

Research Article

Epistemic Injustice in Algorithmic Systems: A Frickerian Analysis of AI-Driven Knowledge Production

Joel Sadiki Charo* , Kennedy Ondimu

Institute of Computing and Informatics, Technical University of Mombasa, Mombasa, Kenya

Abstract

Artificial intelligence is playing a great role in the production, classification, retrieval, and communication of knowledge. Tools like large language models, recommender systems, automated decision-making systems, and search engines do more than just process information; they shape whose testimony is heard, which concepts become available for interpreting experience, and which claims are treated as credible or important. This paper argues that Miranda Fricker's theory of epistemic injustice offers a strong framework to ethically identify shortcomings in AI-enabled knowledge production; however, it also suggests that, the distinction made in classical accounts between testimonial and hermeneutical injustices needs to be extended in order to address them in the context of sociotechnical systems. Testimonial injustice can arise when algorithmic models reproduce credibility deficits through biased training data, proxy variables, ranking practices, moderation systems, or automated classification. On the other hand, hermeneutical injustice can emerge when certain communities are poorly represented in the conceptual and linguistic tools that AI systems use to create explanations, categories, and interpretations. The paper further argues that the opaque nature of many AI systems intensifies these harms because affected persons may be unable to understand, contest, or correct the epistemic processes that affect them. Building on Fricker, Floridi, Dotson, Medina, and contemporary work in AI ethics, the paper develops an epistemic justice framework centred on adequacy in representation, credibility protection, interpretive inclusion, contestability, provenance, and institutional accountability. The proposed framework treats epistemic justice not merely as a desirable output property but as a design and governance requirement for AI systems that participate in consequential knowledge practices.

Keywords

Epistemic Injustice, Algorithmic Systems, Artificial Intelligence, Testimonial Injustice, Hermeneutical Injustice, AI Ethics

1. Introduction

Artificial intelligence (AI) has become key player in today's knowledge landscape. Search and recommendation systems determine which information we get to see; classifiers classify people, documents, images, and behaviours into categories; language models generate explanations and summaries; and

decision-support systems convert complex datasets into outputs that may influence employment, education, healthcare, finance, policing, welfare, and other areas of social life. These systems are frequently described in technical terms as prediction, classification, ranking, or generation systems [16]. Phil-

*Correspondence: Joel Sadiki Charo (sadikicharo7@gmail.com)

Received: 19 September 2026; **Accepted:** 4 October 2026; **Published:** 29 October 2026



Copyright: © The Author(s), 2026. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

osophically, however, they can also be understood as epistemic infrastructures: systems that influence what is knowable, who is represented as credible, which concepts are available for interpreting social experience, and how claims circulate between individuals and institutions [11].

This epistemic dimension raises a problem that cannot be adequately captured by conventional accounts of algorithmic bias alone [14]. Bias analysis often asks whether an AI system produces unequal outcomes across demographic groups. Such analysis is essential, but epistemic injustice concerns something broader: wrongs done to people specifically in their capacity as knowers. Miranda Fricker's influential account distinguishes testimonial injustice, which occurs when a speaker receives less credibility than they deserve because of identity prejudice, from hermeneutical injustice, which occurs when a social group lacks adequate shared interpretive resources for making sense of important experiences [5].

The central claim of this paper is that algorithmic systems can participate in epistemic injustice because they mediate credibility, representation, interpretation, and access to knowledge [9]. The systems do not need to possess human prejudice or consciousness for such injustice to occur [7]. They can inherit and operationalise patterns embedded in historical data, institutional practices, classification schemes, linguistic conventions, and platform incentives. The philosophical significance of this claim is that AI ethics should not be limited to questions of accuracy, privacy, safety, or distributive fairness. It must also ask whether AI-mediated knowledge practices treat persons and communities as legitimate sources of knowledge and whether the conceptual resources used by AI systems allow socially marginalised experiences to become intelligible [6].

The analysis proceeds in five stages. First, it reconstructs Fricker's concepts of testimonial and hermeneutical injustice and clarifies their relevance to digital knowledge production [6]. Second, it examines how recommender systems, large language models, and automated decision systems can generate epistemic harms. Third, it analyses the relationship between AI opacity and epistemic marginalisation. Fourth, it compares the Frickerian approach with information-ethics approaches associated with Luciano Floridi and with broader algorithmic fairness scholarship. Finally, it proposes a set of epistemic justice criteria for AI design and deployment [13]. The objective is not to provide another general literature review, but to develop an analytical framework capable of guiding the evaluation and governance of AI systems as knowledge-producing infrastructures.

2. The Epistemic Turn in AI Ethics

Much of the public discussion of AI ethics has concentrated on familiar concerns: privacy, discrimination, autonomy, accountability, transparency, safety, and security [13]. These concerns remain important, but they do not exhaust the ethical

significance of AI. AI systems increasingly occupy roles traditionally performed by human epistemic agents and institutions. A search engine can determine which sources a user sees first; a recommendation system can influence which political, cultural, or scientific material becomes salient; an automated screening system can decide which applications receive human attention; and a language model can produce an apparently authoritative explanation even when its underlying output is uncertain.

This development changes the conditions under which knowledge is produced and received. Epistemology is traditionally concerned with questions such as what counts as knowledge, how beliefs are justified, who can know, and what makes testimony reliable. In digital environments, these questions increasingly intersect with computation. A system that ranks sources is implicitly making epistemic distinctions. A system that generates a summary is deciding which information is salient. A system that labels a person or event is applying an interpretive category. A system that rejects a claim because it does not fit a learned pattern can affect whose testimony is recognised as credible.

Algorithmic mediation therefore introduces a layer between human knowers. This layer can be beneficial because computational systems can process information at scales impossible for individuals. Yet scale does not guarantee epistemic justice. As Floridi argues in the philosophy of information, information environments are ethically significant because informational entities and processes form part of the environment in which human agency takes place [4]. The epistemic perspective adds a more specific question: how do informational processes distribute the social conditions under which people can know and be known?

A Frickerian approach is particularly useful because it connects knowledge and power. Epistemic injustice is not simply an error in a database. It concerns a wrong committed against someone as a knower. This distinction matters for AI. A prediction may be statistically accurate while the categories used to generate it remain unjust. Conversely, a model may be imperfect without necessarily committing epistemic injustice. The analytical task is therefore to identify the social and epistemic relationships embedded in algorithmic systems rather than treating every error as an ethical wrong of the same kind.

3. Fricker's Framework: Testimonial and Hermeneutical Injustice

3.1. Testimonial Injustice

Fricker defines testimonial injustice around the credibility deficit that occurs when a hearer gives a speaker less credibility than the speaker deserves because of prejudice concerning social identity [5]. The classic structure involves a speaker, a hearer, testimony, and a credibility judgement. The injustice arises when identity prejudice interferes with the appropriate

assessment of testimony. Importantly, the harm is epistemic and ethical at the same time: the speaker is wronged in their capacity as a knower.

Applying this concept to AI requires care because an algorithm is not normally a hearer in the human sense. A recommender system or classifier does not possess human consciousness or necessarily make judgements in the phenomenological sense. Nevertheless, AI systems can operationalise credibility assignments [9]. A ranking model may systematically downrank information associated with particular communities. A moderation system may disproportionately identify particular forms of expression as suspicious or abusive. A language model may reproduce stereotypes when deciding which descriptions of a group are statistically likely. In each case, the computational system can function as part of a broader epistemic process in which some testimony is amplified, discounted, excluded, or rendered less intelligible.

This suggests that algorithmic testimonial injustice is often distributed rather than located in a single machine. The relevant epistemic actor may be a sociotechnical assemblage involving data collectors, developers, platform designers, institutions, users, and automated components. The machine does not have to be morally responsible in order for the system to produce a morally significant credibility deficit. This distinction allows Fricker's insight to be preserved without anthropomorphising algorithms.

3.2. Hermeneutical Injustice

Hermeneutical injustice occurs when a collective interpretive resource is structurally inadequate for understanding an important social experience, particularly where that inadequacy is connected to unequal participation in the production of shared meanings [5]. A community may experience something that is real and consequential but lack the socially available concepts required to articulate and communicate it. Hermeneutical injustice therefore concerns the conditions under which experiences become intelligible.

AI systems intensify this problem because contemporary models learn from existing linguistic and cultural materials. Large language models are trained on enormous corpora, while classification systems depend on categories created by institutions, researchers, and data annotators. If a community is poorly represented in the underlying data, its experiences may be poorly represented in the model's conceptual space. This can produce outputs that are grammatically fluent yet conceptually inadequate.

The problem is not simply that a model may fail to know a particular word. A deeper problem occurs when the categories used by the system repeatedly fail to capture a community's experience. For example, an automated system may force complex identities into binary categories, interpret culturally specific forms of expression through dominant linguistic conventions, or treat a socially contested experience as if the dominant category were neutral. In such circumstances, the model

may reproduce a hermeneutical gap because its representational resources reflect the unequal distribution of interpretive authority in society.

3.3. Beyond Individual Prejudice: Structural Power

Fricker's framework has generated further discussion about the structural dimensions of epistemic injustice. Dotson argues that epistemic injustice can occur through practices of epistemic exclusion [3], while Medina emphasises the dynamics of epistemic resistance and the role of social imagination [10]. These extensions are valuable for AI because algorithmic systems operate within institutions rather than in isolation. The relevant question is therefore not only whether an individual developer has prejudice, but also whether the system's institutional environment systematically excludes certain perspectives from data collection, model development, evaluation, deployment, and governance.

This structural interpretation is essential for understanding AI-driven knowledge production. A system can be developed by well-intentioned engineers and still reproduce epistemic injustice because the underlying data reflect unequal social structures. Similarly, a dataset can contain many examples of a marginalised group while still misrepresenting that group if the examples are selected through categories designed by outsiders. Representation is therefore not reducible to numerical inclusion. Epistemic justice requires attention to who defines the categories, whose interpretations are treated as authoritative, and whether affected communities can challenge the system's representations.

4. Algorithmic Systems as Epistemic Infrastructures

4.1. Recommender and Ranking Systems

Recommendation systems are epistemically significant because they structure attention. Users rarely encounter the total universe of available information. Instead, platforms rank and filter material according to learned patterns, engagement signals, user profiles, business objectives, and other criteria. The resulting information environment can influence which voices are visible and which claims receive repeated exposure.

From a Frickerian perspective, the problem is not that ranking is inherently unjust. Selection is unavoidable when information exceeds human attention. The ethical issue arises when selection systematically interacts with identity, power, or social marginalisation. If certain communities receive less visibility because their language, style, topics, or networks are poorly represented in the training data or because engagement-based objectives favour already dominant sources, their testimony can experience an algorithmic credibility deficit.

Conversely, highly engaging but unreliable content may receive greater visibility because the system optimises for attention rather than epistemic quality.

This illustrates an important distinction between epistemic justice and simple accuracy. A system can maximise prediction accuracy while producing an unjust distribution of epistemic visibility. Similarly, a recommendation algorithm can be individually relevant to users while collectively narrowing the range of perspectives that a public encounters. Epistemic evaluation must therefore consider not only whether individual predictions are correct but also how the system structures opportunities to speak, listen, contest, and learn.

4.2. Large Language Models

Large language models introduce a different but related epistemic problem. Their outputs are generated from statistical patterns learned from large datasets. They can produce explanations, summaries, classifications, translations, and answers in a style that resembles authoritative human communication. This fluency can create a misleading impression that the system possesses a unified and reliable body of knowledge. In reality, generated text can contain factual errors, omissions, stereotypes, and culturally narrow interpretations.

The epistemic risk becomes particularly serious when users cannot easily distinguish between a model's linguistic fluency and the quality of its evidential basis. Bender et al. emphasise that large language models can reproduce problematic patterns from their training data and that scale does not automatically solve the underlying social and representational problems [1]. From an epistemic justice perspective, the concern is that marginalised voices may be reproduced as stereotypes, omitted entirely, or transformed into dominant-language interpretations that erase context.

There is also a problem of epistemic provenance. When a language model generates a statement, users may not know which sources, communities, or linguistic practices contributed to the output. The opacity of provenance makes it difficult for affected people to contest representations. A person may recognise that an AI-generated description of their community is wrong but be unable to determine why the system produced it or how to correct the underlying representation. Thus, opacity can turn a representational error into a continuing epistemic disadvantage.

4.3. Automated Decision Systems

Automated decision systems provide perhaps the clearest connection between epistemic and distributive harms. Systems used in recruitment, credit, education, welfare, healthcare, or other high-impact settings transform information about individuals into classifications or recommendations. The output may then influence how institutions understand and treat the person concerned.

Here, testimonial injustice can occur when a person's own

account is systematically discounted relative to machine-generated indicators. For example, an automated risk score may be treated as more authoritative than an individual's explanation, even when the person has relevant contextual information that the system cannot observe. The problem is not that algorithmic evidence should never be used. Rather, epistemic justice requires institutions to avoid treating computational outputs as epistemically infallible and to preserve meaningful channels through which affected persons can provide evidence and contest interpretations.

This point also exposes a danger in the language of objectivity. Algorithmic outputs may appear objective because they are numerical, repeatable, and produced by software. Yet the categories, variables, labels, objectives, thresholds, and institutional uses of the system are human-designed. Selbst et al. show why fairness problems cannot be separated from the social context in which models operate [8]. A technically fair metric applied to an institutionally unjust setting may fail to address the actual source of harm.

5. The Black-Box Problem and Epistemic Marginalisation

The opacity of algorithmic systems is often described as a transparency problem. From an epistemic justice perspective, opacity is more specifically a problem of unequal access to reasons. A person affected by an automated decision may be unable to determine what information was used, what assumptions shaped the output, or which pathway could lead to correction. The resulting asymmetry is epistemically significant because the institution possesses information about the person while the person lacks equivalent access to the institution's decision process.

Burrell identifies several forms of machine-learning opacity, including secrecy, technical illiteracy, and the intrinsic complexity of some models [2]. These forms should not be treated as identical. A system can be technically complex without being socially secretive, while a simple model can be deliberately hidden. What matters ethically is whether opacity prevents meaningful understanding and contestation by those affected.

The black-box problem can therefore create a secondary layer of epistemic injustice. First, the system may misrepresent a person or community. Second, the opacity of the system may prevent the affected party from demonstrating that the representation is wrong. This creates a feedback loop: the system generates an epistemic judgement; the judgement influences institutional treatment; the person attempts to challenge it; but the lack of access to relevant information weakens the person's ability to contest the judgement.

A response based solely on technical explainability may not be sufficient. An explanation that reveals a mathematical feature importance score may still be unintelligible or irrelevant to the affected person. Epistemic justice requires what may be

called contestable intelligibility: explanations should provide affected persons with enough meaningful information to understand the basis of a decision, identify potentially incorrect assumptions, provide relevant counter-evidence, and obtain review by an appropriate human or institutional process.

This approach also reframes the purpose of transparency. Transparency is often defended as a means of increasing trust. Yet trust should not be the primary goal when a system exercises power. The more fundamental goal is accountability and the preservation of epistemic agency.

6. Frickerian Epistemic Injustice and Floridian Information Ethics

Floridi's philosophy of information provides a complementary framework for understanding the ethical significance of digital environments. His information ethics treats the informational dimension of reality as morally significant and emphasises the preservation and flourishing of informational entities and processes [4]. This perspective is useful because AI systems operate within information environments rather than as isolated tools.

The Frickerian and Floridian approaches, however, ask different primary questions. A Floridian analysis directs attention to the moral status of informational entities, the quality of information environments, and the consequences of information processes. A Frickerian analysis focuses more directly on persons as knowers and on the power relations that shape credibility and interpretation. The two approaches can therefore be combined without collapsing one into the other.

The combination yields a broader ethical model. First, information environments should not systematically damage persons or communities in their epistemic capacities. Second, informational processes should preserve the conditions required for meaningful participation in knowledge practices. Third, AI systems should be evaluated in relation to the social structures within which they operate. This moves the analysis beyond an abstract concern with information quality toward the social justice of knowledge production.

Traditional information-ethics frameworks can sometimes be criticised for treating information primarily as an object of ethical analysis. A Frickerian contribution is to insist that information is also embedded in relations among knowers. The ethical question is therefore not simply whether a dataset is accurate or whether an information object has been properly processed. It is also whether the system's epistemic practices recognise people as legitimate participants in the production and correction of knowledge.

7. From Algorithmic Bias to Epistemic Justice

Algorithmic bias and epistemic injustice overlap, but they should not be treated as synonyms. Algorithmic bias usually

describes systematic deviation, unequal performance, or patterned disadvantage in an algorithmic process. Epistemic injustice identifies a particular kind of wrong: a failure that affects people as knowers or that undermines the interpretive resources through which experience becomes intelligible.

This distinction has practical implications. A fairness audit may demonstrate that a model has different error rates across demographic groups, but it may not reveal whether the model's categories themselves are epistemically inadequate. Conversely, a system may have similar predictive performance across groups while still imposing categories that erase important differences in lived experience. An epistemic justice audit therefore needs to ask questions that conventional performance evaluation may overlook.

The following analytical questions illustrate the difference:

- 1) Who is represented in the system's training and evaluation data?
- 2) Who defines the categories and labels?
- 3) Which forms of testimony are treated as reliable evidence?
- 4) Which experiences are difficult or impossible to express within the system's conceptual vocabulary?
- 5) Can affected persons challenge the system's representation of them?
- 6) Is there a meaningful route for correction?
- 7) Does the system's deployment context amplify or reduce existing power inequalities?

These questions are compatible with established AI fairness research but add an explicitly epistemic dimension. They also support the argument that fairness cannot be reduced to a single metric. Mittelstadt et al. emphasise the ethical complexity of algorithmic decision-making [12], while Selbst et al. demonstrate the limitations of treating fairness as a purely technical property [15]. Epistemic justice extends this contextual approach by making the status of persons as knowers an explicit object of evaluation.

8. Proposed Epistemic Justice Criteria for AI Systems

This paper proposes six criteria for evaluating AI systems that participate in knowledge production. The criteria are intended as an analytical and design framework rather than a universal numerical scoring system. Their purpose is to identify the conditions under which AI-mediated knowledge practices can respect epistemic agency.

8.1. Representational Adequacy

AI systems should represent relevant populations, experiences, languages, and contexts with sufficient conceptual adequacy for the intended use. Numerical representation is not enough. Developers should examine whether categories and labels capture the distinctions that matter to affected communities and whether important experiences disappear through

preprocessing, annotation, or model design.

8.2. Credibility Protection

Systems that rank, classify, filter, or summarise information should be assessed for systematic credibility deficits. Developers should examine whether particular communities or linguistic forms are disproportionately downranked, rejected, misclassified, or associated with negative attributes. The goal is not to ensure that every claim is accepted, but to ensure that credibility judgements are not distorted by identity-based or structurally inherited prejudice.

8.3. Interpretive Inclusion

AI systems should provide adequate interpretive resources for the experiences relevant to their deployment. This requires attention to minority languages, culturally specific meanings, contested categories, and community-generated concepts. Where no single interpretation is appropriate, systems should signal uncertainty or plurality rather than presenting a dominant interpretation as universally valid.

8.4. Epistemic Provenance

Users should have meaningful information about the sources, data categories, model limitations, and institutional processes that shape important AI outputs. Provenance is especially important for generated content. Where feasible, systems should distinguish between retrieved evidence, model inference, uncertainty, and unsupported generation. This can reduce the risk that linguistic fluency is mistaken for evidential authority.

8.5. Contestability and Correction

Affected persons should have practical opportunities to contest consequential AI outputs and provide relevant evidence. Contestability requires more than a generic feedback button. It should include a route to human review, an explanation appropriate to the person's circumstances, and a mechanism through which credible corrections can influence future system performance or institutional practice.

8.6. Participatory Accountability

Communities affected by AI systems should have meaningful opportunities to participate in system definition, evaluation, deployment, and review. Participation should occur before deployment rather than being limited to post-hoc complaints. This criterion follows from the structural nature of hermeneutical injustice: if interpretive resources are socially produced, affected communities should not be excluded from decisions about the categories through which they are represented.

9. A Proposed Epistemic Justice Lifecycle

The six criteria can be integrated into an AI governance lifecycle. At the problem-definition stage, organisations should identify whose knowledge the system is intended to mediate and which communities may be affected. During data design, teams should evaluate representation, provenance, labelling practices, and missing perspectives. During model development, evaluation should include group-specific epistemic tests, including whether the system systematically discounts particular forms of testimony or produces stereotyped interpretations.

Before deployment, organisations should conduct an epistemic impact assessment. Such an assessment should document the system's intended epistemic role, affected communities, potential credibility deficits, conceptual limitations, explanation mechanisms, and correction pathways. During deployment, monitoring should examine real-world feedback and patterns of contestation. After significant incidents, the organisation should review not only model performance but also whether affected people were able to understand and challenge the system's outputs.

This lifecycle approach changes the governance question from 'Is the model fair?' to a more comprehensive set of questions: fair to whom, in what epistemic role, according to whose categories, with what evidence, and with what capacity for correction? It also recognises that epistemic justice cannot be guaranteed entirely at the model level. Institutional procedures, user interfaces, data governance, procurement practices, and legal accountability all contribute to the epistemic properties of the deployed system.

10. Challenges and Objections

10.1. Can Machines Really Commit Epistemic Injustice

A first objection is that Fricker's theory concerns interpersonal epistemic relations and therefore cannot straightforwardly apply to machines. This objection is partly correct. AI systems should not be described as if they possessed human-like intentions, prejudice, or moral consciousness. The more defensible claim is that AI systems can participate in epistemic injustice as components of sociotechnical practices. The injustice belongs to the system of practices and institutions, not necessarily to the machine as a moral subject.

10.2. Is Epistemic Justice Simply Another Name for Fairness

A second objection is conceptual redundancy. If algorithmic fairness already addresses discrimination, why introduce epistemic injustice? The answer is that fairness metrics and

epistemic justice illuminate different dimensions. Fairness can evaluate outcome distributions or error rates; epistemic justice asks whether persons and communities are recognised as knowers and whether their experiences are intelligible within the system. The concepts overlap, but neither fully subsumes the other.

10.3. Could Participatory Design Create Conflicting Demands

A third challenge concerns disagreement. Communities are not internally homogeneous, and different groups may demand incompatible categories or representations. Participatory epistemic governance cannot therefore guarantee a single correct perspective. Its purpose is instead to make disagreement visible, provide mechanisms for negotiation, and prevent powerful institutions from treating their own categories as neutral and final. Systems should sometimes represent uncertainty and contestation rather than manufacture false consensus.

10.4. Does Transparency Solve the Problem

A final objection is that greater transparency could address epistemic injustice. Transparency is valuable, but disclosure alone is insufficient. Technical documentation may be inaccessible to affected persons, and source-code access may not reveal how institutional practices shape outcomes. Epistemic justice requires usable explanation, contestability, participation, and correction. Transparency is therefore best understood as one component of a broader accountability architecture.

10.5. Implications for AI Governance and Professional Practice

The framework has implications for developers, institutions, regulators, and users. For developers, epistemic justice means treating representation and interpretive adequacy as engineering requirements rather than optional ethical additions. Dataset documentation should include information about representation and exclusions. Evaluation should test not only accuracy but also systematic credibility and interpretive failures. Product teams should identify which users will be able to challenge outputs and which users may lack the resources to do so.

For institutions, the framework implies that procurement should include epistemic requirements. Organisations adopting AI should ask vendors how the system was trained, what groups were represented, what limitations are known, how users can contest outputs, and how corrections are incorporated. Human oversight should be meaningful rather than merely ceremonial. A human reviewer who is required to accept a machine output without access to relevant evidence does not provide effective epistemic accountability.

For regulators, epistemic justice supports requirements for

transparency, documentation, human oversight, impact assessment, and redress in high-impact AI applications. Contemporary regulatory developments, including risk-based approaches to AI governance, increasingly recognise that some AI applications require stronger safeguards than low-risk applications. An epistemic justice perspective complements such approaches by focusing specifically on the rights and capacities of people as participants in knowledge practices.

For professional education, the framework suggests that computing and information-technology programmes should teach epistemology alongside technical AI skills. Developers who understand model architecture but lack awareness of testimony, interpretation, social power, and epistemic exclusion may be technically competent while remaining unable to recognise important forms of harm. Philosophy of information can therefore make a practical contribution to responsible computing education.

11. Discussion

The analysis supports a central philosophical conclusion: AI systems should be understood not only as computational instruments but also as epistemic infrastructures. They influence who becomes visible, which claims circulate, how categories are applied, and how institutional knowledge is produced. This makes epistemic justice a distinct dimension of responsible AI.

The Frickerian framework is especially useful because it reveals two different mechanisms of harm. Testimonial injustice concerns credibility: whose claims are believed, amplified, or discounted. Hermeneutical injustice concerns intelligibility: whether people possess or can access the interpretive resources necessary to understand and communicate their experiences. AI systems can affect both. A recommender system may reduce the visibility of certain testimony, while a language model may reproduce dominant conceptual categories that make minority experiences difficult to express. The two mechanisms can also reinforce each other. When a community is poorly represented conceptually, its testimony may appear less credible because the system lacks the categories needed to interpret it.

The black-box problem adds another dimension. If an affected person cannot understand or contest an AI-generated judgement, the original epistemic injustice can become institutionally entrenched. This suggests that epistemic justice requires more than bias reduction. It requires the preservation of epistemic agency: the practical ability of people to contribute knowledge, challenge representations, obtain reasons, and correct errors about themselves and their communities.

The proposed criteria therefore shift the focus from model outputs alone to the entire sociotechnical environment. Representational adequacy, credibility protection, interpretive inclusion, epistemic provenance, contestability, and participatory accountability should be treated as interconnected. A system that performs well statistically but cannot be contested

may still fail epistemically. A system with strong transparency but poor representational adequacy may also fail. Epistemic justice is consequently a relational and institutional property rather than a single technical metric.

12. Conclusion

Artificial intelligence is becoming an active participant in the production and circulation of knowledge. This development creates opportunities for wider access to information and for new forms of human-machine collaboration, but it also introduces new ways in which people can be disadvantaged as knowers. Miranda Fricker's concepts of testimonial and hermeneutical injustice provide a valuable starting point for understanding these risks.

The paper has argued that algorithmic systems can participate in testimonial injustice when they reproduce or operationalise credibility deficits, and in hermeneutical injustice when their representational and interpretive resources systematically fail to capture the experiences of marginalised communities. These harms are intensified when AI systems are opaque and when affected persons lack meaningful opportunities to contest and correct algorithmic representations.

The proposed epistemic justice framework identifies six requirements: representational adequacy, credibility protection, interpretive inclusion, epistemic provenance, contestability and correction, and participatory accountability. Together, these criteria provide a philosophical basis for evaluating AI systems as sociotechnical knowledge infrastructures. They also complement existing approaches to fairness, transparency, information ethics, and responsible AI.

The broader implication is that responsible AI should not ask only whether a system predicts accurately or distributes outcomes fairly. It should also ask whether the system respects people as knowers. AI governance that protects epistemic agency must ensure that communities are represented adequately, that testimony is not systematically discounted, that interpretive resources are inclusive, and that affected persons can challenge institutionalised machine judgements. The ethical design of AI therefore requires not only better algorithms but better epistemic institutions.

Abbreviations

AI	Artificial Intelligence
ACM	Association for Computing Machinery
AAAI	Association for the Advancement of Artificial Intelligence
LL	Large Language Model
ML	Machine Learning and Research
AI	Artificial Intelligence
ACM	Association for Computing Machinery
AAAI	Association for the Advancement of Artificial Intelligence

LL	Large Language Model
ML	Machine Learning and Research

Author Contributions

Joel Sadiki Charo: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Methodology, Project administration

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623. <https://doi.org/10.1145/3442188.3445922>
- [2] Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1). <https://doi.org/10.1177/2053951715622512>
- [3] Dotson, K. (2011). Tracking epistemic violence, tracking practices of silencing. *Hypatia*, 26(2), 236-257.
- [4] Floridi, L. (2013). *The ethics of information*. Oxford University Press.
- [5] Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.
- [6] Furst, M. (2024, January 1). Closing the Conceptual Gap in Epistemic Injustice. *The Philosophical Quarterly*.
- [7] Hull, G. (2023, June 15). Dirty Data Labeled Dirt Cheap: Epistemic Injustice in Machine Learning Systems.
- [8] Kalluri, P. (2020). Don't ask if AI is good or fair, ask how it shifts power. *Nature*, 583, 169. <https://doi.org/10.1038/d41586-020-02003-2>
- [9] Kay, J. K. (2024). Epistemic Injustice in Generative AI. *Proceedings of the AAAI ACM Conference on AI, Ethics, and Society*, 7(1), 684-697.
- [10] Medina, J. (2013). *The epistemology of resistance: Gender and racial oppression, epistemic injustice, and resistant imaginations*. Oxford University Press.
- [11] Miedema, E. (2025). Towards trustworthy artificial intelligence for decision-making: A lifecycle perspective on knowledge- and data-driven artificial intelligence systems.
- [12] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2). <https://doi.org/10.1177/2053951716679679>

- [13] Mollema, W. J. (2025). A taxonomy of epistemic injustice in the context of AI and the case for generative hermeneutical erasure.
- [14] Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.
- [15] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59-68.
<https://doi.org/10.1145/3287560.3287598>
- [16] Xu, Z. (2023). A Reusable Model-agnostic Framework for Faithfully Explainable Recommendation and System Scrutability.