[10] Francis Fukuyama. 1996. *Trust: The social virtues and the creation of prosperity.* Simon and Schuster.

[11] Mike Furr. 2011. *Scale construction and psychometrics for social and personality psychology.* SAGE publications, Ltd., 1 Oliver's Yard, 55 City Road, London EC1Y 1SP. https://dx.doi.org/10.4135/9781446287866

[12] Diego Gambetta et al. 2000. Can we trust trust. *Trust: Making and breaking cooperative relations* 13 (2000), 213–237. http://www.sociology.ox.ac.uk/papers/gambetta213-237.pdf

[13] D Harrison McKnight and Norman Chervany. 2001. While trust is cool and collected, distrust is fiery and frenzied: A model of distrust concepts. In *AMCIS 2001 Proceedings*. 883–888.

[14] D Harrison McKnight and Norman L Chervany. 2001. Trust and distrust definitions: One bite at a time. In *Trust in cyber-societies: Integrating the human and artificial perspectives*, Rino Falcone, Munindar Singh, and Yao-Hua Tang (Eds.). Springer, 27–54. https://doi.org/10.1007/3-540-45547-7_3

[15] Robert R Hoffman, Gary Klein, and Shane T Mueller. 2018. Explaining explanation for "explainable AI". In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 62. SAGE Publications Sage CA, Los Angeles, CA, USA, 197–201. https://doi.org/10.1177/1541931218621047

[16] 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* (Virtual Event, Canada) *(FAccT '21)*. ACM, New York, NY, USA, 624–635. https://doi.org/10.1145/3442188.3445923

[17] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (March 2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

[18] E. F. Juniper. 2009. Validated questionnaires should not be modified. *European Respiratory Journal* 34, 5 (2009), 1015–1017. https://doi.org/10.1183/09031936.00110209 arXiv:https://erj.ersjournals.com/content/34/5/1015.full.pdf

[19] Chester Chun Seng Kam, John P Meyer, and Shaojing Sun. 2021. Why Do People Agree With Both Regular and Reversed Items? A Logical Response Perspective. *Assessment* 28, 4 (2021), 1110–1124. https://doi.org/10.1177/10731911211001931

[20] Frens Kroeger. 2019. Unlocking the treasure trove: How can Luhmann's theory of trust enrich trust research? *Journal of Trust Research* 9, 1 (2019), 110–124. https://doi.org/10.1080/21515581.2018.1552592

[21] Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J. König, and Nina Grgić-Hlača. 2022. "Look! It's a Computer Program! It's an Algorithm! It's AI!": Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems?. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 581, 28 pages. https://doi.org/10.1145/3491102.3517527

[22] Markus Langer, Daniel Oster, Timo Speith, Lena Kästner, Holger Hermanns, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What Do We Want From Explainable Artificial Intelligence (XAI)? A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research. *Artificial Intelligence* 296 (Feb. 2021), 103473. https://doi.org/10.1016/j.artint.2021.103473

[23] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (March 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[24] Roy J Lewicki, Daniel J McAllister, and Robert J Bies. 1998. Trust and distrust: New relationships and realities. *Academy of management Review* 23, 3 (1998), 438–458. https://doi.org/10.2307/259288

[25] Roy J Lewicki, Edward C Tomlinson, and Nicole Gillespie. 2006. Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *Journal of management* 32, 6 (2006), 991–1022. https://doi.org/10.1177/0149206306294405

[26] J. R. Lewis and J. Sauro. 2017. Revisiting the Factor Structure of the System Usability Scale. *Journal of Usability Studies* 12, 4 (2017), 183–192.

[27] James R. Lewis, Brian S. Utesch, and Deborah E. Maher. 2013. UMUX-LITE: When There's No Time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2099–2102. https://doi.org/10.1145/2470654.2481287

[28] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. *Queue* 16, 3 (June 2018), 31–57. https://doi.org/10.1145/3236386.3241340

[29] Niklas Luhmann. 1979. *Trust and Power.* Wiley.

[30] Carol Xiaojuan Ou and Choon Ling Sia. 2009. To Trust or to Distrust, That is the Question: Investigating the Trust-Distrust Paradox. *Commun. ACM* 52, 5 (may 2009), 135–139. https://doi.org/10.1145/1506409.1506442

[31] Sebastian A. C. Perrig, Nicolas Scharowski, and Florian Brühlmann. 2023. Trust Issues with Trust Scales: Examining the Psychometric Quality of Trust Measures in the Context of AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany)

(CHI EA '23). Association for Computing Machinery, New York, NY, USA, 7 pages. https://doi.org/10.1145/3544549.3585808

[32] William J. Pilotte and Robert K. Gable. 1990. The Impact of Positive and Negative Item Stems on the Validity of a Computer Anxiety Scale. *Educational and Psychological Measurement* 50, 3 (1990), 603–610. https://doi.org/10.1177/0013164490503016

[33] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. ACM, New York, NY, USA, Article 237, 52 pages. https://doi.org/10.1145/3411764.3445315

[34] Joseph R Priester and Richard E Petty. 1996. The Gradual Threshold Model of Ambivalence: Relating the Positive and Negative Bases of Attitudes to Subjective Ambivalence. *Journal of personality and social psychology* 71, 3 (1996), 431–449. https://doi.org/10.1037/0022-3514.71.3.431

[35] Julian B Rotter. 1980. Interpersonal trust, trustworthiness, and gullibility. *American psychologist* 35, 1 (1980), 1–7. https://doi.org/10.1037/0003-066X.35.1.1

[36] Mark NK Saunders, Graham Dietz, and Adrian Thornhill. 2014. Trust and distrust: Polar opposites, or independent but co-existing? *Human Relations* 67, 6 (2014), 639–665. https://doi.org/10.1177/0018726713500831

[37] Jeff Sauro and James R. Lewis. 2011. When Designing Usability Questionnaires, Does It Hurt to Be Positive?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2215–2224. https://doi.org/10.1145/1978942.1979266

[38] Nicolas Scharowski, Sebastian AC Perrig, Nick von Felten, and Florian Brühlmann. 2022. Trust and Reliance in XAI–Distinguishing Between Attitudinal and Behavioral Measures. *CHI TRAIT Workshop* (2022), 6 pages. https://doi.org/10.48550/arXiv.2203.12318

[39] Beau G Schelble, Christopher Flathmann, Matthew Scalia, Shiwen Zhou, Christopher Myers, Nathan J Mcneese, Jamie Gorman, and Guo Freeman. 2022. Addressing the Spread of Trust and Distrust in Distributed Human-AI Teaming Constellations. *CHI TRAIT Workshop* (2022), 11 pages.

[40] Neal Schmitt and Daniel M. Stuits. 1985. Factors Defined by Negatively Keyed Items: The Result of Careless Respondents? *Applied Psychological Measurement* 9, 4 (1985), 367–373. https://doi.org/10.1177/014662168500900405

[41] F David Schoorman, Roger C Mayer, and James H Davis. 2007. An integrative model of organizational trust: Past, present, and future. *Academy of Management review* 32, 2 (2007), 344–354. https://doi.org/10.5465/amr.2007.24348410

[42] Chester A. Schriesheim and Kenneth D. Hill. 1981. Controlling Acquiescence Response Bias by Item Reversals: The Effect on Questionnaire Validity. *Educational and Psychological Measurement* 41, 4 (1981), 1101–1114. https://doi.org/10.1177/001316448104100420

[43] Sim B Sitkin and Nancy L Roth. 1993. Explaining the limited effectiveness of legalistic "remedies" for trust/distrust. *Organization science* 4, 3 (1993), 367–392. https://doi.org/10.1287/orsc.4.3.367

[44] Randall D. Spain, Ernesto A. Bustamante, and James P. Bliss. 2008. Towards an empirically developed scale for system trust: Take two. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 52, 19 (Sept. 2008), 1335–1339. https://doi.org/10.1177/154193120805201907

[45] Javier Suárez Álvarez, Ignacio Pedrosa, Luis M Lozano, Eduardo García-Cueto, Marcelino Cuesta, and José Muñiz. 2018. Using reversed items in Likert scales: A questionable practice. *Psicothema* 30, 2 (2018), 149–158. https://doi.org/10.7334/psicothema2018.33

[46] Bert Weijters and Hans Baumgartner. 2012. Misresponse to Reversed and Negated Items in Surveys: A Review. *Journal of Marketing Research* 49, 5 (2012), 737–747. https://doi.org/10.1509/jmr.11.0368

[47] Nancy Wong, Aric Rindfleisch, and James E. Burroughs. 2003. Do Reverse-Worded Items Confound Measures in Cross-Cultural Consumer Research? The Case of the Material Values Scale. *Journal of Consumer Research* 30, 1 (06 2003), 72–91. https://doi.org/10.1086/374697

[48] Xijuan Zhang, Ramsha Noor, and Victoria Savalei. 2016. Examining the effect of reverse worded items on the factor structure of the need for cognition scale. *PloS one* 11, 6 (2016), e0157795. https://doi.org/10.1371/journal.pone.0157795