Responsible AI for Research: Between Moral Philosophy and Philosophy of Science

Alexander Martin Mussgnug[1]

Sabina Leonelli[2]

Shannon Vallor[3]

Forthcoming in The Palgrave Handbook on the Ethics of AI

I Introduction

Today, scientists employ AI to identify relevant literature, facilitate large-scale data analysis, and automate routine experimental procedures. Tomorrow, fully autonomous AI may fundamentally alter the way science is done (Royal Society 2024). Current and future applications of AI in science have motivated a wide range of research exploring how to responsibly integrate AI into scientific practice.

While AI undoubtedly opens up enticing technical possibilities, worries about a decline in quality of scientific procedures and outputs, combined with the increasing inscrutability of research systems, have raised concerns regarding the scientific implications of using AI. The social and ethical implications of such developments are also cause for concern, not least because scientific research often grounds far-reaching political, professional, and personal decision-making. The automation of key moments of evaluation and judgement within the scientific process thus leads to critical questions regarding quality, trust, and accountability beyond the realm of scientific research alone.

The scientific and social dimensions of AI's impact on science are, in our view, inextricably linked. Any problems in the quality and reliability of methods and outputs will reflect on misleading, problematic or downright damaging uses of science in society. Conversely, the ongoing erosion of trust and accountability through AI and the growing resistance toward socially harmful uses of the technology, will ultimately also shape the very conditions for how AI is implemented into scientific research. While philosophers can play a central role in analysing these interrelated issues and their repercussions, discussions around the ethical and epistemic dimensions of AI in science tend to be siloed into different domains of philosophy.

[1] Stanford University, alexander@mussgnug.de

[2] Technical University of Munich, sabina.leonelli@tum.de

[3] University of Edinburgh, Centre for Technomoral Futures, svallor@ed.ac.uk

On the one hand, the epistemic dimensions of AI applications are most comprehensively discussed in the philosophy of science. Among others, philosophers of science have explored how the opacity of deep learning models impacts scientific discovery (e.g., Boge 2022), outlined under what conditions AI models lead to robust results (e.g., Freiesleben & Grote 2023), and debated the potential of fully automated knowledge generation through AI (e.g., Bertolaso & Sterpetti 2020). On the other hand, ethical considerations surrounding the development and use of AI are most extensively explored in the moral philosophy of technology. Within it, scholars have investigated issues such as the fairness and ecological sustainability of AI applications (e.g., Selbst et al 2019; van Wynsberghe 2021), the degree of moral responsibility attributable to AI developers (e.g., Oimann & Tollon 2025), as well as the general effects of AI automation on our social, political, and moral lives (e.g., Vallor 2024).

Our chapter contributes to ongoing research attempting to overcome this division by linking debates in the history and philosophy of science with considerations from the ethics of technology. Section two introduces existing scholarship that connects epistemic and ethical implications of AI. Section three presents our case study. We comment on the background conditions of AI in science, and particularly on questions of inclusivity and equity. In so doing, we provide brief examples of the close connection of epistemic and ethical concerns around the use of AI for research, drawing from scholarship in both the philosophy of science and the moral philosophy of technology. We conclude in section four.

II Between ethics and epistemology

We begin by situating our exploration within existing scholarship acknowledging the close relationship between ethical and epistemic considerations. We focus first on accounts that do so in the context of science, before turning to work on the epistemology and ethics of AI.

II.1 Ethics and epistemology in science

In the history and philosophy of science, the close relationship between epistemic and ethical considerations has been a topic of long-standing concern. Much of this work emerged as a response to the value-free ideal of science. Roughly, the value-free ideal holds that while values can play a role in the choice of a research question and the application of scientific findings, the core of the scientific research process itself should be characterized by the absence of non-epistemic values. Hence only epistemic values such as accuracy, consistency, scope, simplicity, and fruitfulness should play a role in scientific research (cf. Kuhn 1981), and even this is perceived as a possible threat to scientific integrity given the aesthetic and subjective, rather than rational and objective, thrust of such principles.

This view, however, faced challenges from two sides. In his influential article "The Scientist Qua Scientist Makes Value Judgments," Richard Rudner (1953) argues that scientists rarely make decisions under certainty but must decide at which evidential threshold to accept a hypothesis as true — and such is clearly subject to non-epistemic value considerations. For instance, scientists'

assessment of a particular substance as "safe for consumption" is always made in light of their evaluation of the consequences of getting it wrong.[4,5] Moreover, Helen Longino (1995) has questioned the distinction between epistemic and non-epistemic values. According to Longino, various empirical values can conflict in scientific research. Scientists might have to trade off, for instance, between empirical adequacy and simplicity, breadth of scope and accuracy, or consistency and novelty. This choice between epistemic values is necessitates non-epistemic value judgments which must be made also in light of the socio-political contexts of scientific research.[6] In light of this, Longino argues that scientific communities must be diverse and cultivate avenues for criticism in order to represent and render open to scrutiny differently situated knowledge. Longino's work paved the way for an increasingly influential strand of feminist philosophy of science which highlights how scientific knowledge is always situated and shaped by non-epistemic social, cultural, moral, or political factors (Douglas 2009; Elliott 2017).

Such work is often aligned with an "in practice approach" to the philosophy of science. Philosophers of science in practice reframe central question in the philosophy of science as not about the abstract products but the practices of science (cf. Soler et al 2014). Doing so, can further help bring to attention the many non-epistemic features of science. As Ankeny and coauthors (2011) note: "Examining the goals underlying the activities associated with science also forces us to focus not only on epistemological considerations but also on the values, norms, and ideals inherent in the pursuit of scientific knowledge."

II.2 Ethics and epistemology in AI

Whereas philosophers of science have focused on how ethical values shape the epistemic practices of scientists, technology ethicists have concentrated on how epistemic issues can give rise to moral concerns. Among the most researched issues in AI ethics are questions of how deep learning's epistemic opacity can lead to personal and social harms (Burrell 2016; Vaassen 2022). More recently, scholars have explored how epistemic failures (aka "hallucinations) of large learning models can spread harmful stereotypes (Hofmann et al. 2024), undermine social trust (Deng et al. 2025), or threaten democratic stability (Coeckelbergh 2025).

This often-unidirectional exploration of epistemic and ethical issues is challenged by two recent publications. In their 2024 paper, Giorgia Pozzi and Juan M. Durán (2025) oppose what they call the "informativeness account" in AI ethics. The informativeness account limits itself to how epistemological issues can cause ethical concerns without exploring how ethical issues, in turn, might relate to the epistemology of AI. As a result, epistemic considerations are presented as (i) instrumental to potential ethical issues but also as conceptually prior and, thus, (ii) autonomous of them. In contrast, Pozzi and Durán argue that ethical issues can and should bear on the

[4] See also (K. Steele, 2012) for a development of his original argument.

[5] This also applies in cases where scientists provide only a probability, since this is again subject to the risk of getting things wrong, recursing indefinitely.

[6] Johnson (2023) brings these accounts to bear on the value-ladenness of algorithms.

epistemology of AI systems. Focusing on the case of diagnostic AI in medicine, they illustrate how personal values of a patient must feature into the recommendations or explanations the model provides. Ethics and epistemology of AI, so the authors argue, must be assessed in light of one another.

Their arguments follow closely-related work by Federica Russo, Eric Schliesser, and Jean Wagemans (2024). Russo and coauthors argue for an "epistemology-cum-ethics" — an integrated framework that explicitly considers and proactively incorporates values within the design, implementation, and evaluation of AI systems. Their argument takes computational reliabilism as their starting point. Computational reliabilism argues that the reliability of AI systems depends not exclusively on AI outcomes but more broadly on the reliability of the entire process of design, implementation, and use.

Russo and coauthors raise a complication for computational reliabilism. Assessing the reliability of these processes, ideally, involves the inclusion of a wide range of stakeholders with diverse sets of expertise. Directly verifying their epistemic reliability, however, is challenging for those who are not technical experts on AI. In these cases, we must rely on axiological rather than epistemic considerations.[7] We must consider, for instance whether AI models have been designed with care or whether institutional and professional incentives and values are aligned with the reliability of AI systems. These considerations might then indirectly lend credibility to the epistemic reliability of the processes underlying the design, implementation, and use of AI systems. Consequently, their framework of "epistemology-cum-ethics" stresses how ethical matters must form an equipollent complement to epistemic considerations throughout the entire AI pipeline.

II.3 Ethics and epistemology of AI in science

Our approach in this chapter shares strong kinship with these and other explorations of the ethics and epistemology of AI. For instance, we follow existing work in AI ethics and the philosophy of science in practice by emphasizing the close relationship between ethical and epistemic considerations *along the entire design and implementation processes of AI in science* — including the background conditions of scientific research.

At the same time, both our method and aims differ from existing scholarship. Our focus is specifically on the application of AI in science. We seek to show that epistemic and ethical considerations surrounding AI in science are interlinked by *bringing into conversation scholarship in the moral philosophy of technology and the philosophy of science.* We do not pursue a distinct framework for the design, implementation, or evaluation of AI in science. Instead, we primarily seek to *attune* our readers to how responsible AI in science requires detailed engagement with the intricate connections between epistemological and ethical concerns. In doing so, we hope to counter superficial narratives that pit epistemic and ethical concerns against each other.

[7] In fact, such assessment if often challenging even for technical experts.

We favour this approach since ethics and epistemology of AI in science are entangled in ways that are highly complex and deeply situated. Fostering responsible AI requires engaging with how ethical and epistemic considerations play out in light of particular research settings profoundly shaped by practical, historical, cultural, environmental, social, and political dynamics. Our understanding of responsible AI in science is, thus, not reducible to one particular framework but best represented by recent imagery provided by Nancy Cartwright, Eleonora Montuschi and coauthors: the tangle.

Cartwright and coauthors (2022) argue that what matters for science is the reliability of scientific products for a particular purpose and in a particular context. Such reliability emerges from a virtuous *tangle* of diverse, interlinked findings, methods, instruments, and considerations. Scientific products are reliable to the extent that they are enmeshed in such a structurally sound tangle that is (i) rich, (ii) closely intertwined, and (iii) strongly embedded within a network of other successful tangles across domains.

Tangles are concrete. They involve concrete methodologies, findings, procedures, assumptions, models, instruments, and so on. They include not only epistemic but also ethical, social, and institutional features.[8] Moral considerations, institutional incentives, political structures, and social dynamics must interlink with the epistemic products and tools of science in ways that properly *support* and *constrain* scientific practice. Both sides are important. The features must mutually support each other in order to lend stability to successful scientific practice. At the same time, they must also mutually constrain practice in ways that hinder ethically and epistemically problematic science (4.3).

Within this picture, responsibly implementing AI in science requires us to integrate AI tools in ways that support rather than disrupt the structural integrity of these tangles — to carefully and skillfully interweave AI tools within a complex network of epistemic and ethical factors that supports the responsible use of AI in particular contexts and for particular purposes. There is no blueprint for this job. The best we can do here, thus, is to highlight some ways in which ethical and epistemic threads can be entangled with the development and use of AI in science, attune our readers to this complexity, and stress the need to engage with how it might play out in particular applications of AI for particular purposes.

III AI in science

We focus on the background conditions for AI in science, and particularly on the unequal access to computational resources, examining its ethical and epistemic implications. Bringing into conversation debates in the ethics of technology and philosophy of science, we seek to challenge both simplistic win-win narratives intended to justify extractivist practices as well as superficial

[8] The role of moral, social, and institutional factors in the tangle is touched upon only briefly by Cartwright and coauthors (2022, 64), as discussed in a recent review by Anna Alexandrova (2025).

accounts that pit epistemic and ethical concerns against each other. Instead, we illustrate how attention to diversity and inclusion is a requirement of responsible AI in science for both epistemic and ethical reasons. Our explorations are far from exhaustive but merely seek to exemplify how ethical and epistemic concerns interrelate across all stages along the entire design and implementation processes of AI in science — from the background conditions of research, to research practice, and the dissemination of science.

3.1 Compute, inequity, and participation

Developing advanced AI applications for science can require access to advanced computational resources and infrastructure. These resources are, however, disparately distributed, as work in political geography has extensively highlighted. Duncan McDuie-Ra and Kalervo Gulson (2020), for instance, note how "AI is developed in the Global North and in the technology hubs of East Asia. The rest of the world is minimally involved in its development […]." Especially in the Global South, resources required for the development and deployment of advanced AI are scarce due to high infrastructure costs, unreliable electricity, or regulatory constraints (UN General Assembly, 2023). In response, researchers have distinguished between the "Compute North" (which has public compute infrastructure relevant for AI development), the "Compute South" (which has less advanced public compute infrastructure relevant mostly for AI deployment) and "Compute Deserts" (with no public cloud compute particularly suitable for AI which includes all of the world's LMICs and LICs) (Lehdonvirta et al. 2024).

While such geopolitical research underscores the global "AI divide," it can also obscure just how centralized computational resources are. Even within the Global North, the computational resources or "compute" required for the development of large AI models are often only accessible to the most powerful industry players and a handful of the most prestigious research institutions (Kudiabor 2024).[9] This concentration of advanced compute in hands of few and the resulting exclusion of differently resourced communities from the development of advanced AI applications raises a number of ethical issues. Among the most widely discussed in the ethics of technology are the distribution of costs and benefits, as well concerns related to control.

3.2 Benefits, costs, and control

Within popular discourse, AI in science is often presented as a win-win narrative. AI will bootstrap and accelerate scientific achievements in ways that benefit all of humanity — solving climate change, curing disease, and ushering in a new area of "abundance" (cf. Klein & Thompson 2025). Closer inspection, however, reveals how benefits and costs of AI applications in science are often distributed unequally.

[9] Recent developments such as DeepSeek have challenged the presumed relationship between compute and model performance. Moreover, the environmental impact of training increasingly large AI models (in particular LLMs) raises questions regarding the desirability of ever more resource intensive AI development.

Consider, for instance, emerging AI applications in development economics. Under the banner of "data for development, " AI models are used to estimate poverty metrics in the Global South. The hope is that these predictions can help monitor existing relief measures and better target poverty aid. Developers are often quick to acknowledge the risks and costs that such applications entail. AI models commonly rely on sensitive data for their poverty predictions, such as high-resolution satellite imagery and, in particular, mobile phone records. Mobile phone records contain information about the precise location and personal communication patterns of subscribers. This information can not only be used to reveal an individual's intimate social relations but might also be predictive of their political affiliation, sexual orientation, or religious beliefs (cf. Dube et al. 2022). Moreover, communication patterns are challenging to anonymize effectively (de Montjoye et al. 2019; Kohli et al. 2024).

As readily as developers acknowledge ethical concerns, they often dismiss them just as swiftly. Researchers argue that while AI poverty predictions do pose privacy risks, affected populations are also the primary benefactors of AI poverty statistics heralded as paving the way for immediate, effective, and extensive poverty aid. These benefits, so the reasoning goes, categorically outweigh the costs associated with their development. This often-superficial accounting, however, exaggerates the advantages of AI poverty predictions for affected populations and obscures their value for other actors.

AI poverty predictions have only rarely found real-world application and, when employed, have failed to significantly surpass simple category-based targeting methods (cf. Aiken et al. 2022). While the current benefits for affected populations in the Global South are debatable at best, researchers in the Global North stand to profit from the technical capacity built through AI applications which often rely on sensitive informational resources that would be protected if not for their use in the development context (cf. Mann 2018). The capacities and skills acquired in the context of such research often translate rather immediate to commercial uses — a consideration that becomes particularly critical in light of the high frequency with which academics move into the AI industry (Ahmed et al. 2023).

AI poverty prediction stands representative for a dynamic present across AI in science but obscured by overly grandiose narratives of AI's benefit for humanity writ large. The costs of AI development in science, from environmental harms to privacy inflictions, are borne rather widely — and are often most strongly felt by already vulnerable populations. Conversely, economic benefits of AI-enabled science are often concentrated in the hands of those privileged enough to own or access the computational resources required for their development (Hao 2025).

Closely related to these economic benefits are questions regarding control. Unequal ownership and access of AI infrastructure can undermine affected populations' ability to govern how AI applications are developed and used to intervene in their lives. Emerging research explores this under the label of compute-based AI governance and highlights how juridical control is, in some ways, tied to where AI infrastructures are physically located or AI companies are

headquartered (Sastry et al., 2024). AI models in science are often hosted in the Global North and trained on data in the possession of institutions headquartered in the Global North. This results in limited authority of local governments and communities over AI models and their application.

Philosophers of technology have stressed the importance of *who controls* a technology also in light of how technologies themselves can afford *new means of control* and manipulation. Access to data and the development of new technological capacities through AI applications in science can imbue governmental and commercial actors with growing economic power and bureaucratic capacities. Skills and methods developed in research settings often translate immediately to the use of AI in public administration, marketing, surveillance, or even warfare. This highlights the need to understand inequity within AI applications in science also with respect to their potential to reinforce or shift power dynamics and their situatedness within a broader socio-economic and political struggle for control (Couldry and Mejias 2019; Pasquinelli 2023).

3.3 Epistemic injustice

The concentration of advanced computational resources and AI development in hands of few also raises fundamental justice-related questions. This includes issues of epistemic injustice. Epistemic injustice is relevant for us not only as a concern for the responsible use of AI in science but also as one of the few areas that have witnessed significant engagement between philosophy of science and ethics of technology (e.g., Kay et al 2024; Pozzi 2023; Symons & Alvarado 2022).

Miranda Fricker (2007a) originally distinguishes between testimonial and hermeneutical epistemic injustice. She defines testimonial injustice as cases where "a speaker receives an unfair deficit of credibility from a hearer owing to prejudice on the hearer's part" (Fricker, 2007b) and hermeneutical injustice as "the injustice of having some significant area of one's social experience obscured from collective understanding owing to persistent and wide-ranging hermeneutical marginalization" (2007a, p. 154). Inequities in the development of AI in science can interrelate with both testimonial and hermeneutical injustice.

As mentioned, resource constraints in the development of AI can further centralize epistemic activity and expertise in hands of few. This might lead to others being excluded from developing, understanding, and using conceptual and theoretical resources emerging out of this strand of research and, thus, to hermeneutical injustice. Consider, for instance, the case of medicine. When proprietary and opaque AI models are used to allocate healthcare resources or diagnose patients, patients can become unable to interrogate such decisions and articulate potential concerns (Pozzi 2023; Symons and Alvarado 2022). Similar concerns emerge in other areas of science as communities affected by scientific research are unable to grasp and interrogate AI models increasingly central to this research.

The concentration of epistemic activity through AI might also entrench testimonial injustice by playing into science's long-standing devaluation of perspectives from outside the Global North and undermining efforts toward greater inclusion that have just gained traction within the past two

decades (McDuie-Ra and Gulson 2020, 629). In other words, inequality in AI development might contribute to what Michela Massimi has recently called epistemic severing: "the act of cutting off some epistemic communities from the narrative of scientific knowledge production" (Massimi 2021, 1).

3.4 Diversity and reliability in the epistemology of science

We have highlighted how unequal access to computational resources can result in the disparate distribution of costs and benefits of AI applications, centralize new means of control, and entrench epistemic injustice in science. While such issues have received widespread attention, they are often explored exclusively in light of their moral significant absent engagement with the epistemic implications of this inequity. All too easily, the impression can arise that ethical and epistemic consequences of AI are independent if not conflicting. Within this image, increasing participation might be ethically mandated but might hamper the epistemic advancement of science. Attention and resources put toward increasing global participation in AI-enabled research, so the reasoning goes, could after all be invested in those research and industry powerhouses already at the cutting-edge of AI development.

Often, such reasoning is paired with a naïve belief in machine objectivity. AI, after all, is trained on "objective" data — on "hard facts" untainted by human mediation (Anderson 2008). And where there is no human mediation, there is also no need for diversity in perspectives. Such a narrative stands in stark contrast with current research in the history and philosophy of science. In recent years, scholars have stressed how diversity is not only a matter of ethical but also central to the epistemic reliability of science: Science itself is vindicated in diversity.

Drawing from feminist standpoint epistemology, philosophers of science stress how scientific claims, theories, models, and *data* are necessarily situated and shaped by non-epistemic social, cultural, moral, infrastructural, institutional, or political factors.[10] For scientific knowledge claims to be robust, reliable, and objective, they must be rendered subject to critical scrutiny from diverse viewpoints (Longino, 1990a). As Massimi (2022) puts it, science is necessarily perspectival. It is situated (emerging from a particular vantage point) and oriented (aiming toward one or more vanishing points). The reliability of scientific claims increases as intersecting perspectives provide congruent accounts of the same modally robust phenomena (Massimi, 2022). To frame it in yet different terms, science rests on the virtuous tangle of diverse interdependent inquiries relying on an array of methodologies and emerging from different standpoints (Cartwright et al., 2022).

This work in the philosophy of science highlights how a lack of research participation by affected communities is not only ethically but also epistemically problematic (cf. Leonelli 2023). The very robustness and richness of scientific knowledge is put at risk as infrastructural background conditions centralize computational resources required for the development of AI models in the hands of only the most powerful research and industry actors. As these models are starting to

[10] For the case of data, particularly relevant to AI, see (Leonelli 2015; 2019).

become the standard for scientific practice, users and their research necessarily inherit the assumptions, affordances, and understandings baked into these systems. This can result in epistemic uniformity or "scientific monocultures" (Messeri and Crockett 2024). And much like monocultures in farming, scientific monocultures lack resilience and can damage the health and sustainability of the scientific ecosystem. Our brief exploration shows that, rather than in conflict, ethical *and* epistemic considerations both demand greater diversity and inclusion in the development of AI.

IV Conclusion: Responsible Philosophy of AI in and for Science

The 'tangle" of values, technologies and research processes involves an ever-intertwined co-dependence between methods, instruments, assumptions and institutional arrangements, not to speak of the knowledge outputs and incentive structures that support research practices in the first place. Over time, some specific elements of a tangle can become so entrenched as to be part of the 'scaffold' of science, to put it with James Griesemer (R. A. Ankeny et al. 2025). Parts of research that are no longer questioned or even critically examined, since they are taken fully for granted as part of its basic architecture.

When this reification happens to problematic scaffolds, such as the use of AI tools shaped only by the most powerful actors or the overdependence on AI solutions offering convenience at the potential cost of research quality, it is crucial to have conceptual instruments to disentangle the elements at hand and examine those relationships anew (Dupré and Leonelli 2022). It is this timely task that an engaged combination of philosophy of science and ethics of technology can help to achieve: an approach that can both track and understand scientific developments in action, and make them accountable to their broader stakeholders and implications.

In this chapter, we briefly illustrated how responsibly implementing AI in science requires attention to inequities in the background conditions of AI for it to be both epistemically reliable and morally desirable. In doing so, our aim was modest: to attune our readers to how responsible AI in science requires detailed engagement with the intricate connections between epistemological and ethical concerns. Rather than regarding ethics and epistemology of AI as independent or even conflicting, researchers must interrogate and skillfully interweave AI tools within a complex network of epistemic and ethical factors that mutually support its responsible use in particular contexts and for particular purposes. Our brief exploration underscored this intricate relationship. At the same time, complexity should not distract from the importance of a simple heuristic for the responsible integration of AI in science: *Epistemically better AI is often also ethically better AI. And ethically better AI, in the long run, is often also epistemically better AI.*

V References

Ahmed, Nur, Muntasir Wahed, and Neil C. Thompson. 2023. 'The Growing Influence of Industry in AI Research'. *Science* 379 (6635): 884–86. https://doi.org/10.1126/science.ade2420.

Alexandrova, Anna. 2025. 'The Tangle of Science: Reliability Beyond Method, Rigour, and Objectivity, by Nancy Cartwright, Jeremy Hardie, Eleonora Montuschi, Matthew Soleiman, and Ann C. Thresher'. *Mind* 134 (534): 612–20. https://doi.org/10.1093/mind/fzad067.

Anderson, Chris. 2008. 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete'. *Wired*. https://www.wired.com/2008/06/pb-theory/.

Ankeny, Rachel A., Michael R. Dietrich, and Sabina Leonelli. 2025. *Scaffolding: Selected Contributions of James R. Griesemer to History, Philosophy, and Biology*. Springer Nature.

Ankeny, Rachel, Hasok Chang, Marcel Boumans, and Mieke Boon. 2011. 'Introduction: Philosophy of Science in Practice'. *European Journal for Philosophy of Science* 1 (3): 303. https://doi.org/10.1007/s13194-011-0036-4.

Bertolaso, Marta, and Fabio Sterpetti, eds. 2020. *A Critical Reflection on Automated Science: Will Science Remain Human?* Vol. 1. Human Perspectives in Health Sciences and Technology. Springer International Publishing. https://doi.org/10.1007/978-3-030-25001-0.

Boge, Florian J. 2022. 'Two Dimensions of Opacity and the Deep Learning Predicament'. *Minds and Machines* 32 (1): 43–75. https://doi.org/10.1007/s11023-021-09569-4.

Burrell, Jenna. 2016. 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms'. *Big Data & Society* 3 (1): 1. https://doi.org/10.1177/2053951715622512.

Cartwright, Nancy, Jeremy Hardie, Eleonora Montuschi, et al. 2022. *The Tangle of Science: Reliability Beyond Method, Rigour, and Objectivity*. Oxford University Press.

Coeckelbergh, Mark. 2025. 'LLMs, Truth, and Democracy: An Overview of Risks'. *Science and Engineering Ethics* 31 (1): 4. https://doi.org/10.1007/s11948-025-00529-0.

Couldry, Nick, and Ulises A. Mejias. 2019. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. 1st edition. Stanford University Press.

Deng, Chengyuan, Yiqun Duan, Xin Jin, et al. 2025. 'Deconstructing the Ethics of Large Language Models from Long-Standing Issues to New-Emerging Dilemmas: A Survey'. *AI and Ethics*, ahead of print, August 13. https://doi.org/10.1007/s43681-025-00797-3.

Douglas, Heather E. 2009. *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Pre.

Dupré, John, and Sabina Leonelli. 2022. 'Process Epistemology in the COVID-19 Era: Rethinking the Research Process to Avoid Dangerous Forms of Reification'. *European Journal for Philosophy of Science* 12 (1): 20. https://doi.org/10.1007/s13194-022-00450-4.

Elliott, Kevin Christopher. 2017. *A Tapestry of Values: An Introduction to Values in Science*. Oxford University Press.

Freiesleben, Timo, and Thomas Grote. 2023. 'Beyond Generalization: A Theory of Robustness in Machine Learning'. *Synthese* 202 (4): 109. https://doi.org/10.1007/s11229-023-04334-9.

Hao, Karen. 2025. *Empire of AI: Inside the Reckless Race for Total Domination*. Penguin Books Limited.

Hofmann, Valentin, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. 'AI Generates Covertly Racist Decisions about People Based on Their Dialect'. *Nature* 633 (8028): 147–54. https://doi.org/10.1038/s41586-024-07856-5.

Johnson, Gabbrielle M. 2023. 'Are Algorithms Value-Free?' *Journal Moral Philosophy* 21 (1–2): 1–35. https://doi.org/10.1163/17455243-20234372.

Kay, Jackie, Atoosa Kasirzadeh, and Shakir Mohamed. 2024. 'Epistemic Injustice in Generative AI'. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (1): 1. https://doi.org/10.1609/aies.v7i1.31671.

Klein, Ezra, and Derek Thompson. 2025. *Abundance*. Avid Reader Press.

Kohli, Nitin, Emily Aiken, and Joshua E. Blumenstock. 2024. 'Privacy Guarantees for Personal Mobility Data in Humanitarian Response'. *Scientific Reports* 14 (1): 28565. https://doi.org/10.1038/s41598-024-79561-2.

Kudiabor, Helena. 2024. 'AI's Computing Gap: Academics Lack Access to Powerful Chips Needed for Research'. *Nature* 636 (8041): 16–17. https://doi.org/10.1038/d41586-024-03792-6.

Kuhn, Thomas S. 1981. 'Objectivity, Value Judgment, and Theory Choice'. In *Review of Thomas S. Kuhn The Essential Tension: Selected Studies in Scientific Tradition and Change*, edited by David Zaret. Duke University Press.

Lehdonvirta, Vili, Bóxī Wú, and Zoe Hawkins. 2024. 'Compute North vs. Compute South: The Uneven Possibilities of Compute-Based AI Governance Around the Globe'. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (1): 1. https://doi.org/10.1609/aies.v7i1.31683.

Leonelli, Sabina. 2015. 'What Counts as Scientific Data? A Relational Framework'. *Philosophy of Science* 82 (5): 5. https://doi.org/10.1086/684083.

Leonelli, Sabina. 2019. 'What Distinguishes Data from Models?' *European Journal for Philosophy of Science* 9 (2): 2. https://doi.org/10.1007/s13194-018-0246-0.

Longino, Helen E. 1995. 'Gender, Politics, and the Theoretical Virtues'. *Synthese* 104 (3): 383–97.

Mann, Laura. 2018. 'Left to Other Peoples' Devices? A Political Economy Perspective on the Big Data Revolution in Development'. *Development and Change* 49 (1): 3–36. https://doi.org/10.1111/dech.12347.

Massimi, Michela. 2021. 'Epistemic Severing and Epistemic Trademarking. Two Garden Varieties of Epistemic Injustice in Science'. November 10. https://philsci-archive.pitt.edu/19844/.

McDuie-Ra, Duncan, and Kalervo Gulson. 2020. 'The Backroads of AI: The Uneven Geographies of Artificial Intelligence and Development'. *Area* 52 (3): 626–33. https://doi.org/10.1111/area.12602.

Messeri, Lisa, and M. J. Crockett. 2024. 'Artificial Intelligence and Illusions of Understanding in Scientific Research'. *Nature* 627 (8002): 49–58. https://doi.org/10.1038/s41586-024-07146-0.

Montjoye, Yves-Alexandre de, Jake Kendall, and Cameron F. Kerry. 2019. 'Enabling Humantiarian Use of Mobile Phone Data'. *Trusted Data, Revised and Expanded Edition: A New Framework for Identity and Data Sharing*, 167.

Oimann, Ann-Katrien, and Fabio Tollon. 2025. 'Responsibility Gaps and Technology: Old Wine in New Bottles?' *Journal of Applied Philosophy* 42 (1): 337–56. https://doi.org/10.1111/japp.12763.

Pasquinelli, Matteo. 2023. *The Eye of the Master: A Social History of Artificial Intelligence*. Verso Books.

Pozzi, Giorgia. 2023. 'Automated Opioid Risk Scores: A Case for Machine Learning-Induced Epistemic Injustice in Healthcare'. *Ethics and Information Technology* 25 (1): 3. https://doi.org/10.1007/s10676-023-09676-z.

Pozzi, Giorgia, and Juan M. Durán. 2025. 'From Ethics to Epistemology and Back Again: Informativeness and Epistemic Injustice in Explanatory Medical Machine Learning'. *AI & SOCIETY* 40 (2): 299–310. https://doi.org/10.1007/s00146-024-01875-6.

Royal Society. 2024. *Science in the Age of AI*. Royal Society. https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/.

Russo, Federica, Eric Schliesser, and Jean Wagemans. 2024. 'Connecting Ethics and Epistemology of AI'. *AI & SOCIETY* 39 (4): 1585–603. https://doi.org/10.1007/s00146-022-01617-6.

Selbst, Andrew D., Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. 'Fairness and Abstraction in Sociotechnical Systems'. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York, NY, USA), FAT* '19, January 29, 59–68. https://doi.org/10.1145/3287560.3287598.

Soler, Léna, Sjoerd Zwart, Michael Lynch, and Vincent Israel-Jost. 2014. *Science after the Practice Turn in the Philosophy, History, and Social Studies of Science*. Routledge.

Symons, John, and Ramón Alvarado. 2022. 'Epistemic Injustice and Data Science Technologies'. *Synthese* 200 (2): 87. https://doi.org/10.1007/s11229-022-03631-z.

Vaassen, Bram. 2022. 'AI, Opacity, and Personal Autonomy'. *Philosophy & Technology* 35 (4): 88. https://doi.org/10.1007/s13347-022-00577-5.

Vallor, Shannon. 2024. *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*. Oxford University Press.

Wynsberghe, Aimee van. 2021. 'Sustainable AI: AI for Sustainability and the Sustainability of AI'. *AI and Ethics* 1 (3): 213–18. https://doi.org/10.1007/s43681-021-00043-6.