Understanding AI Trustworthiness: A Scoping Review of AIES & FAccT Articles

SIDDHARTH MEHROTRA, University of Amsterdam & Delft University of Technology, The Netherlands
JIN HUANG*, University of Cambridge, United Kingdom
XUELONG FU, University of Amsterdam, The Netherlands
ROEL DOBBE, Delft University of Technology, The Netherlands
CLARA I. SÁNCHEZ, University of Amsterdam, The Netherlands
MAARTEN DE RIJKE, University of Amsterdam, The Netherlands

Background: Trustworthy AI serves as a foundational pillar for two major AI ethics conferences: AIES and FAccT. Current research often adopts techno-centric approaches, focusing primarily on technical attributes such as accuracy, reliability, robustness, and fairness, while overlooking the sociotechnical dimensions critical to understanding AI trustworthiness in real-world contexts.
Objectives: This scoping review aims to examine how the AIES and FAccT communities conceptualize, measure, and validate AI trustworthiness, identifying major gaps and opportunities for advancing a holistic understanding of trustworthy AI systems.
Methods: We conduct a scoping review of the AIES and FAccT conference proceedings to date, systematically analyzing how trustworthiness is defined, operationalized, and applied across different research domains. Our analysis focuses on conceptualization approaches, measurement methods, verification and validation techniques, application areas, and underlying values.
Results: While significant progress has been made in defining technical attributes such as transparency, accountability, and robustness, our findings reveal critical gaps. Current research often predominantly emphasizes technical precision at the expense of social and ethical considerations. The sociotechnical nature of AI systems remains less explored and trustworthiness emerges as a contested concept shaped by those with the power to define it.
Conclusions: An interdisciplinary approach combining technical rigor with social, cultural, and institutional considerations is essential for advancing trustworthy AI. We propose actionable measures for the AI ethics community to adopt holistic frameworks that genuinely address the complex interplay between AI systems and society, ultimately promoting responsible technological development that benefits all stakeholders.

JAIR Track: Surveys

JAIR Associate Editor: Marija Slavkovik

*Corresponding author. This work was partly conducted while the author was with the University of Amsterdam.

Authors' Contact Information: Siddharth Mehrotra, University of Amsterdam & Delft University of Technology, The Netherlands, ORCID: 0000-0003-4633-7023, s.mehrotra@uva.nl; Jin Huang, University of Cambridge, United Kingdom, ORCID: 0000-0001-9273-9037, jh2642@cam.ac.uk; Xuelong Fu, University of Amsterdam, The Netherlands, ORCID: 0009-0001-9608-1954, xlongfu@outlook.com; Roel Dobbe, Delft University of Technology, The Netherlands, r.i.j.dobbe@tudelft.nl; Clara I. Sánchez, University of Amsterdam, The Netherlands, ORCID: 0000-0001-9787-8319, c.i.sanchezgutierrez@uva.nl; Maarten de Rijke, University of Amsterdam, The Netherlands, ORCID: 0000-0002-1086-0202, m.derijke@uva.nl.

DOI: 10.1613/jair.1.20729

JAIR Reference Format:
Siddharth Mehrotra, Jin Huang, Xuelong Fu, Roel Dobbe, Clara I. Sánchez, and Maarten de Rijke. 2026. Understanding AI Trustworthiness: A Scoping Review of AIES & FAccT Articles. *Journal of Artificial Intelligence Research* 85, Article 32 (March 2026), 31 pages. DOI: 10.1613/jair.1.20729

1 Introduction

The rapid advancement and widespread adoption of artificial intelligence (AI) have ushered in a new era of technological innovation, bringing both immense potential and significant challenges. As AI increasingly permeates aspects of our lives, from healthcare to criminal justice, the need for trustworthy AI has become paramount. *Trustworthy* AI, as a concept, encompasses a multifaceted approach to AI systems that prioritizes safety, transparency, and ethical considerations for all stakeholders (Li et al. 2023). It extends beyond technical proficiency, embracing principles like reliability, fairness, explainability, and accountability. This has given rise to dedicated academic venues like the *AAAI/ACM Conference on AI, Ethics, and Society* (AIES) and *ACM Conference on Fairness, Accountability, and Transparency* (FAccT), fostering interdisciplinary discourse on AI ethics. It is now clear that a purely technology-centric view of trustworthiness is not enough. Trustworthy AI requires an interdisciplinary perspective that views AI systems as sociotechnical systems.

The sociotechnical nature of AI systems demands a holistic approach to trustworthiness, one that considers not only the technical aspects but also the complex interplay between AI and the broader social, cultural, and institutional contexts in which it operates. Malle (2022) lays out five elements of trustworthiness as competent, reliable, transparent, benevolent, and having ethical integrity – and calls out to study these elements in a broader sociotechnical setting. This expanded perspective is particularly crucial for the AI ethics community which aims to bridge the gap between computer science and other disciplines in addressing AI's ethical challenges.

In a study by Laufer et al. (2022), when asked about what values 36 self-selecting FAccT affiliates believe FAccT scholarship should address in the near future, participants reflected on the need for broader conceptions for trustworthiness. Central to this discourse is the recognition that the concept of trustworthiness is fundamental to understanding and predicting trust levels in AI systems. It becomes imperative to critically examine how the AIES & FAccT communities conceptualize and communicate trustworthiness, and to what ends these efforts are directed. This examination raises important questions about the commitments that trustworthy AI research in these venues signifies, or should signify, in the broader context of AI ethics and societal impact.

The study of the trustworthiness of AI systems has been a topic of interest for many years, even before the existence of the AIES & FAccT conferences. Scholars from various disciplines have identified values that determine the attribution of trustworthiness, revealing both similarities and differences across fields. For example, in interpersonal trust, competence, predictability, benevolence, and integrity have been highlighted as crucial values (Lahusen et al. 2024). For public institutions, the list extends to competence and reliability, procedures like transparency and accountability, and results including effectiveness and general welfare (Polemi et al. 2024). In the context of AI systems, properties such as reliability, robustness, safety, interpretability, explainability, fairness, transparency, and accountability have been identified as trust-relevant (Kaur et al. 2021; Lee and See 2004).

Trustworthiness has been used to refer to two sides of a coin (e.g., Lee and See 2004; Q. Liao and Sundar 2022; Mayer et al. 1995; Schlicker, Baum, et al. 2025; Visser et al. 2014). On the one hand, trustworthiness has been referred to as an objective attribute of the trustee (e.g., Floridi et al. 2018; Gillis et al. 2024; Jacovi et al. 2021; Kelp and Simion 2023; Zerilli et al. 2022). On the other hand, trustworthiness has been referred to as a trustor's subjective perception of a trustee's attributes (e.g., Lee and See 2004; Mayer et al. 1995; Schlicker, Baum, et al. 2025). Overall, trustworthiness is multifaceted, comprising several elements regardless of whether it is viewed as an inherent quality of the trusted party or as a subjective assessment made by those extending trust (Baer and Colquitt 2018; Dietz and Den Hartog 2006; Jacovi et al. 2021; Lee and See 2004). Therefore, by concentrating on

trustworthiness, we can assess the qualities and behaviors of AI systems that contribute to their reliability, safety, and ethical alignment.

Despite the extensive research on trustworthy AI (Toney et al. 2024), there remains a critical gap in understanding the sociotechnical nature of these systems. Few studies have adequately addressed the complex interplay between technical capabilities and social contexts in which AI operates. This paper aims to address these gaps by conducting a comprehensive scoping review of articles published in AIES & FAccT conferences to date. Through our analysis, we seek to answer several key questions:

(RQ1) How is trustworthiness conceptualized in the context of AI systems within AIES & FAccT proceedings?
(RQ2) What methodologies are employed to measure, verify, and validate trustworthiness?
(RQ3) Which application areas are most prominently represented?
(RQ4) What underlying values and ethical considerations drive this body of work?

We explicitly focus on values because an account of value embodiment in AI aids in assessing whether designed AI systems indeed embody a range of moral values (Mehrotra, Jonker, et al. 2021), e.g., those articulated by the EU High-Level Expert Group (European Union HLEG 2019). To this end, we examine three key dimensions: intended, embodied, and realized values following the framework of Poel (2020) to do a thematic analysis on our corpus. Our motivation to use this framework is that it helps us overcome the limitations of focusing solely on intentions (which may not manifest in the system) or outcomes (which may be influenced by external factors) by emphasizing embodied values—those intentionally and successfully embedded in the system.

Overall, our core contributions are:

(1) A systematic analysis of how trustworthiness is conceptualized, measured, and validated within the AIES & FAccT community.
(2) Identification of major gaps in current research, particularly regarding the sociotechnical aspects of AI systems.
(3) A critical examination of the values and ethical considerations underpinning trustworthy AI research.
(4) A discussion of the intellectual and broader impact of AIES & FAccT conferences in studying AI trustworthiness.

2 Related Work

Trust is a much-discussed topic in algorithmic decision-making, especially in the area of AI (Mehrotra, Degachi, et al. 2024). In the development of trust, the process by which a human assesses the trustworthiness of a system, leading to their perception of trustworthiness, is crucial. Only with an accurate trustworthiness assessment can people base their trust on adequate expectations about a system's capabilities and limitations and make informed decisions. Trustworthiness of AI systems has been studied from multiple disciplines such as computer science (Liu et al. 2023), psychology (Schlicker, Uhde, et al. 2022), public administration (Lahusen et al. 2024), and medicine (J. Zhang and Z.-m. Zhang 2023). Below, we provide a background on studying trustworthiness in four subject areas, namely: computer science, law, social sciences, and humanities. These areas are common in the AIES & FAccT conferences. AIES includes experts from various disciplines such as ethics, philosophy, economics, sociology, psychology, law, history, and politics (Furman et al. 2018), while FAccT brings together scholars from computer science, law, social sciences, and humanities (Friedler and Wilson 2018). These areas serve as the foundation for our interdisciplinary approach to understanding the multifaceted nature of AI trustworthiness.

2.1 Computer Science

Computer science has predominantly focused on technical perspectives for trustworthy AI applications, emphasizing three characteristics, (i) robustness, fortifying AI models against malicious attacks such as adversarial attacks; (ii) generalization, ensuring maintaining performance on unseen out-of-distribution (OOD) data; and

(iii) interpretability, improving understanding of AI model predictions (Mucsányi et al. 2023; J. Wang et al. 2023). Mucsányi et al. (2023) list common pitfalls in evaluating trustworthy machine learning models, e.g., inconsistent coding for evaluation metrics despite being the same mathematically, confounding multiple factors in method comparisons, training and test samples overlap, and lack of validation set. Building on the foundational principles of trustworthy machine learning, recent research has begun to investigate the characteristics that apply to large language models (LLMs) and mechanisms for evaluating their trustworthiness. Liu et al. (2023) survey key dimensions for assessing LLM trustworthiness. These include reliability, safety, fairness, resistance to misuse, explainability and reasoning, adherence to social norms, and robustness. Liu et al. (2023) highlight that a key principle for evaluating an LLM's trustworthiness is the generation of proper test data across the aforementioned dimensions. Finally, Toney et al. (2024) review similarities and differences between governments' and researchers' definitions and frameworks on trustworthy AI. The authors find inconsistencies between policy and research term frequencies, highlighting the different focuses of each group on trustworthy AI, distinct from our work focusing on AI trustworthiness rather than the use of the terminology within the AIES & FAccT community.

Studies in computer science are often tech-centred, focusing primarily on technical methodologies but overlooking socio-cultural, ethical, and legal dimensions of trustworthiness. But achieving trustworthiness demands interdisciplinary collaboration. Thus, we consider AI systems as socio-technical and explore how the AIES & FAccT community is deepening our insight into trustworthiness of AI.

2.2 Social Sciences

Social science provides a broad perspective on AI trustworthiness focusing on its societal, institutional, economic, and political implications and dynamics. E.g., Dacon (2023) investigates the impact of AI trustworthiness on various aspects of society, like the environment, human society, and societal values. These implications are abstracted into three social principles: (i) harm prevention, which focuses on ensuring safety, security, reliability, and privacy; (ii) explicability, which emphasizes explainability and transparency of the system; and (iii) fairness, which consists of accountability and the well-being of society and environment. Thiebes et al. (2021) evaluate AI trustworthiness frameworks and synthesize the recurring themes in five principles: beneficence, non-maleficence, autonomy, justice, and explicability. They also conceptualize the *DaRe4TAI* framework, which structures and directs data-driven research in trustworthy AI while addressing tensions between the five principles. Similar framework-oriented approaches are pursued by other social science scholars, as they form a crucial connection with policymaking (Kusche 2024; Polemi et al. 2024). We adopt a similar lens in this review as Kusche (2024) and Polemi et al. (2024) to discuss the results of our corpus analysis with a socio-technical focus by thinking about the entire ecosystem in which our AI systems operate.

2.3 Law

The question of AI trustworthiness is crucial in the legal domain, where even minor inaccuracies can have significant consequences for individuals navigating the complexities of legal processes. Studies have demonstrated the capabilities of LLMs in understanding and responding to natural language queries. However, the tendency of LLMs to generate inaccurate and incomplete information raises serious concerns about their trustworthiness in providing legal guidance (Kattnig et al. 2024). E.g., J. Tan et al. (2023) investigate ChatGPT's ability to provide legal information using several simulated cases and and find that laypeople often over-trust it. More generally looking at the use of AI systems in law, Steenhuis (2024) raises important questions of exploring trustworthiness of such systems for methods of service delivery, replacement of routine legal tasks and essential legal assistance to those who might otherwise go without. Potential answers to these explorations can be derived from Hagan (2020)'s work who provides legal design testing metrics to explore trustworthiness of AI systems plus methodology and "design deliverables" based on these metrics with California state courts' Self Help Centers (Hagan 2018).

Overall, AI systems hold significant promise for improving accessibility and efficiency in the legal domain but face challenges related to inaccuracies, hallucinations, and over-reliance, raising concerns about trustworthiness.

2.4 Humanities

The study of AI trustworthiness within the humanities emphasizes ethical, philosophical, and cultural aspects, addressing critical issues such as moral agency, responsibility, and the societal impact of AI decision-making (Chun and Elkins 2023; Noller 2024; Rawas 2024). Balmer (2023) critically examines AI's role in society through creative methodologies, while others explore how AI reshapes creativity, authorship, and cultural interpretation (Lim 2018). Historical and philosophical analyses (e.g., Coeckelbergh 2022) provide valuable context for understanding contemporary AI technologies, alongside inquiries into AI consciousness and religious perspectives that highlight cultural and ethical considerations (Alkhouri 2024; Fiore 2024; He 2024). These diverse approaches underline the importance of contextualizing AI systems within broader humanistic and historical frameworks to assess their trustworthiness effectively. The integration of ethical, cultural, and philosophical dimensions in evaluating AI trustworthiness offers a more comprehensive understanding of its implications, enabling more informed and responsible development of AI systems.

3 Methodology and Corpus Overview

3.1 Methodology

We employ a Scoping Literature Review (SLR) following the guidelines by Arksey and O'Malley (2005) to identify articles studying trustworthiness of AI systems. SLRs guide the gathering and identification of papers in a topic area for scrutiny (Kastner et al. 2012), which enables us to perform our thematic investigation of AIES & FAccT scholarship. We use qualitative manual coding and computational corpus analysis, including topic modeling, to extract themes and patterns from the data.

Data Collection. The data consists of peer-reviewed articles published by the AIES & FAccT conferences between 2018–2025, downloaded from the ACM website on September 05, 2025.[1] In total, 241 articles (83 AIES, 158 FAccT) are obtained that include the keyword "trustworthiness" in the full-text excluding references; 6 non-archival extended abstracts are excluded, leaving us 235 articles for screening.

Screening and Selection Criteria. We note several challenges in reviewing the AIES & FAccT papers. The concept of trustworthiness is often embedded throughout papers instead of being the core theme. For example, auditing algorithmic systems or exploring potential biases in AI systems eventually help in understanding their trustworthiness, but this is often a point of discussion rather than the core contribution. Therefore, it is difficult to conclude that these articles engage with the core concept. Hence, we aim for a balanced approach studying trustworthiness, by including two types of article in our analysis: (i) articles that directly address trustworthiness as a central theme, like (Ferrario 2025; Ferrario and Loi 2022; Jacovi et al. 2021; Q. Liao and Sundar 2022), and (ii) articles where trustworthiness is not the main focus, but the authors explain how their findings contribute to a broader understanding of system trustworthiness and, ultimately, user trust.

Therefore, keeping our balanced approach, following prior literature reviews on trust (Benk et al. 2024; Mehrotra, Degachi, et al. 2024; Vereschak et al. 2021) and best research practices to study a particular topic (Ajmani et al. 2023; Proferes et al. 2021), we define our *inclusion criteria* as:

(I1) One of the contributions of the article is linked to understanding trustworthiness of an algorithmic system.
(I2) The article engages with trustworthiness as a component of their measure, directly influenced by the results, or as a key discussion theme.

[1]Note that the proceedings of AIES 2025 were not available when we conducted this study. Hence, AIES 2025 articles are excluded from this review.

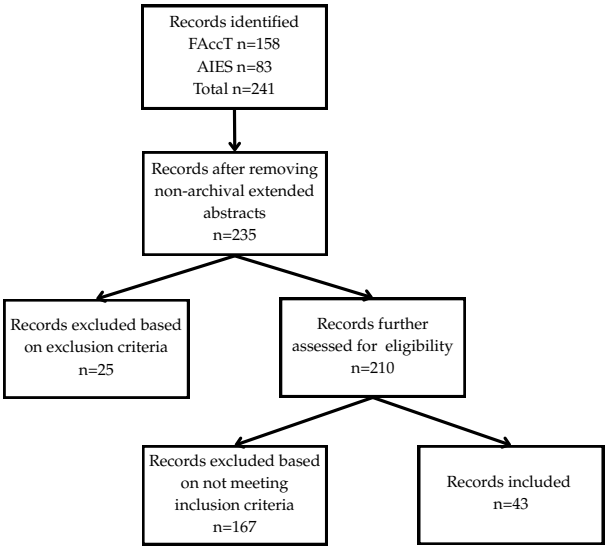

Fig. 1. Flowchart of the articles reviewing process following the PRISMA protocol (Page et al. 2021).

Our *exclusion criteria* are:

(E1) The article discusses the need for AI trustworthiness without directly defining, measuring, or modeling it.
(E2) The article is classified as a survey, scoping review, or literature review.

As we pointed out in the introduction, we did not include articles that discuss the need and nature of trustworthiness without providing a direct measurement or operationalization as a deliberate choice. This methodological decision means the corpus focuses specifically on papers that operationalize trustworthiness through concrete measurement approaches, rather than those that engage with trustworthiness primarily at a theoretical or conceptual level without empirical assessment methods. The final corpus consists of 43 papers after applying our inclusion and exclusion criteria. An overview of our review process following the PRISMA protocol (Page et al. 2021) is shown in Figure 1.

Free-form Questions. Following the goal-question-metric (GQM) approach (Basili et al. 1994), we analyze the articles along the GQM dimensions that are phrased as questions: (i) what is the goal of the paper, (ii) what are different research questions related to this goal, and (iii) what are metrics to measure some aspects/factors of this goal? This framework provides us with a way to categorize the understanding of trustworthiness in the AIES & FAccT community in the form of generic free-form questions closely tied to the research questions from Section 1.

We employ a thematic analysis to analyze the 43 articles and address our research questions. First, we develop a coding scheme to extract relevant information from each paper. For defining trustworthiness (RQ1), we identify key terms, phrases, and concepts used across papers to formulate common definitional themes. We further link the identified themes with the International Organization for Standardization (ISO) standard 5723 on trustworthiness (ISO 2022), allowing us to track frequency and depth of coverage. The ISO standard is chosen as a reference framework as it represents an internationally recognized benchmark for trustworthiness in technical systems (Toney et al. 2024). Furthermore, to understand the drivers of studying trustworthiness, we use both inductive and deductive coding approaches, first allowing themes to emerge naturally from the text then mapping these against common clusters.

For trustworthiness measures (RQ2), we develop a hierarchical coding structure to categorize assessment methods and validation and verification techniques. For application scenarios (RQ3), we employ open coding to identify emerging categories of use cases, followed by axial coding to establish relationships between different application contexts. Finally, to better understand the role of values in AI trustworthiness (RQ4), we analyze intended, embodied, and realized values. We followed Poel (2020)'s work to study the role of values. According to his work, intended values refer to the ethical principles and societal benefits that AI systems aim to achieve, embodied values are those integrated into their design and implementation, and realized values are the actual outcomes observed in practice. We use a two-stage coding process where we first identify explicitly stated intended values related to trustworthiness and then compare these against realized outcomes and embodied values. This involves creating pairs of intended-realized values and analyzing any gaps or alignments between them.

Our coding process is iterative, with initial codes refined over multiple analysis rounds. To ensure reliability, three researchers independently code a subset of papers (25%) and calculate an average Cohen's Kappa coefficient ($\kappa = 0.62$). Through iterative discussions, they update their understanding, recode previous papers to include emergent concepts, include new papers in small batches of five and achieve improved reliability. The team collaborated for about two months to discuss findings and potential discrepancies in coding of the articles. Through multiple discussions, discrepancies were reevaluated and corrected, resulting in the final coefficient of $\kappa = 0.91$.

3.2 Corpus Overview

First, we perform a metadata analysis on the final corpus of 43 articles, focusing on understanding the trustor's background, the object of trustworthiness, variables, and study type. Our meta analysis reveals that there are six interlinked trustors: users of a specific AI system (20),[2] citizens (12), developers (4), designers (3), practitioners (2), and public administrators (2). As to the object of trustworthiness, the corpus revolves around studying trustworthiness of AI linked with data sources (19) and institutions developing the AI system (12). The role of trustworthiness as a study variable is almost equally divided among the papers as independent variable (21) and dependent variable (22). Finally, trustworthiness is studied normatively in only 1/3rd of the studies while the remaining 2/3rd studies it following an empirical study design.

Second, Figure 2 shows a heatmap of the log-normalized, min-max scaled word occurrences of trustworthiness dimensions by ISO (2022), in the full texts of the included corpus. Dimensions like quality, transparency, and accuracy occur frequently throughout the full texts, while authenticity, controllability, and resilience are mentioned relatively less. Some articles, like Paraschou et al. (2025) and Toney et al. (2024), refer to a wide array of dimensions.

Third, we broaden the analysis and discover themes across the initial keyword-based selection of 235 articles, without extended abstracts, through topic modeling. We apply BERTopic (Grootendorst 2022), which leverages c-TF-IDF and the capabilities of the transformer-based language model BERT (Devlin et al. 2018), to analyze topics that are both coherent and contextually meaningful. The generated topics are analyzed using two visualizations as shown in Figure 3. On the one hand, we present the normalized topic distributions over time using a heatmap to capture trends and shifts across publication years. We observe a consistent prevalence and recent increase of topics relating to "explanations" and "accountability", while there is a relative decline in topics referring to ethics and norms. On the other hand, we visualize the representations and relative distance of the topics in a 2D plot. Notably, topics surrounding "fairness," "transparency," and "ethics" lie in closer proximity to each other, while they are more distanced from the topic around "explanations." Also, we find that although fairness and bias are

[2]The number in brackets denote the count of articles corresponding to a specific dimension.

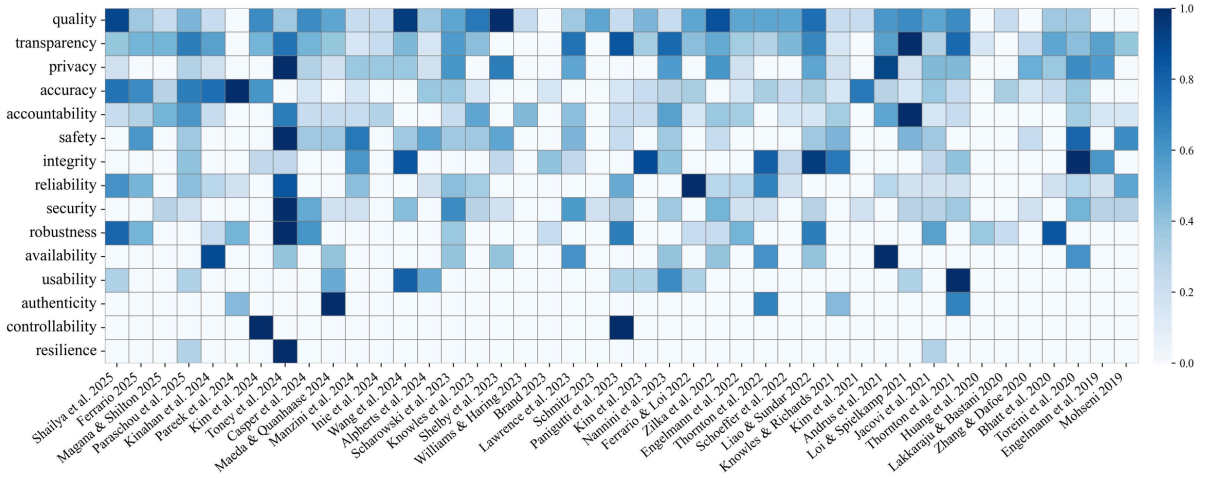

Fig. 2. Heatmap of the normalized frequencies of trustworthiness dimensions in each paper of the final corpus ($N = 43$).

Topic 1 (72): explanations, human, decision, language, study

Topic 2 (85): transparency, accountability, public, human, technologies

Topic 3 (49): fairness, bias, algorithmic, fair, discrimination

Topic 4 (29): ethics, ethical, moral, responsible, norms

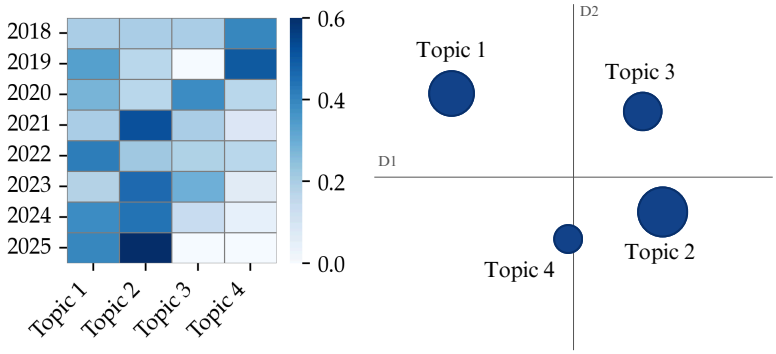

Fig. 3. Visualizations of topics generated using BERTopic applied to the initial keyword-based selection of articles ($N = 235$). First, the heatmap (left) shows normalized topic distributions, grouped by year. Values refer to the proportion of documents associated with the topic. Second, the intertopic distance plot (right) shows the relative positions of topics in a 2D space.[3]

key ethical concerns, the considerable distance between fairness-related topics and ethical themes suggests that papers tend to emphasize one while giving limited attention to the other.

4 Understanding of AI Trustworthiness in the AIES & FAccT Community

4.1 Definition or Conceptualization

To determine how to understand an AI system's trustworthiness, it is crucial to examine how we define or conceptualize it. *Definition* refers to how trustworthiness is formally described and articulated in the selected papers and *conceptualization* encompasses the broader understanding of trustworthiness, e.g., considering its role in systems.

First, the AIES & FAccT community presents a rich and multifaceted understanding of AI trustworthiness, as evidenced across numerous studies. Through our coding and thematic analysis, we identify seven key conceptualization themes, summarized in Table 1 per venue.

An analysis of the conceptual landscape reveals several critical insights about how AI trustworthiness is understood in contemporary scholarship. First, there exists a fundamental tension between transparency-focused approaches (T1) and technical robustness perspectives (T6), suggesting competing philosophies about whether trustworthiness derives primarily from system interpretability or from demonstrated reliability regardless of internal opacity. Second, the prominence of anthropomorphization themes (T3) alongside ethical considerations (T4) indicates growing recognition that trust in AI cannot be divorced from human psychological tendencies to perceive social relationships with technology, a finding that challenges purely technical or procedural definitions. Third, the emphasis on broader societal implications (T5) and regulatory ecosystems signals an important shift from treating trustworthiness as an individual system property to understanding it as embedded within institutional contexts where harms, accountability, and public opinion are distributed unevenly across populations.

Finally, the reliance on established frameworks like Mayer's model and NIST/ISO definitions (T7) alongside emerging concepts of perceived parasocial relationships suggests the field is simultaneously borrowing from organizational trust theory while grappling with AI-specific phenomena that existing models may not adequately capture. Importantly, considering the distribution of articles per venue (AIES: 12, FAccT: 31) the themes are almost equal with AIES being on the higher end on T4 indicating the leaning towards philosophical and standard AAAI/ACM technical crowds, while FAccT attracts a higher density of legal scholars, sociologists, and HCI researchers (Acuna and Liang 2021). These insights collectively reveal that AI trustworthiness is conceptualized not as a unified construct but as a contested terrain where technical, psychological, ethical, and societal dimensions intersect and sometimes conflict.

Table 1. Understanding of AI Trustworthiness: Key themes in definitions and conceptualizations.

Key themes	**Sub-themes**
T1: Emphasis on transparency as a foundational aspect of trustworthiness (AIES: 3,[4] FAccT: 8)	Trustworthiness cues (Q. Liao and Sundar 2022), *Good* explanation (Nannini et al. 2023), Open communication (Brand 2023), Transparency, provenance & connections between them (Thornton et al. 2021), Computational reliabilism & anti-monitoring (Ferrario and Loi 2022), Expressing uncertainty (S. S. Y. Kim, Q. V. Liao, et al. 2024), Transparency with human oversight (Panigutti et al. 2023), Explanation trustworthiness (Mohseni 2019; Shailya et al. 2025), Transparency of algorithmic tools deployed in the criminal justice system (Zilka et al. 2022)
T2: The importance of benchmarks, rigorous auditing and compliance mechanisms (AIES: 1, FAccT: 3)	White- and outside-the box audits (Casper et al. 2024), Certification labels (Scharowski et al. 2023), Performance benchmarks and evaluation measures (Lawrence et al. 2023), Mapping regulatory guidelines to philosophical account of accountability (Loi and Spielkamp 2021)
T3: Perceived anthropomorphization (FAccT: 3)	(de)Anthropomorphized description of AI system (Inie et al. 2024), Perceived parasocial relationships (Maeda and Quan-Haase 2024), Facial AI inferences (Engelmann, Ullstein, et al. 2022)

[3]The representations of topics are reduced to two dimensions using the uniform manifold approximation and projection dimensionality reduction technique, enabling the generation of a 2D plot to visualize the topics.
[4]This value represents the count of articles per venue for the theme.

T4: Ethical considerations, including fairness and respect for user intent (, AIES: 5, FAccT: 4)	Inclusive & intersectional algorithms (E. Kim et al. 2021), Data leakage & reproducibility (Kinahan et al. 2024), Model interpretability (Lakkaraju and Bastani 2020), Understanding of harms and how (and by whom) this has been informed (Knowles, Fledderjohann, et al. 2023), Interaction between risks and the dispositions of the trustee (Williams and Haring 2023), Informational fairness (Schoeffer et al. 2022), Achieving fairness (Schmitz 2023), Fairness by design (Huang et al. 2020), Fair data procurement (Andrus et al. 2021a)
T5: Broader societal implications: Regulatory ecosystem and accountability caused by harms (AIES: 2, FAccT: 5)	Social trustworthiness score (Engelmann, M. Chen, et al. 2019), Trustworthiness criteria for policy-making (Alpherts et al. 2024), Public trust in AI & institutional trust (Knowles and Richards 2021), AI Assistant design & Organizational practices and third-party governance (Manzini et al. 2024), sociotechnical harms (Shelby et al. 2023), Public opinion (B. Zhang and Dafoe 2020), Safe, reliable, and acceptable to users and public (Magaña and Shilton 2025)
T6: Justified reliance on AI's reliability and technical robustness (FAccT: 3)	Reasonable trust in model's output (Bhatt et al. 2020), Warranted trust (Jacovi et al. 2021), AI system must maintain to function correctly within its context of application (Ferrario 2025)
T7: Mayer's (Mayer et al. 1995) and Lee & See's model (Lee and See 2004) & NIST/ISO definition (FAccT: 9)	Ability, benevolence & integrity (S. S. Y. Kim, Q. V. Liao, et al. 2024; S. S. Y. Kim, Watkins, et al. 2023; Q. Liao and Sundar 2022; Paraschou et al. 2025; Pareek et al. 2024; Thornton et al. 2022; R. Wang et al. 2024), ABI+ (ABI, Predictability) framework (Toreini et al. 2020), NIST & ISO Definition (Toney et al. 2024)

We also study the trustworthiness dimensions defined by the ISO trustworthiness standard (ISO 2022) to ensure that our analysis is based on internationally recognized standards. 'Transparency' is the most frequently mentioned ISO dimension, appearing in 21 out of 43 papers, followed by 'accountability' and 'explainability' in 12 papers. The less frequently mentioned ISO dimensions are 'controllability' (5), 'resilience' (4), 'robustness' (4), 'security' (4), 'safety' (3), 'reliability' (3), 'usability' (3), and 'provenance' (2).

While many studies implicitly link trustworthiness to accuracy, reliability, transparency, and fairness, increasing attention is being given to more nuanced conceptualizations, such as stakeholder-centric approaches and systemic perspectives. Frequent references to ISO dimensions, particularly transparency, accountability, and explainability, as also shown earlier in our quantitative analysis, underscore their importance in building trustworthy AI systems. Simultaneously, our topic modeling also showed considerable distances between dimensions like fairness, ethics, and transparency. However, a more comprehensive and standardized approach to defining and measuring AI trustworthiness, which consistently considers multiple dimensions, is needed for the responsible development and deployment of AI systems.

4.2 Drivers

Anjomshoae et al. (2019) have underlined the importance of considering the intended purpose when studying the trustworthiness of AI systems. To this end, this scoping review seeks to understand why the reviewed studies focus on understanding trustworthiness. Most of these works stated their motivation or intended purpose for

studying trustworthiness. Table 2 lists the drivers of the 43 papers included in the review, with some papers having more than one driver.

Increasing users' trust in the system, ethical considerations, and conformity with regulations such as those of (European Union HLEG 2019; Executive Office of the President 2020; Japan Expert Group on How AI Principles Should be Implemented 2022; Science Australia Department of Industry and Resources 2023; The Government of Canada 2023a,b; United Kingdom Information Commissioner's Office 2023; Vought 2020), and accuracy are among the listed motivations for the explanations. The table reveals that trust and ethical considerations are the most prominent drivers of studying AI trustworthiness. Naturally, trust and trustworthiness go hand in hand to increase users' confidence in the system by understanding how its reasoning mechanism works (Anjomshoae et al. 2019). In applications requiring human-AI interaction, honesty and accuracy are among the main drivers for trustworthiness to ensure that AI's decision-making is reliable and fair (Pareek et al. 2024; Schoeffer et al. 2022). For public AI systems, trustworthiness drivers are often linked with conformity with regulations and guidelines, and economic growth and cultural values (Brand 2023; Engelmann, M. Chen, et al. 2019; Knowles and Richards 2021; Loi and Spielkamp 2021; B. Zhang and Dafoe 2020). Finally, Kinahan et al. (2024) identify reproducibility of their results as a driver for studying trustworthiness for their system. Overall, these studies provide a holistic overview of the key drivers for studying AI trustworthiness.

Table 2. The drivers of AI trustworthiness of the primary studies covered by the review.

Drivers	**Representative articles**
D1: Conformity with regulations & Guidelines and ethical considerations (AIES: 8, FAccT: 13)	(Andrus et al. 2021a,b; Brand 2023; Casper et al. 2024; Engelmann, Ullstein, et al. 2022; Ferrario 2025; Huang et al. 2020; Kinahan et al. 2024; Knowles and Richards 2021; Lawrence et al. 2023; Loi and Spielkamp 2021; Mohseni 2019; Nannini et al. 2023; Panigutti et al. 2023; Scharowski et al. 2023; Schmitz 2023; Schoeffer et al. 2022; Shailya et al. 2025; Shelby et al. 2023; Thornton et al. 2021; Toney et al. 2024; Zilka et al. 2022)
D2: Promoting honesty (AIES: 1, FAccT: 2)	(Alpherts et al. 2024; Engelmann, M. Chen, et al. 2019; Williams and Haring 2023)
D3: Trust (AIES: 4, FAccT: 20)	(Alpherts et al. 2024; Bhatt et al. 2020; Ferrario and Loi 2022; Inie et al. 2024; Jacovi et al. 2021; S. S. Y. Kim, Q. V. Liao, et al. 2024; S. S. Y. Kim, Watkins, et al. 2023; Knowles, Fledderjohann, et al. 2023; Knowles and Richards 2021; Lakkaraju and Bastani 2020; Q. Liao and Sundar 2022; Maeda and Quan-Haase 2024; Magaña and Shilton 2025; Manzini et al. 2024; Mohseni 2019; Paraschou et al. 2025; Pareek et al. 2024; Schoeffer et al. 2022; Thornton et al. 2021, 2022; Toreini et al. 2020; R. Wang et al. 2024; Williams and Haring 2023; B. Zhang and Dafoe 2020)
D4: Economic growth & Cultural values (AIES: 1, FAccT: 1)	(Brand 2023; Engelmann, M. Chen, et al. 2019)
D5: Reproducibility (FAccT: 1)	(Kinahan et al. 2024)
D6: Accuracy (AIES: 1, FAccT: 2)	(Alpherts et al. 2024; E. Kim et al. 2021; Mainz et al. 2023)

4.3 Measurement and Verification & Validation

Jacobs and Wallach (2021) have introduced measurement theory for AI/ML systems, where validation and verification are key to ensure we are measuring what we are trying to measure. In the context of AI trustworthiness, measurement refers to quantifying specific attributes or characteristics of the AI system, typically aligned with ISO trustworthiness dimensions. Following ISO 9001 (DNV 2015), verification involves confirming through examination and provision of objective evidence that the specified trustworthiness dimensions are fulfilled. Validation is similar to verification, but must be confirmed under real-world usage conditions.

Measurement. To analyze the measurement techniques for AI trustworthiness, we follow Schlicker, Baum, et al. (2025)'s distinction of actual and perceived trustworthiness. Schlicker, Baum, et al. (2025) define a system's actual trustworthiness as a latent construct that indicates the true value of a system's trustworthiness (in alignment with Realistic Accuracy Model (Funder 1995)), e.g., benevolence, integrity, and ability (Lee and See 2004; Mayer et al. 1995; Mehrotra, Degachi, et al. 2024). Similarly, perceived trustworthiness reflects the result of a trustor's assessment of the trustee's actual trustworthiness.

Analysis of measurement strategies reveals several important insights about the operationalization of AI trustworthiness. First, the field exhibits a problematic conflation between measuring trust (a psychological state) and measuring trustworthiness (system properties that warrant trust). Reliance behavior and Likert scales capture user perceptions while technical metrics assess objective capabilities, yet the relationship between these remains theoretically underdeveloped. Second, the diversity of approaches, from LDA topic modeling to ethical charters to role-playing exercises, suggests that trustworthiness is inherently multi-faceted and resists reduction to single metrics, yet this proliferation also indicates a lack of consensus on what actually constitutes valid measurement. Third, the emergence of documentation artifacts like model cards (Kinahan et al. 2024) and AI assessment catalogs (Schmitz 2023) represents an important shift toward longitudinal, context-aware evaluation rather than snapshot assessments, implicitly acknowledging that trustworthiness is not static but emerges through ongoing scrutiny. Finally, the inclusion of mediator and trustee role factors (social embeddedness, situational normality, structural assurance) indicates growing sophistication in recognizing that measurement must account for the relational and contextual dimensions of trust rather than treating it as residing solely in the AI system. Together, these developments suggest the field is moving toward more holistic measurement paradigms, though integration across approaches remains elusive.

Verification and/or Validation. More than half (34 out of 43) of the papers did not (explicitly) provide or adopt verification or validation approaches. Among the remaining 9 papers, surveys are the most widely used verification methodology, appearing in 5 papers (Engelmann, Ullstein, et al. 2022; Inie et al. 2024; Pareek et al. 2024; Schoeffer et al. 2022; B. Zhang and Dafoe 2020). E.g., Pareek et al. (2024) conduct a survey-based between-subject experiments involving 300 participants to investigate trust development, erosion, and recovery during AI-assisted decision-making. Additionally, B. Zhang and Dafoe (2020) pre-register their experiments to ensure methodological rigor and transparency. Lakkaraju and Bastani (2020) and Alpherts et al. (2024) verify trustworthiness measurement through user studies with domain experts. Finally, only one paper employs a mixed method consisting of qualitative interviews and a subsequent survey to collect quantitative data (Scharowski et al. 2023).

The gap in current research, i.e., the insufficient attention to validation under real-world constraints and the lack of alignment between verification practice and trustworthiness dimensions, can be attributed to two primary factors: the lack of real-world data or representative scenarios (Andrus et al. 2021b) and the absence of multi-disciplinary collaboration (Brundage et al. 2020). The 9 papers that do so represent important but limited work in controlled settings, not diverse real-world environments. Our thoughts resonate with Andrus et al. (2021b) and Brand (2023) who have demonstrated challenges in obtaining representative datasets and how siloed approaches can miss critical sociotechnical dimensions for verification or validation.

Table 3. The measurement of AI trustworthiness of the primary studies covered by the review.

Measurement	Type of measure
Likert scale (FAccT: 3)	McKnight et al.[McKnight et al. (2002)] (S. S. Y. Kim, Q. V. Liao, et al. 2024), Jian et al. [Jian et al. (2000)] (Pareek et al. 2024), Perceived trustworthiness (*Own scale*) (Schoeffer et al. 2022)
Reliance behavior (AIES: 1, FAccT: 6)	Change in people's trust judgment (Q. Liao and Sundar 2022), Agreement between a participant's answer and that of the AI system (S. S. Y. Kim, Q. V. Liao, et al. 2024; S. S. Y. Kim, Watkins, et al. 2023; Magaña and Shilton 2025) (Lawrence et al. 2023; Pareek et al. 2024; R. Wang et al. 2024)
Actual cues (e.g., precision, recall, and f1-scores) (AIES: 1, FAccT: 8)	Positive predictive value (E. Kim et al. 2021), LDA topic modelling (Engelmann, M. Chen, et al. 2019), Stochastic model with surrogates (Nannini et al. 2023), (Casper et al. 2024; Kinahan et al. 2024; Mainz et al. 2023; Panigutti et al. 2023; Toreini et al. 2020), Metrics for plausibility and faithfulness scores (Shailya et al. 2025)
Assessment of impact and alignment with values and concerns (AIES: 2, FAccT: 1)	Ethical charters & internal review bodies (Manzini et al. 2024), sociotechnical harms taxonomy, Moral groundings for public transparency (Loi and Spielkamp 2021) (Shelby et al. 2023)
Levels of monitoring activities (FAccT: 3)	Monitoring-avoiding relation between trustor & trustee (Ferrario and Loi 2022) (Toreini et al. 2020), Trustworthiness level across system accuracy ranges (Ferrario 2025)
Communication style and anthropomorphic features (FAccT: 1)	Role-playing & reciprocal engagement (Maeda and Quan-Haase 2024)
Qualitative assessment (AIES: 8, FAccT: 3)	Interviews & surveys (Alpherts et al. 2024; Andrus et al. 2021a; Bhatt et al. 2020; Engelmann, Ullstein, et al. 2022; Inie et al. 2024; Lakkaraju and Bastani 2020; Scharowski et al. 2023; Thornton et al. 2021; R. Wang et al. 2024; B. Zhang and Dafoe 2020), User satisfaction (Zilka et al. 2022)
Model cards and AI assessment catalog (AIES: 1, FAccT: 2)	Information about data leakage & selective inference for EEG (Kinahan et al. 2024), AI assessment catalog (Schmitz 2023), NIST AI risk management framework & other national frameworks (Toney et al. 2024)
Mediator & Trustee role (FAccT: 1)	Social, institutional and temporal embeddedness, Situational normality, Structural assurance, motivation and ABI (Thornton et al. 2022)

4.4 Application Scenarios

AI techniques are applied across domains, including high-stakes ones, making trustworthiness vital to their success. The AIES & FAccT community demonstrates a comprehensive and diverse exploration of trustworthiness across various application scenarios; see Table 4. High-stake domains, including healthcare and medical applications, financial and economic systems, human resources, governance, law, and justice, are the most prominent application scenarios of studying trustworthiness of AI systems (22 out of 43). 12 out of 43 papers do not centre on a specific application scenario but rather generally discuss AI applications and their impact on society. Additionally, Nannini

et al. (2023) and Panigutti et al. (2023) examine AI policies with a focus on explainability of AI systems. Nannini et al. (2023) perform a thematic and gap analysis of policies and standards on explainability in EU, US, and UK and provide a set of recommendations on how to address explainability in regulations for AI systems. Panigutti et al. (2023) analyze the EU AI Act (European Commission 2021) and argue that it neither mandates explainable AI nor bans the use of black-box AI systems; instead, it seeks to achieve its stated policy objectives with a focus on transparency and human oversight.

Table 4. The application scenarios of AI trustworthiness of the primary studies covered by the review.

Application Scenarios	**Representative articles**
Healthcare and medical applications (FAccT: 7)	(Andrus et al. 2021a,b; S. S. Y. Kim, Q. V. Liao, et al. 2024; Kinahan et al. 2024; Lai et al. 2023; Q. Liao and Sundar 2022; Scharowski et al. 2023; Toreini et al. 2020)
Financial and economic systems (AIES: 1, FAccT: 4)	(Andrus et al. 2021a,b; Bhatt et al. 2020; Brand 2023; Scharowski et al. 2023; Schoeffer et al. 2022)
Human resource (AIES: 1, FAccT: 3)	(Andrus et al. 2021a,b; Engelmann, Ullstein, et al. 2022; Mohseni 2019; Scharowski et al. 2023)
Governance, law, and justice (AIES: 9, FAccT: 2)	(Engelmann, M. Chen, et al. 2019; Huang et al. 2020; Knowles and Richards 2021; Lakkaraju and Bastani 2020; Lawrence et al. 2023; Loi and Spielkamp 2021; Mohseni 2019; Schmitz 2023; Williams and Haring 2023; B. Zhang and Dafoe 2020; Zilka et al. 2022)
Content moderation and information systems (FAccT: 2)	(Bhatt et al. 2020; Longoni et al. 2022)
Personal assistance and interaction (FAccT: 3)	(Manikonda et al. 2024; Manzini et al. 2024; Scharowski et al. 2023)
Environmental science (FAccT: 3)	(Alpherts et al. 2024; S. S. Y. Kim, Watkins, et al. 2023; Thornton et al. 2021)
Domain agnostic (AIES: 3, FAccT: 11)	(Casper et al. 2024; Engelmann, Ullstein, et al. 2022; Inie et al. 2024; Jacovi et al. 2021; E. Kim et al. 2021; Knowles, Fledderjohann, et al. 2023; Knowles and Richards 2021; Maeda and Quan-Haase 2024; Mohseni 2019; Pareek et al. 2024; Shelby et al. 2023; Thornton et al. 2022; Toney et al. 2024; Toreini et al. 2020)
AI policies and standards (FAccT: 2)	(Nannini et al. 2023; Panigutti et al. 2023)
Facial recognition (AIES: 1)	(E. Kim et al. 2021)
Software engineering and programming (FAccT: 1)	(R. Wang et al. 2024)

4.5 Intended, Embedded, and Realized Values

Our analysis reveals a significant gap between the intended, embodied, and realized values of AI systems. While many are developed with good intentions, e.g., promoting transparency, fairness, and accountability, these values are not always reflected in the final product or its impact on society.

Intended Values. The most frequently cited intended values were transparency (26), fairness (19), and accountability (12).[5] Other important values included privacy (8), security (5), and human oversight (5). E.g., Bhatt et al. (2020) emphasize the importance of transparency and trustworthiness as the intended values of explainable AI. Similarly, Casper et al. (2024) highlight the role of AI audits in identifying risks and improving transparency, and Engelmann, M. Chen, et al. (2019) examine the intended value of the social credit system to promote honesty and trust in Chinese society. In abidance with the AI act, Panigutti et al. (2023) examine the role of explainable AI and state the intended value to ensure AI is trustworthy by focusing on transparency and human oversight. This includes enabling users to interpret outputs and manage risks associated with AI systems.

Embodied Values. The embodied values in studying AI trustworthiness are often shaped by the techniques used, the stakeholders involved, and the context of deployment. Nannini et al. (2023) examine the embodied value of explainable AI, which is determined by how well it can fulfill its intended purpose without compromising other desiderata, such as accuracy and privacy. Scharowski et al. (2023) examine the embodied value of certification labels, which is determined by how well they can fulfill their intended purpose by reflecting the AI system's compliance with trustworthiness criteria. Alpherts et al. (2024) examine the embodied value of AI models, determined by how well they can generate explanations that help understand the relations between visual elements in street view imagery and socio-economic variables such as housing pricing.

Realized Values. The realized values of AI systems can differ significantly from their intended values due to various factors, including design flaws, unintended consequences, and societal biases. For example, Ferrario and Loi (2022) find that the realized value of explainability in AI systems may differ from its intended value, particularly when explainability does not lead to a reduction in monitoring or when it fosters an unjustified belief in the AI's trustworthiness and S. S. Y. Kim, Watkins, et al. (2023) find that the realized value of AI systems may differ from their intended value due to factors such as malicious manipulation of trustworthiness cues or ill-designed cues that mislead users.

In general, we found 32 articles in our corpus in which the intended values did not directly translate into the realized value. Our analysis identifies 8 major categories of failure reasons: design flaws & technical limitations (8 articles), malicious manipulation & misleading cues (5 articles), sociotechnical gaps & context mismatch (10 articles), methodological limitations (8 articles), compliance-driven vs. values-driven approaches (6 articles), unintended consequences (5 articles), and lack of stakeholder participation (6 articles). The counts overlap because individual articles often identified multiple contributing factors. The most common failure reason was sociotechnical gaps where systems failed to account for broader social, cultural, and institutional contexts.

To further zoom in as an illustrative example, in (Schoeffer et al. 2022) the intended values center on enabling people to assess fairness of automated systems, while the realized value is a narrow technical measure of explanation effectiveness in controlled settings. What participants in the study found adequate may differ substantially from what vulnerable populations facing actual loan denials would need to exercise meaningful agency. This exemplifies how methodological conventions, controlled experiments, convenience samples, perception metrics can systematically exclude the sociotechnical dimensions necessary to realize justice-oriented values in practice. These gaps between intended, embodied, and realized values in AI systems raise important questions about the effectiveness of current approaches to embedding values in AI design for studying AI trustworthiness. They highlights the need for more robust methods to evaluate AI's societal impact.

[5] 39 articles had more than one intended value for studying AI trustworthiness.

5 Discussion: Challenges, Open Questions, and Future Directions

This review has examined the multifaceted nature of AI trustworthiness as articulated within the AIES & FAccT community. This section examines the challenges, open questions, and future directions that arise from our analysis.

5.1 Why Do We Need to Rethink AI Trustworthiness?

Several critical gaps and limitations emerge from our analysis in Section 4.1. Despite this rich conceptual terrain, the current literature inadequately address: who defines trustworthiness criteria, whose harms count as significant, and how user intent is interpreted remain largely unexamined despite ethical considerations being prominent. Additionally, while there is a convergence around components like accuracy and fairness linking to ISO dimensions, the literature largely neglects the power dynamics and structural inequalities that fundamentally shape trust relationships in AI systems (Dobbe et al. 2021).

Moreover, while benchmarks and auditing mechanisms (T2) are emphasized, there is insufficient attention to how compliance-focused approaches might incentivize "ethics washing" where organizations meet procedural requirements without substantively addressing underlying trust deficits. The conceptualizations also struggle with context-dependency: trustworthiness defined through computational reliability may be necessary but insufficient in domains like criminal justice (as noted in T1) where procedural justice and legitimacy matter as much as accuracy, yet the frameworks provide limited guidance on how to weight these competing demands situationally.

Current conceptualizations, despite their evolution from purely technical metrics to sociotechnical constructs, fail to fully account for how trust can erode or strengthen over time, particularly in response to system failures or successful interventions (Dzhelyova et al. 2012). An emphasis on system-level characteristics and user perceptions, while important, has come at the expense of examining broader institutional and systemic factors that influence trustworthiness, such as regulatory frameworks, corporate governance structures, and market incentives. Without addressing these gaps, current approaches to building trustworthy AI risk perpetuating existing biases and power imbalances while failing to establish sustainable trust relationships across user communities (Osasona et al. 2024). This concern is intensified by an observed gradual decline in explicit attention to fairness, bias, and ethics in the broader corpus, as indicated by our initial topic modeling analysis.

Based on our analyses of the corpus, we propose the following measures:

(1) Future research must implement longitudinal studies that track trust dynamics over time, and design systems with mechanisms to rebuild trust after failures rather than focusing solely on initial trust establishment.
(2) At the heart of AI trustworthiness lies a fundamental distinction between perceived and actual trustworthiness (Schlicker, Uhde, et al. 2022). Future research should clearly distinguish them to foster a more unified understanding of AI trustworthiness.
(3) Develop comprehensive trustworthiness frameworks that explicitly incorporate institutional accountability measures, regulatory compliance mechanisms, and transparent corporate governance structures.

5.2 Is AI Trustworthiness Just a Checklist?

A critical analysis of the drivers of AI trustworthiness revealed oversights in how the field currently conceptualizes motivations for studying this crucial aspect. While the reviewed literature identifies several key drivers (see Section 4.2), the understanding appears superficial and fails to address several fundamental concerns.

First, the heavy emphasis on regulatory compliance and ethical guidelines, while important, suggests a potentially reactive rather than proactive approach to trustworthiness (Tyler 2001). Many studies treat regulations as a checklist (see representative articles for D3 in Table 2) rather than engaging with the deeper philosophical and societal implications of AI trustworthiness. This compliance-driven approach risks creating a false sense of security while potentially missing emerging challenges not yet addressed by current regulations, such as the

evolving nature of biases in dynamic systems, the challenge of ensuring explainability in complex models (Saeed and Omlin 2023), attributing accountability in autonomous decision-making (Busuioc 2021), and impacts on human agency and societal power structures (Santoni de Sio and Mecacci 2021). In real-world industry settings, compliance requirements often act as crucial entry points and motivators for responsible AI work (Díaz-Rodríguez et al. 2023). When organizational resources and attention are limited, framing trustworthiness as a compliance issue can secure support and budget allocations (Caldwell and Clapham 2003).

Second, the classification of drivers reveals a bias toward accuracy and promotion of trust in AI, with a limited focus on end-user needs and societal impacts. While "trust" is frequently cited as a driver, there is an inadequate exploration of building an appropriate level of trust in AI, resulting in avoidance of over- and under-trust (Mehrotra, Degachi, et al. 2024; Mehrotra, Jorge, et al. 2024).

Third, we observed considerable vagueness in discussions of AI systems' safety behavior and in the identification and diagnosis of safety risks within complex social contexts. Safety has been designated as a key component of AI trustworthiness both in the EU AI Act (European Commission 2021) and ISO (ISO 2022) & NIST (National Institute of Standards and Technology 2023) guidelines. However, in our analysis, system safety does *not* appear as a prominent measure for trustworthiness except for the mathematical formalism (Toney et al. 2024). This raises two key questions: (i) Is safety marred by an underlying vagueness where it is hard to establish whether a system is safe or not (Dobbe et al. 2021)? And (ii) does it even make sense to study trustworthiness of an AI system without exploring its safety especially from a sociotechnical lens?

Finally, drivers related to long-term sustainability, environmental impact, and social justice are limited in our analyzed papers. While accuracy is listed as a driver, there is limited attention to the complex relationship between technical accuracy and real-world effectiveness. This limited understanding of drivers profoundly impacts how trustworthiness is studied and implemented in AI systems. Without a more nuanced understanding of why trustworthiness matters, the field risks producing solutions that address surface concerns while failing to engage with deeper systemic challenges. Hence, we propose the following measures:

(1) The ideal approach could involve using compliance frameworks as practical starting points while simultaneously cultivating organizational cultures that recognize the broader societal implications of AI systems beyond regulatory requirements.
(2) Develop frameworks and metrics that address the appropriateness of trust, ensuring users understand both the capabilities and limitations of AI systems they interact with.
(3) Future research must expand beyond these conventional drivers of trustworthiness to include societal concerns, power dynamics, and long-term implications for human-AI interaction.

5.3 Measurement Problem in AI Trustworthiness

The critical analysis of measurement, verification, and validation approaches in AI trustworthiness research in Section 4.3 reveals gaps and methodological shortcomings to establish robust frameworks for evaluating trustworthiness of AI systems and build appropriate level of trust. More precisely, this necessitates: (i) conducting evaluations of both actual and perceived trustworthiness, i.e., establishing valid and reliable measurements, (ii) identifying methods to make valid comparisons between these two dimensions, and also (iii) to achieve alignment between them.

First, current research presents a significant gap concerning evaluation of trustworthiness. Prior studies have described the general challenge of aligning actual and perceived trustworthiness (Q. Liao and Sundar 2022; Mehrotra, Degachi, et al. 2024). They have also identified relevant components that constitute actual and perceived trustworthiness (Schlicker, Lechner, et al. 2025). Additionally, previous work has emphasized the necessity for a communication process that conveys information about an AI system and its actual trustworthiness to shape and align trustworthiness perceptions (Q. Liao and Sundar 2022). However, we still lack an understanding of

how the evaluation of actual trustworthiness and the successful alignment of this actual trustworthiness with perceived trustworthiness would manifest in practice, based on concrete measures and quantitative values for trustworthiness characteristics both for the AI system and user perceptions.

Second, when trustworthiness is measured using only questionnaires or reliance behavior, there is still uncertainty about the trustee's actual trustworthiness (Schlicker, Baum, et al. 2025). Robust methodologies that incorporate objective measures of trustworthiness alongside subjective assessments are essential for developing a comprehensive understanding of trustworthiness in human-AI interaction. However, it is still complex to consider that actual and perceived trustworthiness combined together provides a clear overview of the AI system's overall trustworthiness, perhaps the $2 + 2 = 4$ analogy is not a fit here. We believe the issue is there is no single correct way of doing this, *i.e.*, making individual characteristics commensurable. There is no ground truth to questions like what is the objective trustworthiness value for a given performance or explainability score of an AI system (Mainz et al. 2023).

The third problem differs from the first two in a fundamental way. Rather than dealing with how to combine or compare trustworthiness measures and individual traits, it addresses a more basic question: which specific characteristics should be evaluated when assessing trustworthiness in the first place? The challenge here is not about creating compatible measurement scales. Instead, it involves establishing comparable sets of characteristics between two sides, the AI system itself and the users who perceive it. Critical questions arise from this challenge: Which performance metrics should be considered? What aspects of transparency matter? How should explainability be evaluated? What privacy measures and safety standards should factor into the assessment of AI system trustworthiness? These questions are all fundamentally connected to this obstacle.

The verification and validation issues are even more concerning, with nearly half of the papers failing to address these crucial aspects. For example, what is the source of cues that inform users about an AI system's trustworthiness and how can we verify or validate it? One perspective holds that these cues should originate directly from the systems and their creators. Some researchers advocate for defining good cues that technology developers should implement (Q. Liao and Sundar 2022). These same scholars acknowledge that such indicators must be accurate and honest to be effective. When examining practical applications, they assume this honesty requirement is already met by creators and focus their analysis on additional necessary conditions. This assumption proves problematic in real-world scenarios. System providers cannot simply be expected to present entirely accurate cues about their products' limitations. Commercial entities operate with sales objectives as their primary driver. When users place excessive confidence in a product beyond what is justified, achieving proper confidence calibration would require the provider to communicate information that reduces trust. Yet commercial logic creates a strong disincentive for broadcasting unfavorable information; doing so would directly harm revenue and competitive positioning. It seems unlikely, therefore, that providers would intentionally engineer their systems to broadcast cues indicating mediocre or poor trustworthiness to potential users.

We propose the following measures for effective measurement, validation and verification:

(1) Our analysis suggests a pressing need for AI system designers and developers to explain their rationale for choosing particular trustworthiness attributes and specific metrics to measure those attributes. They also need to justify their methodology for prioritizing and combining these elements to generate a composite trustworthiness score. In doing so, they should be required to acknowledge their inherent perspectives and the criteria they apply when judging an AI system's trustworthiness.
(2) Greater focus on trustworthiness assessment mechanisms within the trust development process would likely enhance future research efforts to clarify how perceived trustworthiness transforms into actual trust and subsequent trusting behaviors.
(3) Those who provide the cues to the AI system users need to elaborate on and give reasons why certain characteristics of the AI system and trustworthiness cues warrant trust. Also, focus on understanding

how cues are relevant for the users, how they detect and utlize it forms a key element for verification and validation of the trustworthiness assessment.

5.4 Values Gap in AI Trustworthiness

Our critical examination of the values in AI trustworthiness research reveals significant oversights and problematic assumptions that limit our understanding of how values materialize in practice. The analysis of intended values demonstrates a concerning preoccupation with surface-level technical attributes such as transparency and fairness, while giving insufficient attention to deeper systemic values such as social justice, environmental sustainability, and cultural preservation. This narrow focus reflects a persistent bias toward engineering solutions rather than addressing fundamental societal challenges (Olteanu et al. 2019). The heavy emphasis on transparency risks becoming a performative gesture rather than a meaningful commitment to openness and accountability, thereby risking to entrench problematic capture of the institutions that are meant to support public values (Gansky and McDonald 2022; Whittaker 2021).

The discussion of embodied values highlights weaknesses in current research. The literature often oversimplifies how values are embedded in technical systems, reducing complex social and ethical considerations to fit technical implementations. This simplification risks a false sense of progress while neglecting the deeper challenges of value implementation (Stern et al. 2002). Particularly concerning is the treatment of realized values, where the analysis lacks a systematic framework for understanding value divergence. The literature notes discrepancies between intended and realized values but inadequately theorizes these gaps, attributing them to superficial issues like "design flaws" or "unintended consequences" without addressing structural barriers that hinder the realization of intended values.

The conclusion that a "more holistic approach" is required appears insufficient given the complexity of the challenges at hand. We need rigorous, dynamic frameworks reflecting the contested and evolving nature of values in AI systems. Such frameworks must move beyond static, universal notions of values to address how they are negotiated, implemented, and evaluated in real-world contexts, confronting structures of power. E.g., these power structures manifest in multiple dimensions that trustworthiness research must address: (i) corporate concentrations of AI development capacity that privilege certain stakeholders' definitions of "trustworthiness," (ii) institutional hierarchies within organizations that determine whose values and concerns shape AI systems, (iii) socio-economic disparities that affect who benefits from or bears risks of AI deployment, and (iv) geopolitical imbalances that enable dominant nations to set global AI governance norms, and professional authority structures that privilege technical expertise over lived experience.

We recommend the following measures to reduce the values gap in AI trustworthiness:

(1) Implementing participatory design methodologies that meaningfully involve marginalized communities in defining trustworthiness criteria as informed by Harrington et al. (2019).
(2) Developing evaluation frameworks that assess differential impacts across diverse populations rather than assuming universal benefits (Carey and Crammond 2017).
(3) Establishing transparency mechanisms that reveal how power influences AI system development decisions and provide mechanisms to contest them as showcased by Ehsan et al. (2021) and Mehrotra, Gadiraju, et al. (2025).
(4) Creating accountability structures that redistribute decision-making authority beyond technical experts (Busuioc 2021) and recognizing that trustworthiness is a contested concept shaped by those with the power to define it (Hardin 2002).

5.5 Operationalizing Recommendations: Use Cases and Transformations

In the four subsections before the present subsection, we have proposed general recommendations and measures for future directions from a sociotechnical perspective. Concretely, these recommendations can be deployed across multiple AI application domains. To make them more actionable for researchers working from technical perspectives, such as machine learning practitioners, we provide several illustrative examples across diverse applications: AI-powered clinical decision support systems and conversational agents for caregiver mental well-being support (AI in healthcare), AI-powered hiring/recruitment systems (AI in recruitment), criminal risk assessment algorithms (AI in criminal justice), and AI-powered adaptive learning platforms and automated assessment systems (AI in education). Table 5 demonstrates how current AI systems are defined, built, and evaluated in these use cases, and illustrates the transformative changes that our proposed recommendations and measures could bring to each domain.

Table 5. Comparison of current AI system practices and practices informed by our proposed recommendations: Illustrative use cases shows how our recommendations address limitations in current practices.

• *Application* AI-powered clinical decision support system
• *Current system* Evaluate overall accuracy metrics (sensitivity/specificity) on aggregated test datasets (Hicks et al. 2022); Measure trust through post-deployment user surveys with general practitioners (M. Chen et al. 2022); Apply fairness metrics (e.g., demographic parity, equalized odds) to measure performance disparities across protected attributes like race and gender (R. J. Chen et al. 2023); Assess system performance using standardized benchmark datasets primarily from well-resourced healthcare settings (Blagec et al. 2023).
• *What our recommendations bring* **Participatory measurement design**: Co-develop trustworthiness criteria with patients from underserved communities (e.g., rural populations, minorities with higher disease prevalence) to identify what "trustworthy screening" means beyond accuracy, including accessibility, cultural sensitivity, and explanation comprehensibility; **Differential impact evaluation**: Stratify performance metrics across race, socioeconomic status, age, and geographic location to reveal disparities (e.g., higher false negative rates in darker skin tones due to training data bias); **Power-aware transparency**: Document and disclose how decisions about training data selection, feature prioritization, and deployment locations were made, revealing potential bias toward well-resourced hospitals; **Redistributed evaluation authority**: Include community health workers, patients, and advocacy groups in defining evaluation success criteria alongside ophthalmologists and data scientists; **Context-sensitive trustworthiness**: Recognize that trustworthiness criteria differ; rural clinic practitioners may prioritize offline functionality and low-bandwidth operation, while medical centers focus on integration with their existing systems.
• *Application* Conversational agents for caregiver mental well-being support
• *Current system* Maintain system safety mainly through content moderation filters (Sarkar et al. 2023); Measure user engagement metrics (session duration, return rate, sentiment scores) (Limpanopparat et al. 2024); Assess conversational quality using expert annotations and standardized mental health screening tools (Dosovitsky et al. 2021); Conduct one-time user satisfaction surveys post-interaction (Limpanopparat et al. 2024); Focus on clinical efficacy benchmarks and harm prevention protocols (Sarkar et al. 2023).

• *What our recommendations bring*

Longitudinal trust dynamics: Track how caregiver trust evolves across crisis moments (e.g., after receiving unhelpful advice during acute stress) and design explicit trust repair mechanisms, such as human handoff protocols when the system fails to provide adequate support; **Perceived vs. actual trustworthiness**: Distinguish between caregivers' confidence in the system (perceived) and its actual clinical appropriateness, measuring whether users over-rely on the agent; **Institutional accountability frameworks**: Establish clear governance structures defining who is responsible when the system provides harmful advice, including regulatory compliance mechanisms, clinical oversight boards, and transparent corporate policies on data usage and care escalation; **Appropriateness of trust calibration**: Develop frameworks that help caregivers understand when to trust the AI (e.g., emotional validation, stress management techniques) versus when to seek human professionals (e.g., suicidal ideation, severe depression), with explicit system boundaries.

• *Application*

AI-powered hiring/recruitment systems

• *Current system*

Evaluate prediction accuracy by measuring how well AI selections match historical hiring decisions and subsequent employee performance (Black and Esch 2020); Assess system efficiency in filtering large applicant pools (Z. Chen 2023); Apply minimal fairness audits primarily to meet legal compliance requirements (Commission et al. 2023).

• *What our recommendations bring*

Participatory measurement design: Co-develop trustworthiness criteria with job seekers from underrepresented groups to identify what "fair screening" means beyond demographic parity; **Power-aware transparency**: Document and disclose whose definition of "qualified candidate" is encoded in the system, revealing whether criteria reflect genuine job requirements or historical biases of previous hiring managers, and how employer preferences versus candidate needs are balanced; **Context-sensitive trustworthiness**: Recognize that trustworthiness criteria differ by stakeholder: employers may prioritize efficiency and cultural fit, while candidates prioritize fairness and transparency in rejection reasons, and marginalized communities may prioritize systems that challenge rather than perpetuate historical exclusion patterns.

• *Application*

Criminal risk assessment algorithms

• *Current system*

Evaluate prediction accuracy by measuring how well risk scores correlate with actual re-arrest or re-offense rates (Dressel and Farid 2018); Conduct post-hoc bias and fairness audits when systems face legal challenges or public scrutiny (Angwin et al. 2022); Measure system adoption rates and judicial satisfaction with decision support (Stevenson and Doleac 2024).

• *What our recommendations bring*

Differential impact evaluation: Assess not just classification accuracy but downstream consequences across racial and socioeconomic groups, including cumulative disadvantage from pretrial detention, access to rehabilitation programs, employment prospects post-release, and intergenerational community impacts of mass incarceration; **Power-aware transparency**: Document and disclose whose conception of "risk" and "public safety" is embedded in the system, revealing whether models perpetuate historical policing patterns that over-surveilled marginalized communities, how feature selection reflects prosecutorial versus defense perspectives, and which stakeholder interests (law enforcement efficiency vs. defendant rights) dominate design decisions; **Redistributed evaluation authority**: Include criminal justice reform advocates, affected community members, defense attorneys, and social workers in defining evaluation success criteria alongside prosecutors.

• *Application*

AI-powered adaptive learning platforms and automated assessment systems

• *Current system*

Assess prediction accuracy of student performance and dropout risk models; Measure engagement metrics (time-on-task, interaction frequency, content completion) (L. Y. Tan et al. 2025); Conduct user satisfaction surveys with teachers and administrators (Strielkowski et al. 2025); Optimize for institutional metrics like pass rates and graduation rates (Du Plooy et al. 2024).

• *What our recommendations bring*

Longitudinal trust dynamics: Track how student and teacher trust in the system evolves across academic failures or successes, designing trust repair mechanisms when the system provides ineffective learning paths or inaccurate assessments; **Differential impact evaluation**: Assess not just average learning gains but stratify outcomes across socioeconomic status, disability status, language background, and prior educational opportunity, revealing whether the system narrows or widens achievement gaps, and whether it channels certain populations toward vocational tracks while others receive enrichment; **Power-aware transparency**: Document and disclose whose pedagogical philosophy is embedded in the system, revealing whether learning objectives reflect standardized testing priorities versus holistic education; **Appropriateness of trust calibration**: Develop frameworks helping students and teachers understand when to trust AI recommendations (e.g., practice problem selection, pacing suggestions) versus when to override them (e.g., when algorithms misinterpret struggle as inability rather than productive challenge, or when standardized paths ignore individual passions and strengths).

Our proposed recommendations represent the first step rather than the end of achieving trustworthiness of AI systems in real-world applications. Realizing this vision requires sustained, collaborative efforts from researchers, practitioners, and stakeholders across multiple disciplines.

5.6 Limitations

First, our focus on AIES & FAccT conferences reflects a particular scholarly perspective. To address this and build a more comprehensive understanding of how trustworthiness is conceptualized across AI research landscape, we propose extending our analysis to major technical AI conferences such as NeurIPS, ICML, IJCAI, and AAAI. This expanded corpus would enable comparison between how trustworthiness is framed in technically-focused venues versus interdisciplinary ones, potentially revealing important differences in priorities, metrics, evaluation methods, and theoretical frameworks. For example, the work by B. Wang et al. (2023) exemplifies the high technical rigor of the computer science community, particularly within the dimensions of fairness (T4) and technical robustness and reliability (T6) (see Table 1), but puts less emphasis on perceived trustworthiness and broader sociotechnical considerations emphasized in ISO standards.

Second, our focus on academic discourse may not adequately represent how trustworthiness is understood by other key stakeholders, including industry practitioners, policymakers, and diverse user communities. These groups may prioritize different aspects of trustworthiness based on their contexts and concerns.

Third, our methodological focus on papers that operationalize trustworthiness through specific metrics or models represents one analytical approach to understanding this concept. An alternative approach for future work could examine theoretical papers that discuss the nature and foundations of trustworthiness without necessarily proposing measurement frameworks. Such papers may offer important conceptual contributions, philosophical perspectives, and normative arguments that complement operationalized approaches. This dual-lens approach might reveal whether theoretical and technical communities prioritize different dimensions of trustworthiness and how these perspectives could inform each other.

Finally, cultural and geographical limitations exist in our analysis, as both AIES & FAccT have historically featured stronger representation from Western institutions. This may result in overlooking important cultural variations in how trustworthiness is conceptualized across societies and knowledge traditions.

6 Conclusion

This study underscores the growing importance of trustworthiness in AI systems as a foundation for ethical and reliable technology. Through a review of AIES & FAccT conference proceedings, we analyzed how trustworthiness is conceptualized, measured, and validated. While progress has been made in defining attributes such as transparency, accountability, and robustness, significant gaps remain, particularly in addressing the sociotechnical nature of AI and integrating safety into trustworthiness discussions.

Our findings highlight the need for an interdisciplinary approach that combines technical precision with social and ethical considerations. Current research often prioritizes technical attributes, overlooking the complex interplay between AI systems and the broader social, cultural, and institutional contexts in which they operate. Addressing these gaps is crucial for the AIES & FAccT community to drive a paradigm shift, advancing holistic approaches to trustworthy AI that genuinely benefit society and promote responsible technological development.

Acknowledgments

We thank our anonymous reviewers for providing valuable suggestions to improve the manuscript.

This research was (partially) supported by the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.1389.20.183, and KICH3.LTP.20.006, and the European Union under grant agreements No. 101070212 (FINDHR), No. 101201510 (UNITE), and No. 101203728 (SOCIALADAPT).

Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of their respective employers, funders and/or granting authorities.

References

D. E. Acuna and L. Liang. 2021. "Are AI Ethics Conferences Different and More Diverse Compared to Traditional Computer Science Conferences?" In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '21). Association for Computing Machinery, Virtual Event, USA, 307–315. ISBN: 9781450384735. doi:10.1145/3461702.3462616.

L. H. Ajmani, S. Chancellor, B. Mehta, C. Fiesler, M. Zimmer, and M. De Choudhury. 2023. "A Systematic Review of Ethics Disclosures in Predictive Mental Health Research." In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '23). Association for Computing Machinery, Chicago, IL, USA, 1311–1323. ISBN: 9798400701924. doi:10.1145/3593013.3594082.

K. I. Alkhouri. 2024. "The Role of Artificial Intelligence in the Study of the Psychology of Religion." *Religions*, 15, 3, 290.

T. Alpherts, S. Ghebreab, Y.-C. Hsu, and N. van Noord. 2024. "Perceptive Visual Urban Analytics is Not (Yet) Suitable for Municipalities." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1341–1354. doi:10.1145/3630106.3658976.

M. Andrus, E. Spitzer, J. Brown, and A. Xiang. 2021a. "What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness." In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21). Association for Computing Machinery, Virtual Event, Canada, 249–260. ISBN: 9781450383097. doi:10.1145/3442188.3445888.

M. Andrus, E. Spitzer, J. Brown, and A. Xiang. 2021b. "What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 249–260. doi:10.1145/3442188.3445888.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. 2022. "Machine Bias." In: *Ethics of Data and Analytics*. Auerbach Publications, 254–264.

S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling. 2019. "Explainable Agents and Robots: Results from a Systematic Literature Review." In: *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*. International Foundation for Autonomous Agents and Multiagent Systems, 1078–1088.

H. Arksey and L. O'Malley. 2005. "Scoping Studies: Towards a Methodological Framework." *International Journal of Social Research Methodology*, 8, 1, 19–32.

M. D. Baer and J. A. Colquitt. 2018. "Why Do People Trust?: Moving toward a More Comprehensive Consideration of the Antecedents of Trust." *The Routledge Companion to Trust*, 163–182.

A. Balmer. 2023. "A Sociological Conversation with ChatGPT about AI Ethics, Affect and Reflexivity." *Sociology*, 57, 5, 1249–1258.

V. R. Basili, G. Caldiera, and H. D. Rombach. 1994. "The Goal Question Metric Approach." *Encyclopedia of Software Engineering*, 528–532.

M. Benk, S. Kerstan, F. von Wangenheim, and A. Ferrario. 2024. "Twenty-four Years of Empirical Research on Trust in AI: A Bibliometric Review of Trends, Overlooked Issues, and Future Directions." *AI & Society*, 1–24.

U. Bhatt et al.. 2020. "Explainable Machine Learning in Deployment." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 648–657. doi:10.1145/3351095.3375624.

J. S. Black and P. van Esch. 2020. "AI-enabled Recruiting: What Is It and How Should a Manager Use It?" *Business Horizons*, 63, 2, 215–226.
K. Blagec, J. Kraiger, W. Frühwirt, and M. Samwald. 2023. "Benchmark Datasets Driving Artificial Intelligence Development Fail to Capture the Needs of Medical Professionals." *Journal of Biomedical Informatics*, 137, 104274.
J. Brand. 2023. "Exploring the Moral Value of Explainable Artificial Intelligence Through Public Service Postal Banks." In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '23). Association for Computing Machinery, Montréal, QC, Canada, 990–992. ISBN: 9798400702310. doi:10.1145/3600211.3604741.
M. Brundage et al.. 2020. "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." *arXiv preprint arXiv:2004.07213*.
M. Busuioc. 2021. "Accountable Artificial Intelligence: Holding Algorithms to Account." *Public Administration Review*, 81, 5, 825–836.
C. Caldwell and S. E. Clapham. 2003. "Organizational Trustworthiness: An International Perspective." *Journal of Business Ethics*, 47, 349–364.
G. Carey and B. Crammond. 2017. "A Glossary of Policy Frameworks: The Many Forms of 'Universalism'and Policy 'Targeting'." *J Epidemiol Community Health*, 71, 3, 303–307.
S. Casper et al.. 2024. "Black-Box Access is Insufficient for Rigorous AI Audits." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2254–2272. doi:10.1145/3630106.3659037.
M. Chen et al.. 2022. "Acceptance of Clinical Artificial Intelligence among Physicians and Medical Students: A Systematic Review with Cross-Sectional Survey." *Frontiers in Medicine*, 9, 990604.
R. J. Chen, J. J. Wang, D. F. Williamson, T. Y. Chen, J. Lipkova, M. Y. Lu, S. Sahai, and F. Mahmood. 2023. "Algorithmic Fairness in Artificial Intelligence for Medicine and Healthcare." *Nature Biomedical Engineering*, 7, 6, 719–742.
Z. Chen. 2023. "Collaboration among Recruiters and Artificial Intelligence: Removing Human Prejudices in Employment." *Cognition, Technology & Work*, 25, 1, 135–149.
J. Chun and K. Elkins. 2023. "The Crisis of Artificial Intelligence: A New Digital Humanities Curriculum for Human-Centred AI." *International Journal of Humanities and Arts Computing*, 17, 2, 147–167.
M. Coeckelbergh. 2022. *The Political Philosophy of AI: An Introduction*. John Wiley & Sons.
E. E. O. Commission et al.. 2023. *Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964.(2023)*. (2023).
J. Dacon. 2023. "Are You Worthy of My Trust?: A Socioethical Perspective on the Impacts of Trustworthy AI Systems on the Environment and Human Society." *arXiv preprint arXiv:2309.09450*.
J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint arXiv:1810.04805*.
N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. L. de Prado, E. Herrera-Viedma, and F. Herrera. 2023. "Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation." *Information Fusion*, 99, 101896.
G. Dietz and D. N. Den Hartog. 2006. "Measuring Trust inside Organisations." *Personnel Review*, 35, 5, 557–588.
G. DNV. 2015. "ISO 9001: 2015 Quality Management Systems-Requirements." *Guidance Document*.
R. Dobbe, T. K. Gilbert, and Y. Mintz. 2021. "Hard Choices in Artificial Intelligence." *Artificial Intelligence*, 300, 103555.
G. Dosovitsky, E. Kim, and E. L. Bunge. 2021. "Psychometric Properties of a Chatbot Version of the PHQ-9 with Adults and Older Adults." *Frontiers in Digital Health*, 3, 645805.
J. Dressel and H. Farid. 2018. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances*, 4, 1, eaao5580.
E. Du Plooy, D. Casteleijn, and D. Franzsen. 2024. "Personalized Adaptive Learning in Higher Education: A Scoping Review of Key Characteristics and Impact on Academic Performance and Engagement." *Heliyon*, 10, 21.
M. Dzhelyova, D. I. Perrett, and I. Jentzsch. 2012. "Temporal Dynamics of Trustworthiness Perception." *Brain research*, 1435, 81–90.
U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz. 2021. "Expanding Explainability: Towards Social Transparency in AI Systems." In: *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–19.
S. Engelmann, M. Chen, F. Fischer, C.-y. Kao, and J. Grossklags. 2019. "Clear Sanctions, Vague Rewards: How China's Social Credit System Currently Defines "Good" and "Bad" Behavior." *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 69–78. doi:10.1145/3287560.3287585.
S. Engelmann, C. Ullstein, O. Papakyriakopoulos, and J. Grossklags. 2022. "What People Think AI Should Infer From Faces." *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 128–141. doi:10.1145/3531146.3533080.
European Commission. 2021. "Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts." https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.
European Union HLEG. 2019. *Ethics Guidelines for Trustworthy Artificial Intelligence (AI)*. Accessed: 2024-01-11. (2019). https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.
Executive Office of the President. 2020. *Presidential Executive Order 13960*. Accessed: 2024-01-11. (2020). https://www.federalregister.gov/documents/2020/02/14/2020-02544/e-o-13960-maintaining-american-leadership-in-artificial-intelligence.
A. Ferrario. 2025. "A Trustworthiness-based Metaphysics of Artificial Intelligence Systems." In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 1360–1370.

A. Ferrario and M. Loi. 2022. "How Explainability Contributes to Trust in AI." In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22). Association for Computing Machinery, Seoul, Republic of Korea, 1457–1466. ISBN: 9781450393522. doi:10.1145/3531146.3533202.

A. Fiore. 2024. "Is Dewey's Aesthetics Critical? A Reflection on the Relationship Between Artificial Intelligence and Art from a Deweyan Perspective." *Contemporary Pragmatism*, 21, 4, 381–398. doi:10.1163/18758185-bja10096.

L. Floridi et al.. 2018. "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines*, 28, 689–707.

S. A. Friedler and C. Wilson, (Eds.) . 2018. *FAT* '18: Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA. https://facctconference.org/2018/.

D. C. Funder. 1995. "On the Accuracy of Personality Judgment: A Realistic Approach." *Psychological Review*, 102, 4, 652.

J. Furman, G. Marchant, H. Price, and F. Rossi, (Eds.) . 2018. *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, USA. ISBN: 9781450360128. https://dl.acm.org/doi/proceedings/10.1145/3278721.

B. Gansky and S. McDonald. 2022. "CounterFAccTual: How FAccT Undermines Its Organizing Principles." In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22). Association for Computing Machinery, Seoul, Republic of Korea, 1982–1992. ISBN: 9781450393522. doi:10.1145/3531146.3533241.

R. Gillis, J. Laux, and B. Mittelstadt. 2024. "Trust and Trustworthiness in Artificial Intelligence." In: *Handbook on Public Policy and Artificial Intelligence*. Edward Elgar Publishing, 181–193.

M. Grootendorst. 2022. "BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure." *arXiv preprint arXiv:2203.05794*.

M. Hagan. 2018. "A Human-centered Design Approach to Access to Justice: Generating New Prototypes and Hypotheses for Interventions to Make Courts User-friendly." *Indiana Journal of Law and Social Equality*, 6, 199.

M. Hagan. 2020. "Legal Design as a Thing: A Theory of Change and a Set of Methods to Craft a Human-centered Legal System." *Design Issues*, 36, 3, 3–15.

R. Hardin. 2002. *Trust and Trustworthiness*. Russell Sage Foundation.

C. Harrington, S. Erete, and A. M. Piper. 2019. "Deconstructing Community-based Collaborative Design: Towards More Equitable Participatory Design Engagements." In: *Proceedings of the ACM on Human-Computer Interaction* CSCW. Vol. 3. ACM New York, NY, USA, 1–25.

Y. He. 2024. "Artificial Intelligence and Socioeconomic Forces: Transforming the Landscape of Religion." *Humanities and Social Sciences Communications*, 11, 1, 1–10.

S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, and S. Parasa. 2022. "On Evaluation Metrics for Medical Applications of Artificial Intelligence." *Scientific Reports*, 12, 1, 5979.

L. Huang, J. Wei, and E. Celis. 2020. "Towards Just, Fair and Interpretable Methods for Judicial Subset Selection." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 293–299. doi:10.1145/3375627.3375848.

N. Inie, S. Druga, P. Zukerman, and E. M. Bender. 2024. "From "AI" to Probabilistic Automation: How Does Anthropomorphization of Technical Systems Descriptions Influence Trust?" *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2322–2347. doi:10.1145/3630106.3659040.

ISO. 2022. *Trustworthiness — Vocabulary*. Accessed 2023-11-20. (2022). https://www.iso.org/obp/ui/#iso:std:iso-iec:ts:5723:ed-1:v1:en.

A. Z. Jacobs and H. Wallach. 2021. "Measurement and Fairness." In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385.

A. Jacovi, A. Marasović, T. Miller, and Y. 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.

Japan Expert Group on How AI Principles Should be Implemented. 2022. *Governance Guidelines for Implementation of AI Principles*. Accessed: 2024-01-11. (2022). https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf.

J.-Y. Jian, A. M. Bisantz, and C. G. Drury. 2000. "Foundations for an Empirically Determined Scale of Trust in Automated Systems." *International Journal of Cognitive Ergonomics*, 4, 1, 53–71. eprint: https://doi.org/10.1207/S15327566IJCE0401_04. doi:10.1207/S15327566IJCE0401_04.

M. Kastner et al.. 2012. "What is the Most Appropriate Knowledge Synthesis Method to Conduct a Review? Protocol for a Scoping Review." *BMC Medical Research Methodology*, 12, 1–10.

M. Kattnig, A. Angerschmid, T. Reichel, and R. Kern. 2024. "Assessing Trustworthy AI: Technical and Legal Perspectives of Fairness in AI." *Computer Law & Security Review*, 55, 106053.

D. Kaur, S. Uslu, and A. Durresi. 2021. "Requirements for Trustworthy Artificial Intelligence–A Review." In: *Advances in Networked-Based Information Systems: The 23rd International Conference on Network-Based Information Systems (NBiS-2020) 23*. Springer, 105–115.

C. Kelp and M. Simion. 2023. "What is Trustworthiness?" *Noûs*, 57, 3, 667–683.

E. Kim, D. Bryant, D. Srikanth, and A. Howard. 2021. "Age Bias in Emotion Detection: An Analysis of Facial Emotion Recognition Performance on Young, Middle-Aged, and Older Adults." In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '21). Association for Computing Machinery, Virtual Event, USA, 638–644. ISBN: 9781450384735. doi:10.1145/3461702.3462609.

S. S. Y. Kim, Q. V. Liao, M. Vorvoreanu, S. Ballard, and J. W. Vaughan. 2024. ""I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 822–835. doi:10.1145/3630106.3658941.

S. S. Y. Kim, E. A. Watkins, O. Russakovsky, R. Fong, and A. Monroy-Hernández. 2023. "Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application." *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 77–88. doi:10.1145/3593013.3593978.

S. Kinahan, P. Saidi, A. Daliri, J. Liss, and V. Berisha. 2024. "Achieving Reproducibility in EEG-Based Machine Learning." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1464–1474. doi:10.1145/3630106.3658983.

B. Knowles, J. Fledderjohann, J. T. Richards, and K. R. Varshney. 2023. "Trustworthy AI and the Logics of Intersectional Resistance." *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 172–182. doi:10.1145/3593013.3593986.

B. Knowles and J. T. Richards. 2021. "The Sanction of Authority: Promoting Public Trust in AI." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 262–271. doi:10.1145/3442188.3445890.

I. Kusche. 2024. "Possible Harms of Artificial Intelligence and the EU AI act: Fundamental Rights and Risk." *Journal of Risk Research*, 1–14. doi:10.1080/13669877.2024.2350720.

C. Lahusen, M. Maggetti, and M. Slavkovik. 2024. "Trust, Trustworthiness and AI Governance." *Scientific Reports*, 14, 1, 20752.

V. Lai, C. Chen, A. Smith-Renner, Q. V. Liao, and C. Tan. 2023. "Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies." *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1369–1385. doi:10.1145/3593013.3594087.

H. Lakkaraju and O. Bastani. 2020. ""How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 79–85. doi:10.1145/3375627.3375833.

B. Laufer, S. Jain, A. F. Cooper, J. Kleinberg, and H. Heidari. 2022. "Four Years of FAccT: A Reflexive, Mixed-Methods Analysis of Research Contributions, Shortcomings, and Future Prospects." In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22). Seoul, Republic of Korea, 401–426.

C. Lawrence, I. Cui, and D. Ho. 2023. "The Bureaucratic Challenge to AI Governance: An Empirical Assessment of Implementation at U.S. Federal Agencies." *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 606–652. doi:10.1145/3600211.3604701.

J. D. Lee and K. A. See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors*, 46, 1, 50–80.

B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou. Jan. 2023. "Trustworthy AI: From Principles to Practices." *ACM Comput. Surv.*, 55, 9, Article 177, (Jan. 2023), 46 pages. doi:10.1145/3555803.

Q. Liao and S. S. Sundar. 2022. "Designing for Responsible Trust in AI Systems: A Communication Perspective." In: *2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22). Association for Computing Machinery, Seoul, Republic of Korea, 1257–1268. ISBN: 9781450393522. doi:10.1145/3531146.3533182.

D. Lim. 2018. "AI & IP: Innovation & Creativity in an Age of Accelerated Change." *Akron Law Review*, 52, 813.

S. Limpanopparat, E. Gibson, and A. Harris. 2024. "User Engagement, Attitudes, and the Effectiveness of Chatbots as a Mental Health Intervention: A Systematic Review." *Computers in Human Behavior: Artificial Humans*, 2, 2, 100081.

Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. G. H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li. 2023. "Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment." *arXiv preprint arXiv:2308.05374*.

M. Loi and M. Spielkamp. 2021. "Towards Accountability in the Use of Artificial Intelligence for Public Administrations." *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 757–766. doi:10.1145/3461702.3462631.

C. Longoni, A. Fradkin, L. Cian, and G. Pennycook. 2022. "News from Generative Artificial Intelligence Is Believed Less." *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 97–106. doi:10.1145/3531146.3533077.

T. Maeda and A. Quan-Haase. 2024. "When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1068–1077. doi:10.1145/3630106.3658956.

M. I. Magaña and K. Shilton. 2025. "Frameworks, Methods and Shared Tasks: Connecting Participatory AI to Trustworthy AI Through a Systematic Review of Global Projects." In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2166–2179.

J. Mainz, L. Munch, and J. C. Bjerring. 2023. "Two Reasons for Subjecting Medical AI Systems to Lower Standards than Humans." *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 44–49. doi:10.1145/3593013.3593975.

B. F. Malle. 2022. "Beyond Fairness and Explanation: Foundations of Trustworthiness of Artificial Agents." In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '22). Association for Computing Machinery, Oxford, United Kingdom, 4. ISBN: 9781450392471. doi:10.1145/3514094.3539570.

L. Manikonda, C. Fosco, and E. Gilbert. 2024. "When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1068–1081. doi:10.1145/3630106.3658956.

A. Manzini, G. Keeling, N. Marchal, K. R. McKee, V. Rieser, and I. Gabriel. 2024. "Should Users Trust Advanced AI Assistants? Justified Trust As a Function of Competence and Alignment." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1174–1186. https://doi.org/10.1145/3630106.3658964.

R. C. Mayer, J. H. Davis, and F. D. Schoorman. 1995. "An Integrative Model of Organizational Trust." *Academy of Management Review*, 20, 3, 709–734.

D. H. McKnight, V. Choudhury, and C. Kacmar. 2002. "The Impact of Initial Consumer Trust on Intentions to Transact with a Web Site: A Trust Building Model." *The Journal of Strategic Information Systems*, 11, 3, 297–323. doi:https://doi.org/10.1016/S0963-8687(02)00020-3.

S. Mehrotra, C. Degachi, O. Vereschak, C. M. Jonker, and M. L. Tielman. Nov. 2024. "A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges." *ACM J. Responsib. Comput.*, 1, 4, Article 26, (Nov. 2024), 45 pages. doi:10.1145/3696449.

S. Mehrotra, U. Gadiraju, E. Bittner, F. van Delden, C. M. Jonker, and M. L. Tielman. 2025. ""Even explanations will not help in trusting [this] fundamentally biased system": A Predictive Policing Case-Study." In: *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization* (UMAP '25). Association for Computing Machinery, NY, USA, 51–62. ISBN: 9798400713132. doi:10.1145/3699682.3728343.

S. Mehrotra, C. M. Jonker, and M. L. Tielman. 2021. "More Similar Values, More Trust? - the Effect of Value Similarity on Trust in Human-Agent Interaction." In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '21). Association for Computing Machinery, Virtual Event, USA, 777–783. ISBN: 9781450384735. doi:10.1145/3461702.3462576.

S. Mehrotra, C. C. Jorge, C. M. Jonker, and M. L. Tielman. Jan. 2024. "Integrity-based Explanations for Fostering Appropriate Trust in AI Agents." *ACM Trans. Interact. Intell. Syst.*, 14, 1, Article 4, (Jan. 2024), 36 pages. doi:10.1145/3610578.

S. Mohseni. 2019. "Toward Design and Evaluation Framework for Interpretable Machine Learning Systems." *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 553–554. doi:10.1145/3306618.3314322.

B. Mucsányi, M. Kirchhof, E. Nguyen, A. Rubinstein, and S. J. Oh. 2023. "Trustworthy Machine Learning." *arXiv preprint arXiv:2310.08215*.

L. Nannini, A. Balayn, and A. L. Smith. 2023. "Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK." *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1198–1212. doi:10.1145/3593013.3594074.

National Institute of Standards and Technology. 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://csrc.nist.gov/publications/detail/sp/1270/final. (2023).

J. Noller. 2024. "Extended Human Agency: Towards a Teleological Account of AI." *Humanities and Social Sciences Communications*, 11, 1, 1–7.

A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman. 2019. "Social Data: Biases, Methodological pitfalls, and Ethical Boundaries." *Frontiers in big data*, 2, 13.

F. Osasona, O. O. Amoo, A. Atadoga, T. O. Abrahams, O. A. Farayola, and B. S. Ayinla. 2024. "Reviewing the Ethical Implications of AI in Decision Making Processes." *International Journal of Management & Entrepreneurship Research*, 6, 2, 322–335.

M. J. Page et al.. 2021. "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews." *BMJ*, 372.

C. Panigutti et al.. 2023. "The Role of Explainable AI in the Context of the AI Act." *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1139–1150. doi:10.1145/3593013.3594069.

E. Paraschou, M. Michali, S. Yfantidou, S. Karamanidis, S. R. Kalogeros, and A. Vakali. 2025. "Ties of Trust: a bowtie model to uncover trustor-trustee relationships in LLMs." In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 1715–1728.

S. Pareek, E. Velloso, and J. Goncalves. 2024. "Trust Development and Repair in AI-Assisted Decision-Making during Complementary Expertise." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 546–561. doi:10.1145/3630106.3658924.

I. van de Poel. 2020. "Embedding Values in Artificial Intelligence (AI) Systems." *Minds and Machines*, 30, 3, 385–409.

N. Polemi, I. Praça, K. Kioskli, and A. Bécue. 2024. "Challenges and Efforts in Managing AI Trustworthiness Risks: A State of Knowledge." *Frontiers in Big Data*, 7, 1381163.

N. Proferes, N. Jones, S. Gilbert, C. Fiesler, and M. Zimmer. 2021. "Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics." *Social Media+ Society*, 7, 2, 20563051211019004.

S. Rawas. 2024. "AI: The Future of Humanity." *Discover Artificial Intelligence*, 4, 1, 25.

W. Saeed and C. Omlin. 2023. "Explainable AI (XAI): A Systematic Meta-survey of Current Challenges and Future Opportunities." *Knowledge-Based Systems*, 263, 110273.

F. Santoni de Sio and G. Mecacci. 2021. "Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them." *Philosophy & Technology*, 34, 4, 1057–1084.

S. Sarkar, M. Gaur, L. K. Chen, M. Garg, and B. Srivastava. 2023. "A Review of the Explainability and Safety of Conversational Agents for Mental Health to Identify Avenues for Improvement." *Frontiers in Artificial Intelligence*, 6, 1229805.

N. Scharowski, M. Benk, S. J. Kühne, L. Wettstein, and F. Brühlmann. 2023. "Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study." In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '23). Chicago, IL, USA, 248–260.

N. Schlicker, K. Baum, A. Uhde, S. Sterz, M. C. Hirsch, and M. Langer. 2025. "How Do We Assess the Trustworthiness of AI? Introducing the Trustworthiness Assessment Model (TrAM)." *Computers in Human Behavior*, 108671.

N. Schlicker, F. Lechner, K. Wehrle, B. Greulich, M. C. Hirsch, and M. Langer. May 2025. "Trustworthy Enough? Examining Trustworthiness Assessments of Large Language Model-based Medical Agents." en. *Technology, Mind, and Behavior*, 6, 2, (May 2025), 1–29. doi:10.1037/tmb0000164.

N. Schlicker, A. Uhde, K. Baum, M. C. Hirsch, and M. Langer. 2022. "Calibrated Trust as a Result of Accurate Trustworthiness Assessment–Introducing the Trustworthiness Assessment Model." *PsyArXiv PPR:PPR547033*.

A. Schmitz. 2023. "Towards Formalizing and Assessing AI Fairness." *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 999–1001. doi:10.1145/3600211.3604762.

J. Schoeffer, N. Kuehl, and Y. Machowski. 2022. ""There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making." In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22). Seoul, Republic of Korea, 1616–1628.

Science Australia Department of Industry and Resources. 2023. *AI Ethics Principles*. https://legalinstruments.oecd.org/en/instruments/oecd-legal0449. Accessed: 2024-01-11. (2023).

K. Shailya, S. Rajpal, G. S. Krishnan, and B. Ravindran. 2025. "LExT: Towards Evaluating Trustworthiness of Natural Language Explanations." In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 1565–1587.

R. Shelby et al.. 2023. "Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction." *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 723–741. doi:10.1145/3600211.3604673.

Q. Steenhuis. 2024. "AI and Tools for Expanding Access to Justice." *SSRN*, 4876633.

J. M. Stern, J. S. Shiely, and I. Ross. 2002. *The EVA Challenge: Implementing Value-added Change in an Organization*. John Wiley & Sons.

M. T. Stevenson and J. L. Doleac. 2024. "Algorithmic Risk Assessment in the Hands of Humans." *American Economic Journal: Economic Policy*, 16, 4, 382–414.

W. Strielkowski, V. Grebennikova, A. Lisovskiy, G. Rakhimova, and T. Vasileva. 2025. "AI-driven Adaptive Learning for Sustainable Educational Transformation." *Sustainable Development*, 33, 2, 1921–1947.

J. Tan, H. Westermann, and K. Benyekhlef. 2023. "ChatGPT as an Artificial Lawyer?" In: *AI4AJ@ ICAIL*.

L. Y. Tan, S. Hu, D. J. Yeo, and K. H. Cheong. 2025. "Artificial Intelligence-Enabled Adaptive Learning Platforms: A Review." *Computers and Education: Artificial Intelligence*, 100429.

The Government of Canada. 2023a. *Directive on Automated Decision-Making*. https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592. Accessed: 2024-01-11. (2023).

The Government of Canada. 2023b. *Responsible Use of Artificial Intelligence Guiding Principles*. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html. Accessed: 2024-01-11. (2023).

S. Thiebes, S. Lins, and A. Sunyaev. 2021. "Trustworthy Artificial Intelligence." *Electronic Markets*, 31, 447–464.

L. Thornton, B. Knowles, and G. Blair. 2021. "Fifty Shades of Grey: In Praise of a Nuanced Approach Towards Trustworthy Design." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 64–76. doi:10.1145/3442188.3445871.

L. Thornton, B. Knowles, and G. Blair. 2022. "The Alchemy of Trust: The Creative Act of Designing Trustworthy Socio-Technical Systems." In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22). Seoul, Republic of Korea, 1387–1398.

A. Toney, K. Curlee, and E. Probasco. 2024. "Trust Issues: Discrepancies in Trustworthy AI Keywords Use in Policy and Research." In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '24). Rio de Janeiro, Brazil, 2222–2233.

E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. van Moorsel. 2020. "The Relationship between Trust in AI and Trustworthy Machine Learning Technologies." In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (FAT* '20). Barcelona, Spain, 272–283.

T. R. Tyler. 2001. "Trust and Law Abidingness: A Proactive Model of Social Regulation." *Boston University Law Review*, 81, 361.

United Kingdom Information Commissioner's Office. 2023. *Guidance on AI and Data Protection*. Accessed: 2024-01-11. (2023). https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/.

O. Vereschak, G. Bailly, and B. 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, 1–39.

E. J. de Visser, M. Cohen, A. Freedy, and R. Parasuraman. 2014. "A Design Methodology for Trust Cue Calibration in Cognitive Agents." In: *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments: 6th International Conference, VAMR 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part I 6*. Springer, 251–262.

T. R. Vought. 2020. *OMB Memorandum M-21-06: Guidance for Regulation of Artificial Intelligence Applications*. Accessed: 2024-01-11. (2020). https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf.

B. Wang et al.. 2023. "DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models." In: *NeurIPS*.

J. Wang, H. Li, H. Wang, S. J. Pan, and X. Xie. 2023. "Trustworthy Machine Learning: Robustness, Generalization, and Interpretability." In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5827–5828.

R. Wang, R. Cheng, D. Ford, and T. Zimmermann. 2024. "Investigating and Designing for Trust in AI-powered Code Generation Tools." In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '24). Rio de Janeiro, Brazil, 1475–1493.

M. Whittaker. Nov. 2021. "The Steep Cost of Capture." *Interactions*, 28, 6, (Nov. 2021), 50–55. doi:10.1145/3488666.

T. Williams and K. S. Haring. 2023. "No Justice, No Robots: From the Dispositions of Policing to an Abolitionist Robotics." *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 566–575. doi:10.1145/3600211.3604663.

J. Zerilli, U. Bhatt, and A. Weller. 2022. "How Transparency Modulates Trust in Artificial Intelligence." *Patterns*, 3, 4.

B. Zhang and A. Dafoe. 2020. "U.S. Public Opinion on the Governance of Artificial Intelligence." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 187–193. doi:10.1145/3375627.3375827.

J. Zhang and Z.-m. Zhang. 2023. "Ethics and Governance of Trustworthy Medical Artificial Intelligence." *BMC Medical Informatics and Decision Making*, 23, 1, 7.

M. Zilka, H. Sargeant, and A. Weller. 2022. "Transparency, Governance and Regulation of Algorithmic Tools Deployed in the Criminal Justice System: a UK Case Study." *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 880–889. doi:10.1145/3514094.3534200.

A: A List of Included Articles in the Final Corpus for Analysis

(1) **[FAccT]** Alpherts et al., Perceptive Visual Urban Analytics is Not (Yet) Suitable for Municipalities, *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.

(2) **[FAccT]** Andrus et al., What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

(3) **[FAccT]** Bhatt et al., Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

(4) **[FAccT]** Casper et al., Black-Box Access is Insufficient for Rigorous AI Audits. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.

(5) **[FAccT]** Engelmann et al., Clear Sanctions, Vague Rewards: How China's Social Credit System Currently Defines "Good" and "Bad" Behavior. *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019*.

(6) **[FAccT]** Engelmann et al., What People Think AI Should Infer From Faces. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

(7) **[FAccT]** Ferrario et al., How Explainability Contributes to Trust in AI. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

(8) **[FAccT]** Ferrario, A. A Trustworthiness-based Metaphysics of Artificial Intelligence Systems. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*.

(9) **[FAccT]** Inie et al., From "AI" to Probabilistic Automation: How Does Anthropomorphization of Technical Systems Descriptions Influence Trust? *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.

(10) **[FAccT]** Jacovi et al., Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

(11) **[FAccT]** Kim et al., "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.

(12) **[FAccT]** Kim et al., Humans, AI, and Context: Understanding End-Users Trust in a Real-World Computer Vision Application. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*.

(13) **[FAccT]** Kinahan et al., Achieving Reproducibility in EEG-Based Machine Learning. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.

(14) **[FAccT]** Knowles et al., Trustworthy AI and the Logics of Intersectional Resistance. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*.

(15) **[FAccT]** Knowles et al., The Sanction of Authority: Promoting Public Trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

(16) **[FAccT]** Liao et al., Designing for Responsible Trust in AI Systems: A Communication Perspective. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

(17) **[FAccT]** Maeda et al., When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.

(18) **[FAccT]** Magaña et al., Frameworks, Methods and Shared Tasks: Connecting Participatory AI to Trustworthy AI Through a Systematic Review of Global Projects. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*.
(19) **[FAccT]** Manzini et al., Should Users Trust Advanced AI Assistants? Justified Trust As a Function of Competence and Alignment. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.
(20) **[FAccT]** Nannini et al., Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*.
(21) **[FAccT]** Panigutti et al., The Role of Explainable AI in the Context of the AI Act. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*.
(22) **[FAccT]** Paraschou et al., Ties of Trust: A Bowtie Model to Uncover Trustor-Trustee Relationships in LLMs. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*.
(23) **[FAccT]** Pareek et al., Trust Development and Repair in AI-Assisted Decision-Making during Complementary Expertise. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.
(24) **[FAccT]** Scharowski et al., Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*.
(25) **[FAccT]** Schoeffer et al., There Is Not Enough Information: On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.
(26) **[FAccT]** Shailya et al., LExT: Towards Evaluating Trustworthiness of Natural Language Explanations. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*.
(27) **[FAccT]** Thornton et al., Fifty Shades of Grey: In Praise of a Nuanced Approach Towards Trustworthy Design. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.
(28) **[FAccT]** Thornton et al., The Alchemy of Trust: The Creative Act of Designing Trustworthy Socio-Technical Systems. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.
(29) **[FAccT]** Toney et al., Trust Issues: Discrepancies in Trustworthy AI Keywords Use in Policy and Research. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.
(30) **[FAccT]** Toreini et al., The Relationship between Trust in AI and Trustworthy Machine Learning Technologies. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.
(31) **[FAccT]** Wang et al., Investigating and Designing for Trust in AI-powered Code Generation Tools. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*
(32) **[AIES]** Brand et al., Exploring the Moral Value of Explainable Artificial Intelligence Through Public Service Postal Banks. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*.
(33) **[AIES]** Huang et al., Towards Just, Fair and Interpretable Methods for Judicial Subset Selection. *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*.
(34) **[AIES]** Kim et al., Age Bias in Emotion Detection: An Analysis of Facial Emotion Recognition Performance on Young, Middle-Aged, and Older Adults. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*.
(35) **[AIES]** Lakkaraju et al., "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*.
(36) **[AIES]** Lawrence et al., The Bureaucratic Challenge to AI Governance: An Empirical Assessment of Implementation at U.S. Federal Agencies. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*.
(37) **[AIES]** Loi et al., Towards Accountability in the Use of Artificial Intelligence for Public Administrations. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*.

(38) **[AIES]** Mohseni et al., Toward Design and Evaluation Framework for Interpretable Machine Learning Systems. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.*
(39) **[AIES]** Schmitz et al., Towards Formalizing and Assessing AI Fairness. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society.*
(40) **[AIES]** Shelby et al., Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*
(41) **[AIES]** Williams et al., No Justice, No Robots: From the Dispositions of Policing to an Abolitionist Robotics. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society.*
(42) **[AIES]** Zhang et al., U.S. Public Opinion on the Governance of Artificial Intelligence. *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society.*
(43) **[AIES]** Zilka et al., Transparency, Governance and Regulation of Algorithmic Tools Deployed in the Criminal Justice System: A UK Case Study. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society.*

B. Research Ethics and Social Impact

B.1 Ethical Considerations Statement

This paper is primarily a literature review and theoretical contribution, thus we did not engage in any human subjects research, systems development or deployment. Our work has been guided by considerable efforts of the AIES & FAccT community in the ethical ramifications of AI systems and their impact on human societies. This scoping review aims to contribute to a better understanding of AI trustworthiness, acknowledging the potential biases and limitations. We are committed to fostering responsible AI development and deployment, and this research is intended to support that goal.

B.2 Adverse Impact Statement

With this work, we seek to spark conversations across disciplines about the social and technical inequalities being seen in the literature while studying AI trustworthiness. Rather than "calling out" specific articles, we hope this critique serves as a "call in" other scholars to join together and work collectively towards critically understanding the trustworthiness of AI systems.

B.3 Social Impact

One of the primary impacts of AIES & FAccT scholarship is the enhancement of public literacy regarding AI technologies. By critically examining the articles focusing on understanding trustworthiness of AI systems, we hope to have fostered a greater understanding of the complexities involved in trustworthy AI deployment, particularly concerning issues of comprehensive measurement frameworks, socio-technical integration and contribution to SDGs.

Received 16 October 2025; accepted 13 January 2026