

Understanding AI Trustworthiness: A Scoping Review of AIES & FAccT Articles (Abstract Reprint)

Siddharth Mehrotra^{1,2}, Jin Huang³, Xuelong Fu¹, Roel Dobbe², Clara I. Sánchez¹ and Maarten De Rijke¹

¹ University of Amsterdam

² Delft University of Technology

³ University of Cambridge

Abstract Reprint. This is an abstract reprint of a journal article by [Mehrotra *et al.*, 2026].

jke. Understanding AI trustworthiness: A scoping review of AIES & FAccT articles. *Journal of Artificial Intelligence Research*, 85, 2026.

Abstract

Background: Trustworthy AI serves as a foundational pillar for two major AI ethics conferences: AIES and FAccT. However, current research often adopts techno-centric approaches, focusing primarily on technical attributes such as 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 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.

References

[Mehrotra *et al.*, 2026] Siddharth Mehrotra, Jin Huang, Xuelong Fu, Roel Dobbe, Clara I. Sánchez, and Maarten de Rijke.