THE LEGAL LIFE OF TRUST:

WHAT IS LOST WITH "TRUSTWORTHY AI"

María P. Angel [*]

EXTENDED ABSTRACT

"Trust"—in its colloquial sense—has become the master concept of digital governance. Privacy law scholars have long called for fiduciary duties to rebuild trust between users and data intermediaries (Waldman, 2018; Balkin, 2020; Richards & Hartzog, 2021). Until recently, online platforms largely governed themselves through "Trust & Safety" teams responsible for content moderation and the mitigation of online harms (Klonick, 2018; Citron & Waldman, 2025; Moran et al., 2025). Today, policymakers around the world promote "trustworthy AI" as the foundation of emerging AI governance frameworks (European Commission, 2019; National Institute of Standards and Technology, 2023; European Parliament & Council, 2024).

Despite its ubiquity, "trust" is an unstable and migrating concept whose object has shifted across governance domains. In privacy and online platform governance, trust has been tied to organizational actors with clear legal identities and potential regulatory obligations. Yet AI governance marks a striking departure from this pattern. Here, "trust" is relocated onto the AI system itself. Policymakers and industry leaders increasingly speak of "trustworthy AI," as though the artifact, rather than the institution developing or deploying it, were the agent capable of earning trust. This conceptual move is reinforced by technical governance practices—audits, evaluations, safety benchmarks, and alignment techniques—that target system performance rather than institutional behavior (Durán & Pozzi, 2025; Zanotti, 2025).

This shift in the object of trust has already been criticized for anthropomorphizing AI and diverting responsibility from those who develop and deploy it (Ryan, 2020). But the stakes are deeper. Drawing on policy analysis and doctrinal research, this Article traces the legal life of "trust" in the governance of digital technologies. It argues that by framing trustworthiness as a property of the AI system rather than of the corporations behind it, this latest shift strips trust of its analytical and normative value. Technical trustworthiness promises actionability, standardization, and auditability, but at the cost of flattening and downplaying two of the three defining features of trust relationships: power asymmetries and vulnerability.

When trust is framed as relating only to the artefact, the imbalance of power at stake risks being reduced to a merely computational register. By contrast, when power is understood as residing with AI companies, a broader political economy comes into view—one in which informatic power enables firms to accumulate and exercise not only technical capacity, but also market, cultural, and political power. In this context, vulnerability does not refer solely to the risk of technical error or

[*] Post-doctoral Resident Fellow, Yale's Information Society Project (ISP).

system malfunction, but also to conditions of market monopolization, citizen manipulation, and democratic destabilization.

Therefore, the Article calls for restoring "trust" as an analytical and normative lens that places the spotlight on the institutional incentives, organizational behaviors, and power structures that shape how AI systems are designed, deployed, and governed (Zuboff, 2019; Cohen, 2019; Kapczynski, 2020; Mejias & Couldry, 2024). Early efforts to mobilize the concept of "trust" in the privacy domain were rightly criticized for individualizing and legitimizing corporate power, thereby obscuring the broader political economy of digital technologies (Khan & Pozen, 2019). Building on that critique, the Article proposes a reorientation of AI governance that re-centers attention not on the power of AI companies alone, but on the broader socio-technical system that enables the development and diffusion of AI. By shifting the focus to systemic power, the Article charts a path for digital governance that redirects attention to the structural power asymmetries and system-level vulnerabilities that make AI governance most urgent.

References

Balkin, J. M. (2020). The fiduciary model of privacy. *Harvard Law Review Forum*, 134, 11-28.

Citron, D. K., & Waldman, A. E. (2025). *The evolution of trust and safety* (Virginia Public Law & Legal Theory Research Paper No. 2025-65). Emory Law Journal (forthcoming).

Cohen, J. E. (2019). *Between truth and power: The legal constructions of informational capitalism* (2nd ed.). Oxford University Press.

Durán, J. M., & Pozzi, G. (2025). *Trust and trustworthiness in AI*. *Philosophy & Technology, 38*(1), Article 16.

European Commission, High-Level Expert Group on Artificial Intelligence. (2019, April 8). *Ethics guidelines for trustworthy AI*. Shaping Europe's Digital Future.

European Parliament & Council. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council on artificial intelligence* [Artificial Intelligence Act]. *Official Journal of the European Union, L (...)*.

Kapczynski, A. (2020). *The law of informational capitalism*. *Yale Law Journal, 129*(5), 1276–1599.

Khan, L. M., & Pozen, D. E. (2019). A skeptical view of information fiduciaries. Harvard Law Review, 133, 497–541.

Klonick, K. (2018). *The new governors: The people, rules, and processes governing online speech*. *Harvard Law Review, 131*(6), 1598–1670.

Mejias, U. A., & Couldry, N. (2024). *Data Grab: The New Colonialism of Big Tech and How to Fight Back*. University of Chicago Press.

Moran, R. E., Schafer, J., Bayar, M. C., & Starbird, K. (2025). The end of "Trust & Safety"?: Examining the future of content moderation and upheavals in professional online safety efforts. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1).

Richards, N. M., & Hartzog, W. (2016). *Taking trust seriously in privacy law*. Stanford Technology Law Review, 19, 450.

Richards, N. M., & Hartzog, W. (2021). *A duty of loyalty for privacy law*. *Washington University Law Review, 99*(4), 961–1048.

Ryan, M. (2020). *In AI we trust: Ethics, artificial intelligence, and reliability*. *Science and Engineering Ethics, 26*(5), 2749–2767.

Waldman, A. E. (2018). *Privacy as trust: Information privacy for an information age*. Cambridge University Press.

Zanotti, G. (2025). *AI systems should be trustworthy, not trusted*. *AI & Society*.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.