UNDONE TRANSPARENCY:

HOW TO ADDRESS BLIND SPOTS IN AI GOVERNANCE CAUSED BY SELF-REPORTING

Shlomi Hod, Weizenbaum Institute, Germany

Maayan Perel, Netanya Academic College, Israel

Yonatan Lourie, Tel Aviv University, Israel

Niva Elkin-Koren, Tel Aviv University, Israel

Transparency has become a central governance tool in Artificial Intelligence (AI) regulatory frameworks (Bloch-Wehba, 2019; OECD, 2024; Sloane & Wüllhorst, 2025) and it is widely presented as the cure for the opacity of AI systems (Ingrams & Klievink, 2022; Novelli, 2024). Existing legal frameworks, including the recently enacted European AI Act and the Digital Services Act (DSA), seek to enhance AI transparency primarily by imposing various disclosure duties (Sloane & Wüllhorst, 2025). Those include transparency reports of aggregate statistics, statement-of-reason microdata, and periodical bias audits—all reflecting the assumption that detailed data sharing obligations would yield procedural fairness and enable meaningful oversight by impacted stakeholders (Wagner et al., 2020).

Nevertheless, transparency that is operationalized primarily through self-reported disclosures of summary statistics and data—Transparency-by-Disclosure (TbD)—fails to facilitate the full range of potential systemic investigations needed to meaningfully hold AI systems accountable. Thus, the TbD approach risks reproducing the very opacity it seeks to undo. Specifically, transparency reports generated by the same corporations that design or deploy AI systems are limited to what those corporations are willing to reveal, rather than what society needs to know (Perel & Elkin-Koren 2017; Douek, 2022). Indeed, prior research found serious deficiencies in self-reported data, including partial, vague and inconsistent entries, unreliable data quality and inconsistencies between different types of self-reported data (Sekwenz, Wagner, and Bruijn 2025; Shahi et al. 2025; Trujillo, Fagni, and Cresci 2025). Moreover, TbD does little to surface the underlying value choices encoded in AI systems. Their models learn patterns, make predictions, generate content, and operationalize norms that are not directly observable from disclosure outputs (Goodman & Trehu, 2022; Rieder & Hofmann, 2020). Making these latent normative choices legible to those humans impacted by the system adds another layer of complexity. Overall, TbD creates and perpetuate systemic gaps in knowledge about how AI systems actually function, whom they affect, and which values they encode.

Accordingly, we propose Transparency-by-Middleware (TbM) as a complementary approach to TbD. Middleware is an intermediary software that interfaces between stakeholders and platforms, enabling external functions such as filtering, curation, auditing, or customization without modifying the platform's core architecture (Fukuyama et al. 2020; Hogg et al. 2024). TbM positions transparency not as a data-availability exercise, but as a distributed practice of multiple stakeholders developing testable inquiries about AI systems, guided by their own interests and values and operationalized by their own tools. In TbM, the power to define what matters is

hence redistributed outwards from the platform to society: The questions come first; the required data, methodology and access mechanisms follow from those inquiries.

In this work, we explore the potential of TbM to address a critical sociotechnical transparency challenge: how AI-based content moderation systems on social media platforms generate and enforce speech norms. We focus on two building-blocks required to implement TbM effectively. First, we propose a typology of inquires that helps stakeholders structure the inquiries that matter: those that allow stakeholders to translate abstract social values (i.e., accountability or fairness) into concrete, well-defined, measurable questions about AI systems. This typology structures inquiries across three dimensions: (1) The type of information sought (2) The technical component being examined (e.g., classifier, enforcer, or the system as a whole) and (3) The speech norm (or set of norms) being operationalized at scale. Second, we analyze the access mechanism, namely, the layer that mediates between the stakeholder and the platform, enabling these inquiries. These access mechanisms must meet legal and technical requirements, including safeguards for privacy and data protection, security and intellectual property. Overall, we offer a structured approach to identifying and evaluating key questions, providing methods to assess how content moderation systems encode, enforce and reproduce speech norms.

Our work makes three main contributions. First, we re-conceptualize transparency as a process of independent inquiries, rather than static disclosures, extending the idea of middleware to the transparency domain and emphasizing a tool-building mindset that empowers stakeholders to pursue their own inquires. Second, we develop a typology that decomposes systemic inquiries about AI content moderation systems in accordance with the system's components and demonstrate its practical feasibility. Third, we outline legal policy implications, advocating a shift from a mere right to access platform data toward a right to systematically interrogate opaque systems, thereby enhancing accountability while reducing regulatory burdens.

Hod, Shlomi and Perel, Maayan and Lourie, Yonatan and Elkin-Koren, Niva, *Transparency by Inquiry: Unveiling Speech Norms in AI Content Moderation*. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6094466

References

Bloch-Wehba, H. (2020). "Access to Algorithms". *Fordham Law Review* 88: 1265.

Douek, E. (2022). "Content moderation as systems thinking". *Harvard Law Review*, 136: 526-607.

European Commission (2022). Regulation (EU) 2022/2065 (Digital Services Act).Official Journal of the European Union. url: https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng.

Fukuyama, et al. (2020). Middleware for Dominant Digital Platforms: A Technological Solution to a Threat to Democracy. Stanford Cyber Policy Center, https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/cpc-middleware_ff_v2.pdf

Goodman, E. P., Trehu, J. (2023). "Algorithmic auditing: chasing ai accountability". *Santa Clara High Technology Law Journal*. 39(3): 289-338.

Hogg, L., DiResta, R., Fukuyama, F., Reisman, R., Keller, D., Ovadya, A., Thorburn, L., Stray, J., Mathur, S. (2024). "Shaping the Future of Social Media with Middleware". *ArXiv, abs/2412.10283*.

Ingrams, Alex, and Bram Klievink, 'Transparency's Role in AI Governance', in Justin B. Bullock, and others (eds), *The Oxford Handbook of AI Governance*, Oxford Handbooks (2024; online edn, Oxford Academic, https://doi.org/10.1093/oxfordhb/9780197579329.013.32

Novelli, C., Taddeo, M. & Floridi, L. (2024) Accountability in artificial intelligence: what it is and how it works. *AI & Soc* 39, 1871–1882. https://doi.org/10.1007/s00146-023-01635-y

OECD. 2024. OECD AI Principles. https://www.oecd.org/en/topics/sub-issues/ai-principles.html.

Perel, M., Elkin-Koren, N. (2017). "Black box tinkering: Beyond disclosure in algorithmic enforcement". *Florida Law Review*. 69: 181-221. https//:scholarship.law.ufl.edu/flr/vol69/iss1/5.

Rieder, B., & Hofmann, J. (2020). Towards platform observability. Internet Policy Review, 9(4). https://doi.org/10.14763/2020.4.1535

Sekwenz, MT., Wagner, B., and De Bruijn, H., (2025). "From Reports to Reality: Testing Consistency in Instagram's Digital Services Act Compliance Data." arXiv preprint arXiv:2507.01787.

Shahi, GK, Tessa, B., Trujillo, A, and Cresci ,S. (2025). "A Year of the DSA Transparency Database: What it (Does Not) Reveal About Platform Moderation During the 2024 European Parliament Election." arXivpreprint arXiv:2504.06976.

Sloane, M., & Wüllhorst, E. (2025). A systematic review of regulatory strategies and transparency mandates in AI regulation in Europe, the United States, and Canada. Data & Policy, 7, e11. doi:10.1017/dap.2024.54

Trujillo, A., Fagni, T., and Cresci. S. (2025). "The DSA Transparency Database: Auditing self-reported moderation actions by social media." *Proceedings of the ACM on Human-Computer Interaction*, 9, 2, 1–28.

Wagner, B., Rozgonyi, K., Sekwenz, M.T., Cobbe, J., Singh, J. (2020). "Regulating transparency? Facebook, Twitter and the German Network Enforcement Act". *In Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency* pp. 261-272.