Harvard Data Science Review • Issue 7.4, Fall 2025

Justice at the Core of a New AI Paradigm

Giuseppe Primiero[1]

[1]Logic, Uncertainty, Computation and Information Lab, PhilTech Research Center for the Philosophy of Technology, Department of Philosophy, University of Milan, Milan, Italy

The MIT Press

Published on: Oct 31, 2025

DOI: https://doi.org/10.1162/99608f92.3f78ed8e

License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)

A Significant Rethinking of What AI Is

Leonelli (2025) offers a compelling environmental perspective on the mission of creating artificial intelligence (AI), urging us to reconsider the broader scientific and societal implications of technological innovation. Among the many issues she raises, one particularly critical point deserves everyone's closer attention: the recognition that "there are severe and increasing disparities not only in who has access to life-changing technologies but also and most importantly in who has the means to shape such innovations." I fully endorse this observation, as well as the subsequent claim that "what is needed is the incorporation of affirmative action into technology design to address digital and material divides."

This perspective calls for a renewed and expanded sense of responsibility—one that builds upon, but goes beyond, the ethical obligations traditionally associated with scientific practice. This means embracing a vision of AI that does not merely avoid exacerbating social inequalities but actively seeks to mitigate them. Even more strongly: AI can be considered truly intelligent only if it contributes to the creation of a more equitable and just human society. Whose responsibility this is, and which tools are needed to realize it are major questions for our society, and for the science and humanities communities.

Levels of Responsibility

A call for responsibility in the development and deployment of AI requires articulation at distinct levels (Primiero, 2025). First, there is a primary responsibility borne by developers—those who design models, curate and process training data, invest substantial computational resources to optimize performance, and make critical decisions regarding the application and deployment of AI systems. This level of responsibility concerns the ethical and technical integrity of the systems themselves, particularly in ensuring that their outputs are fair and socially beneficial.

The second level of responsibility pertains to oversight and verification. It involves the continuous monitoring of AI systems to ensure that the commitments made at the first level are upheld—namely, that the systems have been developed and trained in ways that promote fairness and mitigate bias, and moreover that the deployment does not have unexpected side effects. This form of verification must be embedded throughout the entire life cycle of machine learning (ML) systems, from initial design to postdeployment evaluation. Crucially, the question of who assumes responsibility for this oversight is not merely technical but political. It depends on regulatory frameworks and policy decisions that govern the economic and institutional structures surrounding AI development and distribution. Determining accountability in this phase requires deliberate choices about governance, transparency, and the equitable allocation of power in shaping technological futures.

A third level of responsibility can be assigned to clients, users, and deployers, acknowledging that their choices impact the environment, the community, and the infrastructure of our social life. Being mindful of which

technologies are selected, the contexts in which they are used, and the purposes they serve is a crucial step toward understanding their limitations and potential harms.

From Functional Correctness to Technological Justice

Since its inception, the history of computer science and technology has been fundamentally shaped by the imperative to control and predict the behavior of computational systems. As artefacts situated at the intersection of scientific rigor and creative design (De Mol & Primiero, 2018), computers have been engineered to execute programs in accordance with formally specified functionalities (Primiero, 2019; Turner, 2011). The role of formal methods in this endeavour remains until today paramount: alongside extensive empirical testing, the abstraction of programs as mathematical entities enables the formal verification of properties that help prevent malfunctions and unintended behavior. This methodological framework is essential for ensuring the reliability and safety of computational systems—especially given the extent to which our daily lives are governed by countless automated operations.

The advent of ML methods as the new paradigm of AI in the last decade has introduced two major novelties in this picture: nondeterminism and higher societal impact. The nature of ML-based classifiers, predictors, and generators is inherently probabilistic, and as such the verification of their behavior has increased in complexity. Formal verification of nondeterministic systems has been studied for a long time, and the application to ML systems is growing steadily. But the high variability in output of an ML system depends in large part on factors that go beyond mathematical modeling: the data used to train it, the number of parameters, and not the least the adherence of the input data to the data the model has previously seen as its image of the world. Hence, beyond the novel challenge of combining statistical analysis with logical inference to determine what an ML model will produce as output, the required understanding of correctness for such systems has essentially changed in view of their increased societal impact.

Today we are beyond safety-critical, we have reached justice-critical. As Eubanks (2018) and Noble (2018) among others have documented in detail, algorithmic systems deployed across public assistance programs, hiring platforms, criminal justice tools, and educational technologies often reproduce and even exacerbate existing social inequalities. These systems are not only products of biased data but also of institutional and epistemic frameworks that favor surveillance, control, and efficiency over equity and care. As such, algorithmic systems have become vectors of what Fricker (2017) calls epistemic injustice—excluding marginalized groups from the very processes that define them, judge them, and make decisions about them. *Technological injustice* is a better moniker for what we should count as an unacceptable misfunctioning: the systematic exclusion, misrepresentation, or devaluation of individuals or communities through automated systems. We should not only invest in new methods and tools to identify it, but rather deploy affirmative action to prevent it and correct it. Justice is not just an add-on to smart technologies, it is an essential part of their correctness.

Care as an Objective, Corrective Abduction as a Means

A constructive and supportive relation between developers, deployers, and users, emphasizing attentiveness, responsibility, and responsiveness should recognize that technologies are not neutral tools but socially situated entities shaped by—and shaping—human practices, a main element in Leonelli's understanding of environmental intelligence (EI). Understanding computational systems as ontologically embedded in social contexts allows for a more nuanced assessment of their ethical and political dimensions. This includes developing frameworks that trace responsibility across complex sociotechnical networks, assess the justice implications of algorithmic design, and integrate care-based ethics into system development. As Leonelli puts it: "Mechanisms need to be in place to ward against authoritarian monopolies of power and foster a multitude of voices within political and public discourse."

In this light, what do we require from current correctness verification tools to become tools for verifying responsible development and caring deployment of our technologies? While fairness metrics—such as statistical parity, equalized odds, and counterfactual fairness—offer formalized, quantifiable approaches to evaluating bias (Mehrabi et al., 2021; Verma & Rubin 2018), they often lack the normative grounding necessary for restorative or justice-oriented correction. These metrics can detect disparities but do not inherently guide systems toward ethical remediation or contextual fairness. To address this gap, symbolic logic provides a promising avenue for embedding normative reasoning directly into computational inference. Current efforts to include normative inference in nonmonotonic and statistical reasoning (Baldi et al., 2025; Kyburg et al., 2006) resonate with contemporary research on hybrid AI systems that integrate symbolic and subsymbolic reasoning (Nawaz et al., 2025), aiming to combine the interpretability and structure of logic with the adaptability of neural networks. Such integration is crucial for developing systems that can reason about AI trustworthiness, fairness, responsibility, and correction in a principled manner (D'Asaro et al., 2025). Bias detection and correction tools, both those based on purely statistical analyses (Google Responsible AI Team, 2024; IBM Research AI, 2024; Microsoft Research, 2024; Saleiro et al., 2018) and those combining subsymbolic with symbolic procedures (Coraglia et al., 2023) remain currently limited in a key challenge: the development of an abductive reasoning mechanism capable of identifying miscomputations thorough implementing causal information and proposing corrective actions aligned with normative goals. Abductive reasoning—inferring the best explanation for observed anomalies—offers a powerful foundation for such a mechanism. Formal models of abduction should be extended to support ethical inference, enabling systems to not only detect when a decision deviates from fairness norms but also to hypothesize and implement corrective strategies. This would mark a significant step toward computational systems that are not only technically robust but also ethically responsive.

If specifying for trustworthiness is already on the agenda of the formal scientists (Abeywickrama et al., 2023), specifying for justice should be our next objective.

Conclusion

Leonelli's appeal to reconsider the foundational principles of AI, particularly her insistence on integrating environmental concerns and the ethics of care as essential, nonnegotiable dimensions, constitutes an invitation for both humanists and scientists to initiate research agendas guided by renewed inspiration. For formal methods and the philosophy of computer science, this entails an urgent need to reconceptualize notions of correctness and the design of verification strategies in ways that explicitly incorporate and pursue justice. The present moment offers a critical opportunity to reorient the fundamental aims of research in accordance with this objective, rather than treating it as external to technical concerns.

Disclosure Statement

This research was supported by the Ministero dell'Università e della Ricerca (MUR) through PRIN 2022 Project SMARTEST – Simulation of Probabilistic Systems for the Age of the Digital Twin (20223E8Y4X), and through the Project "Departments of Excellence 2023-2027" awarded to the Department of Philosophy "Piero Martinetti" of the University of Milan.

References

Abeywickrama, D. B., Bennaceur, A., Chance, G., Demiris, Y., Kordoni, A., Levine, M., Moffat, L., Moreau, L., Mousavi, M. R., Nuseibeh, B., Ramamoorthy, S., Ringert, J. O., Wilson, J., Windsor, S., & Eder, K. (2023). On specifying for trustworthiness. *Communications of the ACM*, *67*(1), 98–109. https://doi.org/10.1145/3624699

Baldi, P., Corsi, E., & Hosni, H. (2025). A logical framework for data-driven reasoning. *Logic Journal of the IGPL*, *33*(3), Article jzae113. https://doi.org/10.1093/jigpal/jzae113

Coraglia, G., D'Asaro, F. A., Genco, F. A., Giannuzzi, D., Posillipo, D., Primiero, G., & Quaggio, C. (2023). Brioxalkemy: A bias detecting tool. In G. Boella, F. A. D'Asaro, A. Dyoub, L. Gorrieri, F. A. Lisi, C. Manganini & G. Primiero (Eds.), *AI*IA 2023: Proceedings of the 2nd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence* (Vol. 3615, CEUR Workshop Proceedings; pp. 44–60). CEUR-WS.org. https://ceur-ws.org/Vol-3615/paper4.pdf

D'Asaro, F., Genco, F., & Primiero, G. (2025). Checking trustworthiness of probabilistic computations in a typed natural deduction system. *Journal of Logic and Computation*, *35*(2), Article exaf003. https://doi.org/10.1093/logcom/exaf003

De Mol, L., & Primiero, G. (2018). Programming systems: In search of historical and philosophical foundations. In L. De Mol & G. Primiero (Eds.), *Reflections on programming systems*. Springer. https://doi.org/10.1007/978-3-319-97226-8_1

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Fricker, M. (2017). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Google Responsible AI Team. (2024). *Fairness indicators: A tool for evaluating AI fairness in TensorFlow models*.

IBM Research AI. (2024). *AI Fairness 360: An extensible toolkit for detecting and mitigating bias in machine learning models*.

Kyburg, Jr., H. E., & Teng, C. M. (2006). Nonmonotonic logic and statistical inference. *Computational Intelligence*, *22*(1), 26–51. https://doi.org/10.1111/j.1467-8640.2006.00272.x

Leonelli, S. (2025). Environmental intelligence: Redefining the philosophical premises of AI. *Harvard Data Science Review*, *7*(4). https://doi.org/10.1162/99608f92.ac7c1504

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), Article 115. https://doi.org/10.1145/3457607

Microsoft Research. (2024). *Fairlearn: A toolkit for assessing and improving fairness in machine learning*.

Nawaz, U., Anees-ur-Rahaman, M., & Saeed, Z. (2025). A review of neuro-symbolic AI integrating reasoning and learning for advanced cognitive systems. *Intelligent Systems with Applications*, *26*, Article 200541. https://doi.org/10.1016/j.iswa.2025.200541

Noble, S. (2018) *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Primiero, G. (2019). *On the foundations of computing*. Oxford University Press.

Primiero, G. (2025). BRIO: il ruolo della logica nella costruzione di un'Intelligenza Artificiale equa, volume *Spiegabilità e Intelligenza Artificiale. Prospettive tecnologiche, etiche e normative*, Serie *Etica della ricerca, bioetica, biodiritto e biopolitica* [BRIO: The Role of Logic in Building Fair Artificial Intelligence, in *Explainability and Artificial Intelligence: Technological, Ethical and Regulatory Perspectives*, part of the series *Ethics of Research, Bioethics, Biolaw and Biopolitics*]. CNR Edizioni.

Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., & Ghani, R. (2018). *Aequitas: A bias and fairness audit toolkit*. ArXiv. https://doi.org/10.48550/arXiv.1811.05577

Turner, R. (2011). Specification. *Minds and Machines*, *21*(2), 135–152. https://doi.org/10.1007/s11023-011-9239-x

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In Y. Brun, B. Johnson, & A. Meliou (Eds.), *FairWare '18: Proceedings of the International Workshop on Software Fairness* (pp. 1–7). ACM. https://doi.org/10.1145/3194770.3194776

©2025 Giuseppe Primiero. This article is licensed under a Creative Commons Attribution (CC BY 4.0) International license, except where otherwise indicated with respect to particular material included in the article.

References

- Abeywickrama, D. B., Bennaceur, A., Chance, G., Demiris, Y., Kordoni, A., Levine, M., Moffat, L., Moreau, L., Mousavi, M. R., Nuseibeh, B., Ramamoorthy, S., Ringert, J. O., Wilson, J., Windsor, S., & Eder, K. (2023). On specifying for trustworthiness. *Communications of the ACM*, *67*(1), 98–109. https://doi.org/10.1145/3624699 ↩
- Baldi, P., Corsi, E., & Hosni, H. (2025). A logical framework for data-driven reasoning. *Logic Journal of the IGPL*, *33*(3), Article jzae113. https://doi.org/10.1093/jigpal/jzae113 ↩
- Coraglia, G., D'Asaro, F. A., Genco, F. A., Giannuzzi, D., Posillipo, D., Primiero, G., & Quaggio, C. (2023). Brioxalkemy: A bias detecting tool. In G. Boella, F. A. D'Asaro, A. Dyoub, L. Gorrieri, F. A. Lisi, C. Manganini & G. Primiero (Eds.), *AI*IA 2023: Proceedings of the 2nd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence* (Vol. 3615, CEUR Workshop Proceedings; pp. 44–60). CEUR-WS.org. https://ceur-ws.org/Vol-3615/paper4.pdf ↩
- D'Asaro, F., Genco, F., & Primiero, G. (2025). Checking trustworthiness of probabilistic computations in a typed natural deduction system. *Journal of Logic and Computation*, *35*(2), Article exaf003. https://doi.org/10.1093/logcom/exaf003 ↩
- De Mol, L., & Primiero, G. (2018). Programming systems: In search of historical and philosophical foundations. In L. De Mol & G. Primiero (Eds.), *Reflections on programming systems*. Springer. https://doi.org/10.1007/978-3-319-97226-8_1

↩

- Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

↩

- Fricker, M. (2017). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

↩

- Google Responsible AI Team. (2024). *Fairness indicators: A tool for evaluating AI fairness in TensorFlow models*.

↩

- IBM Research AI. (2024). *AI Fairness 360: An extensible toolkit for detecting and mitigating bias in machine learning models*.

↩

- Kyburg, Jr., H. E., & Teng, C. M. (2006). Nonmonotonic logic and statistical inference. *Computational Intelligence*, *22*(1), 26–51. https://doi.org/10.1111/j.1467-8640.2006.00272.x

↩

- Leonelli, S. (2025). Environmental intelligence: Redefining the philosophical premises of AI. *Harvard Data Science Review*, *7*(4). https://doi.org/10.1162/99608f92.ac7c1504

↩

- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Sur*veys, 54(6), Article 115. https://doi.org/10.1145/3457607

↩

- Microsoft Research. (2024). *Fairlearn: A toolkit for assessing and improving fairness in machine learning.*

↩

- Nawaz, U., Anees-ur-Rahaman, M., & Saeed, Z. (2025). A review of neuro-symbolic AI integrating reasoning and learning for advanced cognitive systems. *Intelligent Systems with Applications*, *26*, Article 200541. https://doi.org/10.1016/j.iswa.2025.200541

↩

- Noble, S. (2018) *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

↩

- Primiero, G. (2019). *On the foundations of computing*. Oxford University Press.

↩

- Primiero, G. (2025). BRIO: il ruolo della logica nella costruzione di un'Intelligenza Artificiale equa, volume *Spiegabilità e Intelligenza Artificiale. Prospettive tecnologiche, etiche e normative*, Serie *Etica della ricerca, bioetica, biodiritto e biopolitica* [BRIO: The Role of Logic in Building Fair Artificial Intelligence, in *Explainability and Artificial Intelligence: Technological, Ethical and Regulatory Perspectives*, part of the series *Ethics of Research, Bioethics, Biolaw and Biopolitics*]. CNR Edizioni.

↩

- Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., & Ghani, R. (2018). *Aequitas: A bias and fairness audit toolkit*. ArXiv. https://doi.org/10.48550/arXiv.1811.05577

↩

- Turner, R. (2011). Specification. *Minds and Machines*, *21*(2), 135–152. https://doi.org/10.1007/s11023-011-9239-x

↩

- Verma, S., & Rubin, J. (2018). Fairness definitions explained. In Y. Brun, B. Johnson, & A. Meliou (Eds.), *FairWare '18: Proceedings of the International Workshop on Software Fairness* (pp. 1–7). ACM. https://doi.org/10.1145/3194770.3194776

↩