

HUMAN-AI EPISTEMIC PARTNERSHIPS: INVESTIGATING THE DYNAMICS OF TRUST AND AUTHORITY

Terzungwe Emmanuel Igyom

Department of Philosophy, Benue State University
emmanueligyom@gmail.com

Abstract

The speedy incorporation of Artificial Intelligence (AI) into knowledge-generation and decision-making processes has reshaped traditional epistemic relationships. This article explores the emerging concept of Human-AI epistemic partnerships, focusing on the evolving dynamics of trust and authority within these hybrid systems. As AI systems increasingly participate in domains requiring judgment, interpretation, and inference, such as scientific research, healthcare, and law, the question of how humans assign epistemic authority to non-human agents becomes urgent. Drawing from interdisciplinary perspectives in epistemology, cognitive science, and human-computer interaction, the article examines how trust is cultivated, maintained, or eroded in human-AI collaborations. We analyze case studies where AI systems are not merely tools but are treated as epistemic collaborators or agents whose outputs are trusted, questioned, or even challenged by human users. Special attention is given to the asymmetry in understanding: while humans often lack full transparency into AI decision-making, they nonetheless rely on these systems in high-stakes scenarios. The

article proposes a framework for evaluating trustworthiness in AI systems, distinguishing between justified trust and epistemically irresponsible reliance. Furthermore, we interrogate the shifting boundaries of epistemic authority and what it means to "know" or "believe" something in a context where AI contributes to the formation of those beliefs. Ultimately, the article argues for a distinct model of epistemic partnership that accounts for both the affordances and limitations of AI. Such a model is essential for guiding the ethical and effective integration of AI into knowledge practices that shape human understanding and societal outcomes.

Keywords: Human, AI, Epistemic partnerships, Trust, Authority.

Introduction

In recent years, artificial intelligence (AI) has moved from the periphery of technological innovation to the center of knowledge production and decision-making across diverse domains. Once limited to computational tools that augmented human labor, AI systems are increasingly taking on roles that resemble epistemic agents or entities that generate, evaluate, and justify knowledge. This evolution marks a significant shift in how humans relate to machines, positioning AI not merely as a passive tool but as an active partner in epistemic practices.

However, this shift brings with it complex challenges. As AI systems begin to influence and in some cases,

shape human understanding, the dynamics of trust between human users and AI agents become increasingly fraught. Questions arise regarding who or what holds epistemic authority in these interactions. Can AI systems be considered credible knowers? And if so, under what conditions should their outputs be trusted or contested? The opacity of AI decision-making processes and the speed at which these technologies evolve further complicate the relationship, raising concerns about autonomy, accountability, and the legitimacy of knowledge generated in these hybrid systems.

This article seeks to explore how trust and authority are negotiated within human-AI epistemic partnerships. It examines the conceptual and practical tensions that emerge when humans and AI co-produce knowledge, drawing on interdisciplinary insights from philosophy, cognitive science, and technology studies.

Ultimately, this study argues that the human-AI epistemic relationship is a dynamic negotiation of trust and authority, one that has profound implications for the legitimacy of knowledge, the autonomy of human agents, and the distribution of epistemic responsibility in an increasingly AI-mediated world.

Historical and Philosophical Foundations of Human-AI Epistemic Relationships

Human Trust in Tools and Technology: Human beings have long engaged in the use of tools and instruments to extend their cognitive reach. The historical arc from rudimentary instruments to modern artificial intelligence (AI) technologies reveals a consistent interplay between technological innovation and epistemic transformation. One significant thing here is the question of trust: how and why humans come to place confidence in technologies to produce or support knowledge claims.

A pivotal historical analogy emerges when comparing the adoption of the telescope in the early modern scientific revolution to the integration of statistical models and algorithms in contemporary scientific practice. When Galileo first used the telescope to observe celestial bodies, his findings were met with skepticism, not solely due to their radical nature but because of the perceived unreliability of the instrument itself. As Shapin notes, “the acceptance of telescope-aided observations required not only trust in the device but also a reconfiguration of epistemic norms to accommodate technologically mediated seeing” (Shapin 38). This shift in trust did not occur overnight; it necessitated a cultural and philosophical adjustment that gradually reshaped scientific methodology.

In a similar vein, the adoption of statistical tools in the 19th and 20th centuries brought about a recalibration of what constituted reliable evidence. The rise of probabilistic reasoning in medicine, social sciences, and physics demonstrated a growing comfort with forms of knowledge that were not strictly deterministic but were instead probabilistically warranted (Porter 45). Today, we see a comparable moment unfolding with AI and machine learning systems, which often operate as "black boxes" offering outputs without transparent rationales. Despite or perhaps because of their opacity, these systems are increasingly trusted in high-stakes domains such as diagnostics, risk assessment, and scientific discovery.

The philosophical challenge here is not new: throughout history, new technologies have challenged prevailing epistemic norms. Just as the telescope required a rethinking of the nature of observation, AI demands a rethinking of the nature of justification. Trust in AI systems reflects a broader historical pattern, where tools first generate skepticism, then acceptance, and finally epistemic dependence, often without sufficient philosophical reflection on the foundations of that trust.

The Epistemic Status of Machines: The growing reliance on AI systems for decision-making and knowledge generation invites renewed scrutiny of their epistemic status. Are these systems merely

sophisticated calculators or can they be considered epistemic agents in their own right? To answer this, one must turn to foundational philosophical questions, particularly those posed by David Hume and René Descartes, whose contrasting views offer valuable lenses for analyzing AI.

David Hume famously challenged the reliability of inductive reasoning, arguing that our beliefs about causation and the future are based not on rational deduction but on habituated expectations (Hume 89). This skeptical view poses a significant challenge to any epistemic system, human or artificial, that relies on patterns of data to make predictions. AI, especially in its machine learning forms, is deeply entrenched in inductive reasoning; it identifies correlations and extrapolates from them, often without access to the underlying causal mechanisms.

From a Humean standpoint, the outputs of AI systems cannot be justified by appeal to necessity or certainty but only by their pragmatic success in past applications. As Binns et al. argue, “machine learning systems replicate Hume’s model of inference by learning regularities without understanding underlying causes” (Binns et al. 546). This leads to a tension: while such systems can be extraordinarily effective, their epistemic authority remains philosophically tenuous. Hume’s skepticism urges caution in equating predictive success with genuine understanding or rational justification.

Moreover, Hume's insights into probability suggest that even probabilistic knowledge is grounded in subjective experience and belief, rather than objective necessity. In the context of AI, this underscores a fundamental limitation: probabilistic outputs do not constitute knowledge in the traditional philosophical sense, but rather sophisticated forecasts whose reliability is always contingent and revisable.

Contrasting with Humean empiricism is the Cartesian model of knowledge, which emphasizes clarity, certainty, and rational deduction. Descartes' project aimed to ground knowledge in indubitable foundations, a goal that seems increasingly elusive in the age of probabilistic AI. If Descartes sought certainty through introspective reasoning, AI systems achieve functionality through data-driven generalizations, an approach that would likely have failed to meet Cartesian standards of epistemic rigor.

However, some have sought to reconcile Cartesian and empiricist elements in the evaluation of AI. For example, rationalist approaches to explainable AI attempt to reintroduce transparent, rule-based structures that mimic Cartesian clarity (Lipton 3). Yet these efforts often run up against the practical limitations of complex neural networks, which resist easy explanation or reduction. As Burrell observes, "the opacity of algorithmic systems is not merely a technical problem, but a philosophical one that

challenges the very idea of accountability and rational justification” (Burrell 1).

The epistemic frameworks of Descartes and Hume thus provide contrasting evaluations of AI systems. From a Cartesian view, their opacity and lack of foundational certainty are deeply problematic. From a Humean perspective, their reliance on past regularities and inductive reasoning mirrors human cognition, but it also inherits its philosophical vulnerabilities. As such, AI systems may be best understood not as epistemic agents in their own right, but as extensions of human epistemic practices, tools that amplify our abilities while reproducing our philosophical limitations.

Contemporary Scholarly Debates on AI as Epistemic Agents

The emergence of artificial intelligence (AI) as a partner in knowledge production and decision-making has triggered intense scholarly debates about its epistemic status and moral implications. As AI systems become increasingly integrated into scientific research, healthcare, legal frameworks, and public policy, a central question arises: can these systems be considered epistemic agents in their own right, or are they merely tools enhancing human inquiry? These debates pivot around two primary axes: those who champion the epistemic capabilities of AI, and those who raise concerns about the

implications of assigning epistemic authority to non-human entities.

Proponents of AI as Epistemic Agents: Supporters of AI as epistemic agents emphasize the technological affordances of machine learning systems, arguing that they possess unique strengths in objectivity, consistency, and performance, qualities long sought in epistemic practices. These scholars posit that AI's ability to process vast datasets, identify non-obvious patterns, and update models based on new data allows it to surpass human limitations in many domains.

Luciano Floridi, a key proponent of this view, argues that AI systems constitute a form of “artificial epistemology,” not merely replicating but expanding the boundaries of human knowledge (Floridi 15). Floridi introduces the concept of “distributed epistemic agency,” wherein AI is seen as part of an integrated epistemic network that contributes substantively to knowledge production. He maintains that, although AI lacks consciousness or self-reflective awareness, its role in contributing to justified true beliefs in scientific and social contexts cannot be dismissed.

Shannon Vallor also contributes to this debate by examining how AI systems can exhibit traits associated with epistemic virtue, such as reliability and impartiality. She suggests that the epistemic

authority of AI systems lies not in their independence but in their ability to enhance human capabilities through virtuous collaboration (Vallor 202). For Vallor, epistemic agency is not an all-or-nothing category but a gradient that includes both human and non-human entities working in epistemically fruitful ways.

Proponents also underscore the ability of AI to avoid some of the cognitive biases that plague human reasoning. Unlike humans, who are susceptible to confirmation bias, availability heuristics, and motivated reasoning, properly designed AI systems can make decisions based solely on evidence and statistical inference. This quality is particularly valued in fields like medicine, finance, and criminal justice, where objectivity is critical to ethical and accurate decision-making.

Critics and Concerns: While the optimism surrounding AI's epistemic capacities is compelling, numerous scholars caution against uncritical acceptance of AI as a full epistemic agent. Critics emphasize the opaqueness of machine learning systems, the challenges of accountability, and the moral and social risks of over-reliance on AI.

One major concern is the “black box” nature of many AI systems. As algorithms become more complex, especially those using deep learning techniques, their internal decision-making processes become

increasingly opaque to users and even to their designers. This raises critical issues of accountability: if an AI system makes a harmful or erroneous decision, who is responsible? Without transparency, it becomes difficult to trace the epistemic justifications behind AI outputs, thereby undermining trust in these systems as epistemic agents (Burrell 9).

The question of moral agency further complicates the issue. Unlike human knowers, AI lacks intentions, desires, and the capacity for moral judgment. Scholars like Joanna Bryson argue that conferring agency on AI risks absolving humans of responsibility, creating a dangerous moral gap in contexts where ethical reasoning is essential (Bryson 173). As machines increasingly participate in high-stakes decisions from parole recommendations to military targeting, the absence of moral accountability becomes a pressing concern.

Furthermore, critics like Miranda Fricker have brought attention to the problem of epistemic injustice, which arises when certain groups are unfairly discredited or marginalized in knowledge practices. Although Fricker's work originally focused on human-to-human interactions, scholars have extended her insights to AI systems. If the datasets used to train AI are biased, reflecting historical patterns of discrimination, then the outputs will perpetuate and even exacerbate existing injustices (Fricker 1). In such cases, AI not only fails as an

epistemic agent but also becomes a vehicle for epistemic harm.

Another significant issue is the potential for human deskilling and epistemic dependency. As AI systems take over more cognitive tasks, humans may lose the skills necessary for critical thinking, interpretation, and decision-making. This concern echoes Socratic warnings about the externalization of memory and knowledge, but in a far more technologically sophisticated context. If people begin to defer to AI systems uncritically, they risk becoming passive recipients rather than active participants in epistemic processes. This phenomenon leads to what some have called “epistemic outsourcing,” where human agents lose both the ability and the responsibility to know (Coeckelbergh 189).

These concerns point to a need for a nuanced understanding of AI's role in epistemic systems. While AI may possess certain epistemic virtues such as speed, consistency, and pattern recognition, it lacks others, including the capacity for contextual understanding, moral discernment, and socio-political sensitivity.

Trust-Building Mechanisms in Human-AI Systems

As artificial intelligence systems increasingly permeate domains such as healthcare, finance, military strategy, and education, fostering trust between humans and AI agents becomes paramount.

Trust is not merely a byproduct of advanced performance but a product of complex socio-technical interactions that involve transparency, explainability, human oversight, and mutual learning.

Transparency and Explainability: Transparency and explainability are fundamental to trust in AI systems. Interpretability, the degree to which a human can understand the cause of a decision, serves as a conduit for users to make sense of AI behavior. As Doshi-Velez and Kim define it, interpretability involves “the ability to explain or to present in understandable terms to a human” the decision-making processes of an AI model (Doshi-Velez and Kim 2). For example, in high-stakes domains like medical diagnostics, clinicians are more likely to trust AI tools when the rationale behind the output is clear and coherent with human expertise.

Interpretable systems bridge the epistemic gap between opaque algorithmic reasoning and human cognitive models. In this context, symbolic AI models such as decision trees and rule-based systems are often favored for their intelligibility over complex deep learning models (Guidotti et al. 56). However, interpretability is not only about providing technical explanations; it is also about tailoring those explanations to the cognitive needs of diverse users. Ribeiro et al.'s LIME (Local Interpretable Model-Agnostic Explanations) is a prime example of a method that explains predictions by approximating

the local behavior of the model in human-readable form, thus enhancing user confidence (Ribeiro, Singh, and Guestrin 1135).

Trade-Offs Between Accuracy and Explainability:

However, an ongoing challenge in the field is balancing model accuracy with explainability. Deep learning models, particularly in computer vision and natural language processing, often achieve superior performance at the cost of interpretability. Lipton refers to these models as “black boxes,” emphasizing the difficulty in tracing how outputs are derived (Lipton 2). Consequently, developers must weigh whether the marginal gains in predictive power justify sacrificing interpretability, especially in ethically sensitive or legally regulated applications.

Moreover, simplistic explanations may lead to overconfidence in models that are not robust under all circumstances. This phenomenon, known as the “illusion of explanatory depth,” can give users a false sense of security (Rozenblit and Keil 521). As such, transparent systems must also communicate their limitations and uncertainty margins. As Ananny and Crawford argue, transparency should be reframed not as a guarantee of truth but as a means to cultivate epistemic humility in the face of complex AI reasoning (Ananny and Crawford 974).

Human Oversight and Calibration

Beyond algorithmic transparency, human oversight plays a crucial role in establishing dynamic and adaptive trust in AI systems. Trust calibration involves aligning a user's level of trust with the actual reliability of an AI system (Lee and See 51). Overtrust and undertrust both undermine effective collaboration: the former can lead to uncritical acceptance of flawed outputs, while the latter results in the underutilization of valuable tools.

Adaptive trust frameworks integrate user feedback, environmental context, and system performance to modulate trust appropriately. These frameworks are particularly vital in semi-autonomous systems such as aviation or autonomous driving, where the human operator must maintain situational awareness and be prepared to intervene. For instance, Endsley's model of situation awareness is often cited as foundational in designing systems that support human operators in real-time decision-making (Endsley 35).

Feedback Loops and Mutual Learning: Establishing feedback loops where both humans and AI systems learn from each other is essential for sustaining long-term trust. Mutual learning implies that AI systems adapt to human preferences, values, and behaviors over time, while users develop a better understanding of AI capabilities and constraints. This symbiotic process requires mechanisms for user input, behavior modeling, and interactive learning.

Designing AI that evolves based on user interaction echoes the principles of human-centered AI (HCAI), as advocated by Shneiderman, which emphasizes the importance of augmenting human abilities rather than replacing them (Shneiderman 6). For example, AI tutoring systems in education adapt to individual learners by analyzing feedback and adjusting instructional strategies accordingly. Similarly, in collaborative filtering systems, user ratings contribute to refining recommendations, thus forming a feedback-rich trust loop.

Furthermore, explainability itself can be a product of feedback, as users indicate which types of explanations are most useful to them. This reflects a participatory approach to AI design, where trust is not pre-built but co-constructed through continuous dialogue between humans and machines.

Implications and Future Directions

Rethinking Epistemic Agency: The growing integration of artificial intelligence (AI) in knowledge production and dissemination compels a reassessment of what it means to be an epistemic agent. Traditionally, epistemic agency has been ascribed to human individuals who possess the capacity for reason, intention, and accountability. However, with AI systems increasingly generating data-driven insights, forming predictions, and even engaging in interpretive tasks (such as language translation or

legal analysis), the question emerges: Should AI be granted epistemic status?

Some scholars advocate for a reconfiguration of epistemic roles, suggesting that AI systems, while lacking consciousness, play a functional epistemic role in collaborative human-AI systems. According to Rolin, epistemic agency need not be fully autonomous but can be distributed across networks, including non-human actors (Rolin 168). From this perspective, AI systems contribute epistemically insofar as they participate in processes that result in knowledge generation and validation, albeit mediated through human interpretation and oversight.

Yet, the philosophical tradition remains cautious. Miranda Fricker's work on epistemic injustice underscores the importance of moral and social responsibility in the practice of epistemic agency (Fricker 1). Since AI lacks moral accountability and cannot engage in dialogue or reason about its beliefs, many argue that attributing epistemic agency to AI risks diluting the concept beyond recognition. As Binns points out, there is a critical distinction between "epistemic function" and "epistemic authority", the latter involving a normative weight that AI, at least currently, cannot bear (Binns 221).

Nonetheless, as AI systems become increasingly autonomous in their operation and are trusted in

high-stakes decisions, such as medical diagnoses or criminal justice assessments, the idea of “epistemic reliance” becomes crucial. Trust in these systems often mirrors trust in human experts, suggesting a shift in how epistemic agency is socially constructed (Zerilli et al. 72). Future research must thus navigate the blurry boundaries between tool and co-agent, particularly in terms of responsibility, trust, and interpretability.

Ethical and Political Dimensions: The ethical and political dimensions of AI are deeply entangled with questions of bias, fairness, and societal trust. Machine learning systems trained on historical data risk perpetuating and amplifying existing societal inequalities. A well-documented example is facial recognition technology, which has been shown to have significantly higher error rates for darker-skinned individuals, particularly women (Buolamwini and Gebru 87). These technical failures reflect deeper issues of epistemic exclusion: who gets represented and whose knowledge counts in AI systems?

Furthermore, fairness in algorithmic decision-making is not merely a technical issue but also a normative and political one. As Selbst et al. argue, attempts to de-bias algorithms often overlook the contextual and institutional factors that generate injustice in the first place (Selbst et al. 59). Rather than assuming that fairness can be coded into algorithms through

mathematical formalism, they call for a broader socio-technical approach that includes affected communities in the design and evaluation of AI systems.

Societal trust in AI is likewise fragile. Scandals involving biased algorithms or opaque decision-making, such as the COMPAS algorithm used in U.S. courts, have revealed a general unease about the legitimacy of AI judgments. Public trust requires more than transparency; it demands accountability structures, meaningful oversight, and the ability to contest and redress algorithmic decisions (Eubanks 143).

To cultivate an ethically responsible AI ecosystem, future directions must emphasize participatory design, interdisciplinary collaboration, and the embedding of normative values into both the design and deployment phases of AI systems. This entails moving beyond compliance checklists toward a genuinely reflective, justice-oriented practice.

Governance and Norms: The rapid evolution of AI technologies outpaces the development of corresponding governance frameworks. Existing norms and legal regimes often fail to address the unique challenges posed by autonomous and semi-autonomous systems. For example, the General Data Protection Regulation (GDPR) in the EU introduced the right to explanation, yet it remains unclear how

such rights can be operationalized in the context of complex machine learning models (Wachter, Mittelstadt, and Floridi 75).

Governance of AI must therefore be both proactive and adaptive. Scholars and policymakers have proposed multi-level governance models that combine international standards, national regulations and local ethical review boards (Floridi et al. 691). These models emphasize the need for transparency, accountability, and contestability in AI systems and recommend mechanisms such as algorithmic audits, impact assessments, and public engagement initiatives.

Crucially, the norms that underpin AI governance must reflect democratic values and avoid technosolutionism. As Crawford warns, over-reliance on technical fixes can obscure deeper societal problems and depoliticize what are fundamentally political choices (Crawford 194). Governance frameworks should aim not only to minimize harm but also to promote epistemic justice, inclusivity, and deliberative democracy.

Looking forward, the responsible integration of AI into society depends on a delicate balance: embracing innovation while safeguarding human rights, promoting fairness, and ensuring that epistemic authority remains accountable. This

requires continuous dialogue between technologists, ethicists, legal scholars, and the public.

Conclusion

The evolving landscape of human-AI epistemic partnerships has underscored a critical need to understand how trust and authority are negotiated between human agents and artificial systems. This investigation has examined key dimensions of that relationship, including the epistemological frameworks that underlie trust in algorithmic outputs, the conditions under which authority is granted or withheld from AI systems, and the broader ethical and social implications that arise from these interactions. Through analysis of both philosophical perspectives and empirical research, it is clear that the dynamics of trust are deeply contextual and dependent not only on technical reliability but also on transparency, accountability, and the user's interpretive agency.

Reaffirming the central thesis, this study contends that effective and responsible human-AI epistemic partnerships depend on a nuanced, critically informed engagement with trust and authority. Trust in AI should not be blind nor merely functional, but must be anchored in a reflexive understanding of both human and machine limitations. Authority, likewise, must remain contingent, subject to ongoing evaluation rather than uncritically conferred.

Ultimately, as AI systems continue to mediate knowledge production and decision-making across domains, the challenge lies not in whether to trust these systems, but in how to construct frameworks of trust that are resilient, transparent, and responsive to human values. Cultivating such partnerships will require continuous dialogue between technologists, philosophers, and users, ensuring that AI remains a tool for epistemic empowerment rather than a source of unquestioned authority.

Works Cited

Ananny, Mike, and Kate Crawford. "Seeing without Knowing: Limitations of the Transparency

Ideal and Its Application to Algorithmic Accountability." *New Media & Society*, vol. 20, no. 3, 2018, pp. 973–989.

Angwin, Julia, et al. "Machine Bias." *ProPublica*, 23 May 2016

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

Baier, Annette. *Moral Prejudices: Essays on Ethics*. Harvard UP, 1994.

Binns, Reuben, et al. "It's Reducing a Human Being to a Percentage: Perceptions of Justice in

Algorithmic Decisions." *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ACM, 2018, pp. 1—14.

Binns, Reuben. "On the Apparent Conflict Between Individual and Group Fairness." *Proceedings*

- of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 221—229.
- Bryson, Joanna J. "Robots Should Be Slaves." *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, edited by Yorick Wilks, John Benjamins Publishing, 2010, pp. 63—74.
- Buolamwini, Joy and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research*, vol. 81, 2018, pp. 77—91.
- Burrell, Jenna. "How the Machine Thinks: Understanding Opacity in Machine Learning Algorithms." *Big Data & Society*, vol. 3, no. 1, 2016, pp. 1—12.
- Crawford, Kate. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale UP, 2021.
- Doshi-Velez, Finale, and Been Kim. "Towards A Rigorous Science of Interpretable Machine Learning." arXiv preprint, arXiv:1702.08608, 2017.
- Elliott, Kevin C. and David B. Resnik. "Science, Policy, and the Transparency of Values." *Environmental Health Perspectives*, vol. 122, no. 7, 2014, pp. 647—50.
- Endsley, Mica R. "Toward a Theory of Situation Awareness in Dynamic Systems." *Human Factors*, vol. 37, no. 1, 1995, pp. 32—64.

- Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018.
- Floridi, Luciano, and Josh Cowls. "A Unified Framework of Five Principles for AI in Society." *Harvard Data Science Review*, vol. 1, no. 1, 2019, pp. 1—15.
- Floridi, Luciano, et al. "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines*, vol. 28, no. 4, 2018, pp. 689—707.
- Floridi, Luciano. *The Logic of Information: A Theory of Philosophy as Conceptual Design*. Oxford UP, 2019.
- Fricker, Miranda. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford UP, 2007.
- Guidotti, Riccardo, et al. "A Survey of Methods for Explaining Black Box Models." *ACM Computing Surveys*, vol. 51, no. 5, 2018, pp. 1—42.
- Hardin, Russell. *Trust and Trustworthiness*. Russell Sage Foundation, 2002.
- Hume, David. *An Enquiry Concerning Human Understanding*. Edited by Tom L. Beauchamp, Oxford UP, 2000.
- Jumper, John, et al. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature*, vol. 596, no. 7873, 2021, pp. 583—89.
- Kanu, A. I., Addidi, T. & Kanu, C. C., "Artificial intelligence and cybercrime in Nigeria: Towards an

Ethical framework". *Dialogue and Universalism*. 32(1)
143-158, 2024.

Kanu, I. A. (2020). "Igwebuiké philosophy of science
and technology". *Nnadiabube Journal of Education*.
Vol. 5. No. 4, pp. 64-74.

Lee, John D., and Katrina A. See. "Trust in
Automation: Designing for Appropriate Reliance."

Human Factors, vol. 46, no. 1, 2004, pp. 50—
80.

Lipton, Zachary C. "The Mythos of Model
Interpretability." *Queue*, vol. 16, no. 3, 2018, pp.
30—57.

Mittelstadt, Brent D., et al. "The Ethics of Algorithms:
Mapping the Debate." *Big Data &*

Society, vol. 3, no. 2, 2016, pp. 1—21.

Porter, Theodore M. *Trust in Numbers: The Pursuit of
Objectivity in Science and Public Life*.

Princeton UP, 1995.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos
Guestrin. "Why Should I Trust You?:

Explaining the Predictions of Any Classifier."

Proceedings of the 22nd ACM SIGKDD

*International Conference on Knowledge
Discovery and Data Mining*, 2016, pp.

1135—1144.

Rini, Regina. "Deepfake Belief: The Epistemic Threat
of Deepfakes." *Philosophy & Technology*, vol. 33,
2020, pp. 61—77.

Rolin, Kristina. "The Epistemic Roles of Non-Human
Agents." *Studies in History and*

Philosophy of Science Part A, vol. 71, 2018, pp. 168—175.

Ross, Casey, and Blake Swetlitz. "IBM Pitched Watson as a Revolution in Cancer Care. Its Nowhere Close." *STAT News*, 5 Sept. 2017, <https://www.statnews.com/2017/09/05/watson-ibm-cancer/>.

Rozenblit, Leonid, and Frank Keil. "The Misunderstood Limits of Folk Science: An Illusion of Explanatory Depth." *Cognitive Science*, vol. 26, no. 5, 2002, pp. 521—562.

Selbst, Andrew D. "Disparate Impact in Big Data Policing." *Georgia Law Review*, vol. 52, 2017, pp. 109—95.

Selbst, Andrew D., et al. "Fairness and Abstraction in Sociotechnical Systems." *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 2019, pp. 59—68.

Shapin, Steven. *The Scientific Revolution*. U of Chicago P, 1996.

Shneiderman, Ben. "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy." *International Journal of Human-Computer Interaction*, vol. 36, no. 6, 2020, pp. 495—504.

Taddeo, Mariarosaria, and Luciano Floridi. "How AI Can Be a Force for Good." *Science*, vol.

361, no. 6404, 2018, pp. 751—52.

Topol, Eric J. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*.

Basic Books, 2019.

Vallor, Shannon. "Moral Deskillling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character." *Philosophy & Technology*, vol. 28, no. 1, 2015, pp. 107—124.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." *International Data Privacy Law*, vol. 7, no. 2, 2017, pp. 76—99.

Zerilli, John, et al. "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?" *Philosophy & Technology*, vol. 32, no. 4, 2019, pp. 661—683.

Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, 2019.