Epistemic Algorithmic Drift: AI-Mediated Scholarly Infrastructures and the Reproduction of Gendered and Intersectional Inequalities

Paola Panarese
Sapienza University of Rome (Italy)
paola.panarese@uniroma1.it

DOI: https://doi.org/10.13165/WinS-26-01

Abstract: Artificial intelligence is reshaping scientific knowledge production in ways that risk reinforcing existing inequalities. This article introduces the concept of *epistemic algorithmic drift* to explain how AI-mediated scholarly infrastructures, trained on historically stratified data, generate cumulative and often imperceptible redistributions of epistemic visibility, amplifying gendered, linguistic, and geographic disparities.
The contribution is theory-building: drawing on feminist epistemology, intersectionality theory, and critical algorithm studies, it conceptualizes drift as a self-reinforcing feedback loop between structured data and AI-mediated processes of discovery, evaluation, and recognition. In doing so, it distinguishes epistemic algorithmic drift from related notions such as algorithmic bias, concept drift, and path dependency.
The article specifies empirically observable indicators, examines the process through an intersectionally informed perspective, and derives governance considerations grounded in its structural dynamics. It argues that addressing these inequalities requires longitudinal, intersectional, and infrastructural interventions rather than isolated technical fixes.

Keywords: Artificial intelligence; algorithmic bias; gender inequality; scientific publishing; epistemic infrastructures; intersectionality; cumulative advantage; feminist epistemology.

Introduction

In recent years, artificial intelligence (AI) has shifted from the margins to the core of scientific practice, evolving from a set of auxiliary tools into a more deeply embedded component of knowledge production (Amayreh & Amayreh, 2025; Eren & Perez, 2025). This transition signals a broader transformation: not simply the adoption of new technologies, but their integration into the very processes through which scientific value is assigned and recognition distributed.

Central to this shift is the growing role of machine learning systems in mediating key decision-making processes within academic publishing and research evaluation. These systems shape which papers are encountered in literature searches, how manuscripts are screened and routed through peer review, and how research proposals are assessed and ranked (Beel et al., 2016; Heaven, 2018; Kousha & Thelwall, 2024; Lund et al., 2024). In doing so, AI does

not merely support scientific activity but actively shapes its trajectories. Yet the functioning and criteria of these systems remain only partially transparent, even to the institutions that rely on them.

From a feminist perspective on knowledge production, this reconfiguration concerns the conditions under which epistemic authority is produced and distributed. Within this framework, Sandra Harding (1986, 1991) reconceptualizes objectivity as situated, arguing that "strong objectivity" requires explicit recognition of the social position of the knower. Building on this shift, Helen Longino (1990, 2002) locates scientific validity in structured processes of critical interaction rather than individual methodological rigor alone. Extending this critique, Donna Haraway (1988) challenges the illusion of a neutral "view from nowhere," showing how claims to objectivity obscure their own partiality, while Lorraine Code (1991) emphasizes that epistemic authority circulates within socially organized practices of credibility. Taken together, these contributions establish that AI cannot be understood as a neutral enhancement of scientific practice: it functions as a socio-technical mediator embedded in existing structures of power.

This argument gains further force when situated within Kimberlé Crenshaw's (1989, 1991) concept of *intersectionality*, which shows how categories such as gender, race, and class operate as interlocking systems with irreducible effects, and Patricia Hill Collins's (1990) notion of a *matrix of domination*, which highlights the mutually reinforcing nature of institutional and discursive power. Applied to AI-mediated systems, this perspective suggests that their effects are neither neutral nor uniform: the organization of visibility, evaluation, and recognition is shaped by linguistic, geographic, and positional inequalities, which are both reflected and reinforced through algorithmic processes.

Against this background, mechanisms of cumulative advantage in science acquire renewed relevance, as Robert K. Merton (1968) showed in his formulation of the Matthew effect, whereby recognition tends to concentrate around already established scholars. Within this framework, Margaret W. Rossiter (1993) identified its gendered counterpart, the Matilda effect, naming the systematic undervaluation and misattribution of women's scientific contributions. These dynamics are not merely theoretical but are consistently supported by empirical evidence: large-scale bibliometric analyses show that publications authored by women receive fewer citations than comparable work by men across most scientific fields (Larivière et al., 2013); a landmark peer review study found that female applicants required significantly stronger publication records to receive equivalent evaluation scores (Wennerås & Wold, 1997); and more recent work confirms that disparities accumulate over time, shaping trajectories of visibility and contributing to the persistent underrepresentation of women among the most highly cited scholars (Huang et al., 2020; King et al., 2017).

Thus, the growing integration of AI-mediated systems reshapes the conditions under which inequality is reproduced. Systems trained on historically stratified data risk encoding and amplifying existing disparities, while their opacity makes these processes less visible and more difficult to

contest — a shift from socially mediated to infrastructurally embedded forms of cumulative disadvantage.

To account for this shift, the concept of epistemic algorithmic drift is introduced, capturing the process through which pre-existing inequalities, once mediated by AI-driven scholarly infrastructures, are translated into cumulative and often imperceptible redistributions of visibility, recognition, and epistemic authority. The aim is to move from naming an emerging phenomenon to specifying its generative mechanism, to distinguish it from adjacent notions, and to identify empirically observable indicators through which it can be analyzed.

This analytical orientation defines the nature of the contribution: the article develops a conceptual framework grounded in a theoretically informed narrative review of existing research, with the aim of specifying a generative mechanism and deriving analytically grounded governance implications.

On this basis, the argument unfolds across four interconnected movements. The first examines the infrastructural role of AI in scholarly communication, focusing on the forms of epistemic filtering through which visibility and relevance are algorithmically organized. The second considers the unequal conditions under which AI integration unfolds across the research community. The third develops the concept of epistemic algorithmic drift itself, outlining its analytical dimensions, intersectional implications, and empirically observable indicators. The fourth derives governance considerations from the structural analysis, distinguishing between responses adequate to drift and those that address only its symptoms.

Methodological Note and Analytical Positioning

The analysis combines conceptual synthesis with a theoretically informed narrative review of recent empirical research on artificial intelligence in academic work and scientific publishing. Rather than aiming at exhaustive coverage, the review seeks to identify recurring mechanisms through which AI-mediated systems interact with existing inequalities in knowledge production, with a specific focus on gender as a relational and intersectional dimension.

The literature search was conducted between November 2025 and January 2026 using Scopus, Web of Science, and Google Scholar, combining terms related to artificial intelligence, scholarly communication, and inequality in Boolean configurations[1]. In addition, forward citation tracking was conducted on key contributions in feminist epistemology, intersectionality theory, and critical algorithm studies. The corpus was supplemented by foundational studies on gender disparities in peer review and citation practices, and by

[1] The Boolean search string was constructed as follows: (“artificial intelligence” OR “machine learning” OR “generative AI”) AND (“academic publishing” OR “scholarly communication” OR “peer review” OR “research evaluation”) AND (“gender” OR “women” OR “gender inequality” OR “gender disparities” OR “intersectionality” OR “algorithmic bias”).

selected theoretical works necessary to situate the proposed concept within adjacent sociological and STS frameworks.

The initial search yielded approximately 160 records across the three databases. After duplicate removal, the remaining records were screened in two stages (title/abstract screening followed by full-text assessment). The final corpus includes 32 sources: peer-reviewed journal articles, books and book chapters, and a limited number of reports or preprints addressing recent developments not yet consolidated in the peer-reviewed literature.

The distribution of results varied across platforms. Scopus and Web of Science returned the majority of peer-reviewed articles in social sciences and science and technology studies. Google Scholar contributed additional materials, including book chapters, interdisciplinary contributions, and recent preprints. Purely technical papers focused exclusively on model performance, as well as opinion pieces lacking explicit analytical grounding, were excluded unless directly relevant to AI-mediated scholarly practices.

Selection was guided by three criteria: (1) relevance to the role of AI in mediating visibility, evaluation, and participation in knowledge production; (2) sufficient methodological transparency, understood as the explicit description of data sources, research design, or analytical procedures allowing assessment of evidential robustness; and (3) coverage across disciplinary perspectives, including social sciences, STS, and computer science. Studies were excluded if they addressed AI in contexts unrelated to academic knowledge production, lacked minimal methodological specification, or did not provide implications for scholarly practice or epistemic inequality.

The review adopts a comparative interpretive approach rather than quantitative aggregation. Several limitations are acknowledged. Industry sources are used cautiously, as indicative of broader distributional patterns rather than as primary evidence, and preprints are included selectively to capture emerging developments while remaining analytically qualified throughout. The uneven availability of gender-disaggregated empirical data on AI in academic contexts limits the precision with which intersectional dynamics can be reconstructed. This constraint is not only empirical but analytically relevant, as it contributes to the limited observability of the processes under investigation.

Accordingly, the article adopts intersectionality as an analytical orientation rather than as a fully operationalized empirical design: while gender, geography, language, career stage, and institutional position can be specified as relevant dimensions of exposure to drift, the available literature does not yet allow systematic analysis of their intersections with racialized forms of inequality across scholarly infrastructures.

A further limitation concerns the epistemological status of the concept introduced here. Epistemic algorithmic drift is proposed as a theoretical construct supported indirectly by evidence on its constituent mechanisms (historically stratified data, algorithmic amplification of visibility), and unequal conditions of AI adoption, rather than directly demonstrated as a fully observable process. Establishing it empirically would require longitudinal,

disaggregated data linking AI system deployment to changes in visibility and evaluation patterns within specific scholarly infrastructures. While partial datasets exist across bibliometric archives and editorial systems, they remain fragmented, unevenly accessible, and not yet available in integrated forms that would allow robust, intersectionally disaggregated analysis of these dynamics. This paper therefore contributes by specifying the mechanism in analytically explicit terms, enabling its future empirical investigation.

Epistemic Infrastructures and Algorithmic Mediation

AI systems increasingly function as epistemic infrastructures within scholarly communication, shaping not only what becomes accessible but also the conditions under which it is encountered. As Susan Leigh Star (1999) observed, infrastructures are not neutral supports; they are relational arrangements that embed standards, classifications, and conventions. In doing so, they actively configure what can be known, even as they remain largely taken for granted. Indeed, their relative invisibility is not incidental but constitutive of their power, allowing them to organize epistemic processes while remaining only partially accessible to scrutiny.

Within this infrastructural configuration, the growing reliance on measurable proxies, such as citation counts, publication frequency, institutional prestige, and linguistic fluency, follows less from their capacity to capture epistemic quality than from their compatibility with computational optimization. In other words, what is most easily formalized becomes what is most readily valued. As D'Ignazio and Klein (2020) argue, data systems embed particular worldviews while presenting themselves as neutral and universal, since decisions about what is counted, how it is categorized, and what remains invisible are incorporated into technical architectures. Consequently, the organization of knowledge is not simply mediated by AI systems, but actively structured through the categories and metrics they operationalize.

A first domain in which these infrastructural effects become empirically observable is literature discovery. Here, AI-assisted review tools do not merely accelerate existing practices; rather, they reorganize them by redefining how relevance is identified, how topics are clustered, and, crucially, which sources become visible at all. Because recommender systems are typically trained on citation networks, they tend to privilege already highly cited publications, thereby reinforcing cumulative advantage dynamics (Beel et al., 2016). The resulting effect is a form of epistemic pre-filtering that shapes the field of possible knowledge in advance of any explicit evaluative judgment.

At this point, however, an important qualification is required. Research on network homophily shows that even in the absence of algorithmic intervention, structurally unbalanced networks tend to concentrate visibility around majority-group members. Karimi et al. (2018), for instance, demonstrate that when minority groups are underrepresented, even individually unbiased linking behaviour can produce systematically unequal outcomes. Applied to citation networks, this implies that AI systems inherit not only explicit biases but also the structural effects of homophily embedded

in the data. For this reason, it becomes analytically necessary to distinguish between concentration that would emerge from network structure alone and concentration that is further amplified through recursive AI-mediated feedback. Epistemic algorithmic drift designates precisely this latter process: the additional redistribution of visibility produced through the iterative interaction between system outputs and scholarly practices.

These dynamics resonate with a broader body of research indicating that automated classification and recommendation systems tend to reproduce existing social hierarchies under the guise of efficiency and objectivity (Benjamin, 2019; Crawford, 2021; Noble, 2018). At the same time, longstanding debates in science and technology studies caution against treating technological effects as inevitable or detached from institutional contestation and user agency (Pinch & Bijker, 1984; Winner, 1980). In this perspective, epistemic algorithmic drift is not conceived as an inevitable outcome, but as a structural tendency whose effects depend on how systems are designed, implemented, and governed.

Importantly, these mechanisms are not confined to literature discovery but extend to editorial infrastructures more broadly. Systems for reviewer identification and manuscript routing, for example, are increasingly built on historical publication records, reviewer activity, and editorial assignment data (Beel et al., 2016; Elsevier, 2024; Zhao & Zhang, 2022). Because these data are generated within academic environments long documented as structurally unequal, their use raises concerns about the reproduction of existing disparities within AI-assisted editorial decision-making. In parallel, the growing use of AI-assisted writing introduces unresolved questions concerning authorship and provenance, while the increasing entanglement of human and machine judgment complicates the evaluative boundaries of peer review (Hosseini & Horbach, 2023).

The extent to which these infrastructural effects can be empirically observed, however, varies across disciplinary contexts. In fields such as biomedical research, computer science, and physics, AI-assisted tools are already embedded in the operational workflows of major publishers and preprint platforms, making their distributional consequences, at least in principle, traceable. By contrast, in the social sciences and humanities, AI integration into editorial processes remains more uneven and less systematically documented, while bibliometric data disaggregated by gender and institutional position are less consistently available. As a result, disciplinary unevenness shapes not only the exposure to epistemic algorithmic drift but also the feasibility of monitoring it: fields in which AI integration is most advanced are simultaneously those most exposed to drift and those best positioned to generate the data needed to detect it.

Overall, these developments point to a structural transformation rather than a set of isolated distortions. Epistemic gatekeeping is progressively delegated to systems that operate outside the normative frameworks through which scientific communities have traditionally regulated evaluation and recognition. To be sure, algorithmic mediation can also introduce corrective effects, for instance, automated double-blind review may reduce reputational

bias, while AI-assisted translation can expand access across linguistic boundaries. Yet, in the absence of explicit design choices and accountability mechanisms, these benefits remain unevenly distributed. Ultimately, the same systems that can mitigate inequality under certain conditions can also intensify it under others, depending on how they are embedded, governed, and evaluated.

Unequal Conditions of AI Integration

The mechanisms of epistemic filtering described above do not affect all researchers in the same way. Their effects depend on who can engage with these systems and under what conditions. AI integration into academic work therefore unfolds within pre-existing structures of inequality that shape both access to these tools and their consequences for epistemic participation, at the intersection of gender, career stage, institutional position, language, and geography (Collins, 1990; Crenshaw, 1989, 1991).

A first axis of differentiation concerns the organization of academic labour. Time, as a key resource for research, is unevenly distributed: women researchers tend to carry disproportionate teaching, administrative, and mentoring responsibilities, which fragment the sustained time required for both intellectual work and experimentation with new technologies (Guarino & Borden, 2017; Schiebinger, 1999). The capacity to integrate AI into research practices thus depends on the availability of time, stability, and institutional support; in this sense, the relationship between gender and technology is mediated by the organization of labour (Wajcman, 2004).

Empirical evidence shows that engagement with AI tools varies systematically within this structured landscape. Dorta-González et al. (2024), in a large-scale cross-disciplinary and cross-national study, find that women researchers report higher levels of concern regarding the reliability, transparency, and ethical implications of generative AI, alongside more cautious patterns of use. This orientation can be understood as shaped by structural position: researchers whose work is subject to heightened scrutiny face stronger incentives to avoid practices whose legitimacy is not yet stabilized, consistent with evidence on gendered evaluation asymmetries in science (Wennerås & Wold, 1997). Over time, such caution may affect familiarity with emerging research practices and competitiveness in contexts where AI tools are increasingly integrated (Lund et al., 2024; Nicholas et al., 2024).

These dynamics are particularly pronounced among early-career researchers, for whom AI represents both an opportunity and a source of uncertainty. Nicholas et al. (2024) show that junior scholars use AI to accelerate aspects of research production while expressing concerns about authorship, originality, and evaluative legitimacy. When career precarity intersects with gender, these tensions may become more pronounced, limiting the capacity to absorb the risks associated with AI adoption.

Global and linguistic asymmetries further compound these dynamics. Women remain underrepresented in the global research workforce, with

significant regional disparities (UNESCO Institute for Statistics, 2024), while the large language models that underpin most AI tools are predominantly trained on English-language data, contributing to the centrality of standardized academic English in scholarly communication (Raitskaya & Tikhonova, 2025). Researchers working in other languages must navigate tools optimized for different linguistic environments while competing in systems that privilege linguistic conformity as a proxy for quality.

These dynamics interact with disciplinary structures in ways that compound rather than simply add their effects. In fields with strong international standardization — such as experimental biology or quantitative economics — the pressure toward linguistic and methodological conformity is already institutionally embedded, making AI-assisted standardization an accelerant of pre-existing tendencies. In fields characterized by greater methodological pluralism and stronger traditions of non-English scholarly production — including sociology, history, and area studies — the same tools may function as instruments of disciplinary boundary-drawing, privileging outputs that conform to dominant Anglophone conventions while rendering heterodox or regionally specific contributions less visible. The mechanism of drift is the same across these contexts; its rate and its consequences for epistemic diversity are not.

Considered jointly, these dimensions point to a differentiated landscape of AI integration in which constraints interact rather than simply accumulate. While the empirical literature has not yet fully captured these configurations, each of their constituent dynamics is well documented, and their interaction helps explain how these processes remain closely aligned with the inequalities that structure academic work more broadly.

Epistemic Algorithmic Drift: Concept, Mechanism, and Indicators

Against this background, the concept of *epistemic algorithmic drift* can be specified as a recursive process through which AI-mediated scholarly interactions progressively reorganize the distribution of visibility, recognition, and epistemic authority. Rather than referring to a single distortion or identifiable bias, the concept captures a dynamic: the gradual reorganization of epistemic opportunity that unfolds as AI-mediated outputs are integrated into scholarly practices and recursively incorporated into the infrastructures that shape subsequent outputs.

At its core, this process can be understood as a self-reinforcing feedback loop operating across four stages. First, AI systems are trained on scholarly data — citation networks, publication records, reviewer activity — that encode historically uneven distributions of recognition, including well-documented gender disparities (Larivière et al., 2013; Wennerås & Wold, 1997). Second, these systems operationalize such distributions through proxies of relevance, visibility, and expertise: highly cited work is more likely to be recommended, established publication patterns function as benchmarks, familiar reviewer profiles are more likely to be selected (Beel et al., 2016; Kousha & Thelwall,

2024). Third, researchers interact with these outputs, generating new data traces. Fourth, as these traces are incorporated into subsequent system updates, later outputs tend to reinforce earlier patterns. Over successive iterations, relatively small asymmetries can become more pronounced, leading to increasing concentration of visibility and recognition. No single decision need be overtly discriminatory for this to occur; the effect emerges from the recursive interaction between system outputs and scholarly practices.

It is this cumulative and recursive character of the process that makes it necessary to distinguish epistemic algorithmic drift from algorithmic bias. Algorithmic bias refers to systematic disparities in the outputs or decisions of algorithmic systems arising from the interaction of data, model design, target definitions, and deployment conditions (Barocas & Selbst, 2016; Suresh & Guttag, 2021). In this sense, it concerns the distributional properties of outputs at a given point in time. Epistemic algorithmic drift, by contrast, captures the temporal dynamics through which those outputs, once taken up in practice, contribute to reorganizing the distribution of epistemic opportunity over time. Bias may contribute to drift, but drift is not reducible to it.

Nor does the concept coincide with *concept drift* in machine learning, which refers to situations in which the underlying data-generating distribution shifts so that a model trained under earlier conditions becomes less reliable when applied to later data (Gama et al., 2014). There, the problem is primarily statistical: the environment changes and the model must be adapted. Epistemic algorithmic drift addresses a different phenomenon: not that the model loses accuracy, but that its repeated use progressively reorganizes the social distribution of visibility, recognition, and opportunity.

The concept also differs from familiar representations of inequality. The metaphor of *the glass ceiling* presupposes a discrete barrier that can be located and removed. Drift operates instead through incremental adjustments in probability that, over time, narrow trajectories without any single point of exclusion. A closer parallel can be found in *path dependency* (Mahoney, 2000; Pierson, 2000), which explains how past decisions constrain future possibilities through self-reinforcing mechanisms. Yet while path dependency captures institutional persistence, epistemic algorithmic drift describes an ongoing redistribution of opportunity generated through the continuous operation of AI systems on historically stratified data. Addressing path dependency involves reforming institutional arrangements; addressing drift requires intervening in the feedback dynamics through which inequalities are continually reproduced.

What distinguishes epistemic algorithmic drift from adjacent frameworks is its focus on a process that links infrastructural design, user interaction, and cumulative epistemic outcomes. Research on algorithmic discrimination (Barocas & Selbst, 2016) identifies conditions under which protected attributes affect outputs but remains oriented toward model-level properties. Work on feedback loops (Crawford, 2021) highlights self-reinforcing dynamics yet does not specify their epistemic consequences. Infrastructure studies (Bowker & Star, 1999; Star, 1999) explain how such processes become naturalized but do

not provide a mechanism for tracing their accumulation over time. Epistemic algorithmic drift brings these perspectives together by specifying how these dynamics converge within AI-mediated scholarly communication.

The phenomenon can be examined through patterned distributions over time, provided that data are available at sufficient granularity and over extended periods. Four indicators are proposed, each specifying an empirical referent, a data requirement, and a minimum analytical design. Their value lies not in isolation but in triangulation, and in each case the challenge of distinguishing drift from other sources of concentration must be confronted directly.

The most directly traceable indicator is citation concentration: a longitudinal increase in the share of AI-recommended or AI-retrieved literature attributed to already highly cited authors, controlling for publication output (Larivière et al., 2013). Operationally, this requires linking bibliometric data from Scopus or Web of Science to recommendation system logs across a pre- and post-deployment window of at least five years within specific disciplinary fields.

A complementary indicator is demographic narrowing in editorial roles: reduced gender and geographic diversity in AI-assisted reviewer selection relative to the pool of qualified scholars available in a given field (Helmer et al., 2017). This requires longitudinal access to journal manuscript systems and reviewer databases disaggregated by gender and institutional affiliation, tracked across editorial rounds over at least three years, and considering policy shifts independent of AI adoption.

Another indicator is linguistic standardization, understood as the convergence of AI-assisted manuscripts toward dominant forms of standardized academic English. This can be measured through stylometric analyses of corpora of submitted and published manuscripts, comparing periods before and after the documented adoption of automated writing tools. The main analytical challenge lies in distinguishing convergence driven by drift from patterns of author self-selection or from shifts in journal language policies that may have tightened independently, in line with broader evidence on linguistic inequalities in scholarly publishing (Raitskaya & Tikhonova, 2025).

Finally, agenda narrowing, considered as declining methodological and theoretical diversity among highly visible publications, can be operationalized through topic modelling of titles, abstracts, and keywords across bibliometric archives spanning at least ten years, segmented by citation quartile. The primary confound is genuine paradigm consolidation within a field, which requires qualitative disciplinary judgment to distinguish from algorithmically amplified concentration (Fortunato et al., 2018).

None of these indicators is conclusive in isolation. Their evidential value depends on triangulation across dimensions, on intersectional disaggregation of results by gender, geography, and career stage, and on the availability of data linking AI system deployment to observed distributional shifts. Considering that such data remain largely unavailable at present, these indicators are offered not as a measurement protocol ready for

implementation but as a specification of what an empirical research programme addressing epistemic algorithmic drift would need to produce.

The intersectional dimension of drift is not an additional layer but a defining feature of the process. The feedback dynamics described above do not operate uniformly: their effects vary across configurations of structural position. As Crenshaw (1989) has shown, approaches that consider inequality along a single axis risk overlooking compounded forms of marginalization; similarly, auditing AI systems for gender or geographic bias separately may fail to capture how these dimensions interact in practice. A further dimension concerns disciplinary position: exposure to drift is highest where AI integration into editorial and discovery infrastructures is most advanced, where bibliometric proxies carry the greatest institutional weight, and where the gap between dominant and marginal methodological traditions is widest. Future empirical work will therefore need to be disciplinarily situated rather than cross-field in its initial design.

The implications extend beyond distributive concerns to the epistemic quality of science itself. A research system that progressively privileges already dominant perspectives (whether methodological, linguistic, or theoretical) risks narrowing the range of questions that can be asked and the forms of knowledge that can be recognized as valid. This does more than reproduce inequality: it reduces the system's capacity for reflexivity, critique, and innovation. Epistemic algorithmic drift is therefore not only a matter of fairness, but of the long-term robustness of knowledge production.

Discussion: Implications for Governance

The analysis developed in this paper shifts the focus of governance from isolated errors or access gaps to the cumulative effects of AI-mediated systems. If the problem were limited to measurable disparities in a model's outputs, technical correction might constitute part of the response. If it were limited to unequal access to technology, redistributive measures might be sufficient. Epistemic algorithmic drift exceeds both diagnoses. It is a process through which the routine operation of technically sound and widely accessible AI systems, under conditions of pre-existing inequality, produces progressive redistributions of epistemic opportunity that are cumulative, intersectional, and largely invisible to the actors they most affect. Governance responses adequate to this diagnosis must be structural, longitudinal, and intersectional in design. Three levels of response can be distinguished, each corresponding to a different locus of intervention: institutional governance, system design, and scholarly practice.

At the level of institutional governance, this diagnosis implies requirements that go beyond the transparency commitments that publishers and funders typically invoke. Binns (2018) argues that fairness in machine learning is not a purely technical problem but a political one, requiring the prior specification of normative criteria before any technical implementation can proceed. Barocas, Hardt, and Narayanan (2023) demonstrate that different formalisations of fairness are mathematically incompatible: a system

optimized for individual fairness cannot simultaneously satisfy group fairness criteria, and the choice among them cannot be delegated to the algorithm. These insights translate into concrete governance requirements: algorithmic impact assessments must specify in advance which conception of fairness a tool is designed to realize; monitoring of citation and recommendation data must be intersectional rather than single-axis; reporting on reviewer pools must be disaggregated along multiple axes simultaneously; and investment in language-assistance training corpora must be accompanied by explicit commitments to linguistic diversity, rather than treating the dominance of English as a neutral technical baseline (Oermann et al., 2025).

At the level of system design, a growing body of research on fairness-aware recommender systems demonstrates that equity-oriented objectives can be incorporated into recommendation algorithms without prohibitive costs to performance (Burke, 2017). These contributions show that the choice of optimization target is a normative decision with measurable distributional consequences: systems optimized exclusively for citation-based relevance will systematically disadvantage minority-group scholars in homophilic networks, while multisided fairness criteria can partially counteract this effect. Translating these findings into scholarly communication infrastructures requires that publishers and platform providers specify in advance which conception of fairness their systems are designed to realize.

Another level of response concerns scholarly practice, though its limits must be acknowledged from the outset. The concept of epistemic algorithmic drift motivates what might be called *infrastructural reflexivity*: a systematic attentiveness, on the part of researchers and research communities, to the AI-mediated systems through which their work is shaped, circulated, and evaluated. The term is used here as distinct from the broader notion of reflexivity in social science methodology: it refers not to awareness of one's own positionality, but to critical engagement with the technical and institutional arrangements that mediate epistemic outcomes — arrangements that, as Star (1999) observed, typically recede from view precisely when they function most effectively. Zhu, Wang, and Searsmith (2025) propose pedagogical frameworks for critical engagement with generative AI in academic writing that address how AI participation in knowledge production should be understood and communicated. Such frameworks are necessary but insufficient: infrastructural reflexivity at the individual level cannot compensate for drift that operates at the system level. Thus, individual reflexivity is a condition of awareness, not a mechanism of structural change.

A risk cuts across the previous levels: the assumption that identifying bias is equivalent to addressing it, as illustrated by the "audit-as-endpoint" problem. (Metcalf et al., 2021). Sandvig et al. (2014) distinguish between process audits and outcome audits, arguing that neither is sufficient without institutional mechanisms capable of acting on findings. Raji et al. (2020) propose end-to-end internal auditing frameworks that integrate impact assessment into system development cycles rather than treating it as an external check applied after deployment. Applied to AI-mediated scholarly infrastructures, this implies that auditing epistemic algorithmic drift requires not only

longitudinal measurement of its indicators but institutional arrangements within which publishers, funders, and research communities can deliberate on findings and mandate corrective intervention. Without such arrangements, monitoring merely documents the progression of drift. Effective governance therefore requires intervention in the conditions that produce it, specifically, in the design of training data, the weighting of recommendation criteria, and the accountability structures that determine who benefits when AI-mediated systems perform well and who bears the costs when they do not. By their nature, these interventions are political as well as technical, requiring forms of collective deliberation that current AI governance frameworks in scholarly communication have not yet systematically enabled.

Conclusions

Proposed as a theoretical construct requiring empirical validation, epistemic algorithmic drift accounts for the cumulative, feedback-driven redistribution of epistemic visibility and authority produced by AI-mediated scholarly infrastructures under conditions of pre-existing inequality.

Three principal conclusions follow. First, epistemic algorithmic drift is a structural feature of AI-mediated scholarly infrastructures, not a correctable error. The Matthew and Matilda effects are not displaced by algorithmic systems; they are encoded in training data, reproduced in recommendations, and reinforced through the data trails that AI-mediated interactions generate. What changes is the mode of operation: from socially recognizable processes of recognition and attribution, visible and therefore contestable, to infrastructurally embedded feedback mechanisms, diffuse and therefore resistant to identification and challenge. Addressing drift requires intervening in the feedback loop, not merely auditing its outputs.

Second, the effects of epistemic algorithmic drift are irreducibly intersectional. Gender, geography, language, career stage, and institutional position do not produce independent and additive risks; they constitute configurations of structural location whose combined effects are qualitatively distinct from the sum of their parts. An adequate empirical account of drift therefore requires intersectionally disaggregated data, that capture not which categories are marginalized but which configurations of position are most exposed to the compounding dynamics described here. The current absence of such data makes drift structurally less legible and therefore less contestable. Generating such data is a prerequisite for effective governance.

Third, the governance responses adequate to epistemic algorithmic drift are structural, normative, and deliberative. They require the prior specification of which conception of epistemic fairness AI-mediated systems should realize; longitudinal and intersectional monitoring frameworks capable of detecting drift before it consolidates into entrenched advantage; and forms of collective deliberation among the communities most affected. The feminist epistemological tradition, and the intersectionality framework in particular, offer not only a diagnostic lens but a normative horizon: the goal is not merely

the correction of bias, but the construction of epistemic infrastructures that actively support plurality, reflexivity, and the equitable distribution of the conditions under which knowledge can be produced and recognized. Developing auditing frameworks, diversity-aware training protocols, and deliberative governance structures constitutes a concrete agenda for future research. Operationalizing this agenda will require longitudinal and intersectionally disaggregated studies capable of testing the indicators specified above across concrete AI-mediated scholarly infrastructures. Without this evidence base, the construct remains analytically productive but empirically unconfirmed, a condition this paper acknowledges and treats not as a deficiency but as the starting point for a research programme.

The question is not whether artificial intelligence will shape the epistemic future of science. The question that matters, analytically and politically, is whether the infrastructural conditions under which AI operates will be deliberately designed to distribute epistemic visibility more equitably than the hierarchies it currently inherits. The intersectional governance frameworks outlined here do not resolve that question; they attempt to specify, with sufficient analytical precision, what resolving it would require.

Funding

No external funding was received.

Author Contribution

Paola Panarese is the sole author of this paper and carried out all stages of the research, including conceptualization, theoretical development, analysis, and manuscript preparation.

Tools of Artificial Intelligence

AI tools were used exclusively for linguistic revision, stylistic refinement, and structural editing of the manuscript. The tools employed were general-purpose large language model assistants applied solely at the level of prose editing.

References

Scientific research

1. Amayreh, A., & Amayreh, M. (2025). The interactive theory of artificial intelligence in academic knowledge production. *International Journal of Academic Research in Business & Social Sciences, 15*(5), 794–811. https://doi.org/10.6007/IJARBSS/v15-i5/25468
2. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732. https://doi.org/10.15779/Z38BG31

3. Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.
4. Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Research-paper recommender systems: A literature survey. *International Journal on Digital Libraries, 17*(4), 305–338. https://doi.org/10.1007/s00799-015-0156-0
5. Benjamin, R. (2019). *Race after technology*. Polity.
6. Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 149–159. https://doi.org/10.1145/3287560.3287583
7. Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. MIT Press.
8. Burke, R. (2017). Multisided fairness for recommendation. *Proceedings of the Conference on Fairness, Accountability and Transparency*, 129–137.
9. Code, L. (1991). *What Can She Know? Feminist Theory and the Construction of Knowledge*. Cornell University Press
10. Collins, P. H. (1990). *Black feminist thought*. Routledge.
11. Crawford, K. (2021). *Atlas of AI*. Yale University Press.
12. Crenshaw, K. (1989). Demarginalizing the intersection of race and sex. *University of Chicago Legal Forum*, 139–167.
13. Crenshaw, K. (1991). Mapping the margins. *Stanford Law Review, 43*(6), 1241–1299.
14. D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press.
15. Dorta-González, P., López-Puig, A. J., Dorta-González, M. I., & González-Betancor, S. M. (2024). Generative artificial intelligence usage by researchers at work: Effects of gender, career stage, type of workplace, and perceived barriers. *Telematics and Informatics, 94*, 102187. https://doi.org/10.1016/j.tele.2024.102187
16. Eren, M. E., & Perez, D. M. (2025). *Rethinking science in the age of artificial intelligence*. arXiv. https://doi.org/10.48550/arXiv.2511.10524
17. Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., & Barabási, A.-L. (2018). Science of science. *Science, 359*(6379), eaao0185. https://doi.org/10.1126/science.aao0185
18. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys, 46*(4), 44. https://doi.org/10.1145/2523813
19. Guarino, C. M., & Borden, V. M. H. (2017). Faculty service loads and gender. *Research in Higher Education, 58*(6), 672–694. https://doi.org/10.1007/s11162-016-9454-2
20. Haraway, D. (1988). Situated knowledges. *Feminist Studies, 14*(3), 575–599.
21. Harding, S. (1986). *The science question in feminism*. Cornell University Press.

22. Harding, S. (1991). *Whose science? Whose knowledge?* Cornell University Press.
23. Heaven, D. (2018). *AI peer reviewers unleashed to ease publishing grind*. Nature, 563(7733), 609–610. https://doi.org/10.1038/d41586-018-07245-9
24. Helmer, M., Schottdorf, M., Neef, A., & Battaglia, D. (2017). Gender bias in scholarly peer review. *eLife, 6*, e21718. https://doi.org/10.7554/eLife.21718
25. Hosseini, M., & Horbach, S. P. J. M. (2023). Fighting reviewer fatigue or amplifying bias? *Research Integrity and Peer Review, 8*, 4. https://doi.org/10.1186/s41073-023-00133-5
26. Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers. *Proceedings of the National Academy of Sciences, 117*(9), 4609–4616. https://doi.org/10.1073/pnas.1914221117
27. Karimi, F., Génois, M., Wagner, C., Singer, P., & Strohmaier, M. (2018). Homophily influences ranking of minorities in social networks. *Scientific Reports, 8*, 11077. https://doi.org/10.1038/s41598-018-29405-7.
28. King, M. M., Bergstrom, C. T., Correll, S. J., Jacquet, J., & West, J. D. (2017). Men set their own cites high. *Socius, 3*. https://doi.org/10.1177/2378023117738903
29. Kousha, K., & Thelwall, M. (2024). Artificial intelligence to support publishing and peer review: A review. *Learned Publishing, 37*(1), 4–12. https://doi.org/10.1002/leap.1570
30. Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Global gender disparities in science. *Nature, 504*, 211–213. https://doi.org/10.1038/nature12695
31. Longino, H. (1990). *Science as social knowledge*. Princeton University Press.
32. Longino, H. (2002). *The fate of knowledge*. Princeton University Press.
33. Lund, B. D., Lamba, M., & Oh, S. H. (2024). The impact of AI on academic research and publishing. *arXiv*. https://doi.org/10.48550/arXiv.2406.06009
34. Mahoney, J. (2000). Path dependence in historical sociology. *Theory and Society, 29*(4), 507–548.
35. Merton, R. K. (1968). The Matthew effect in science. *Science, 159*(3810), 56–63.
36. Metcalf, J., Moss, E., & boyd, d. (2021). Owning ethics. *Social Studies of Science, 51*(4), 568–588. https://doi.org/10.1177/03063127211013804
37. Noble, S. U. (2018). *Algorithms of oppression*. NYU Press.
38. Nicholas, D., Świgoń, M., Clark, D., Abrizah, A., Revez, J., Herman, E., Rodríguez Bravo, B., Xu, J., & Watkinson, A. (2024). The impact of generative AI on the scholarly communications of early career researchers: An international, multi-disciplinary study. *Learned Publishing, 37*(4), Article e1628. https://doi.org/10.1002/leap.1628
39. Oermann, M., Owens, J., Carter-Templeton, H., Peterson, G., & Bailey, H. (2025). Using artificial intelligence for scholarly writing. *The American*

Journal of Nursing, 125(11), 52–55. https://doi.org/10.1097/AJN.0000000000000179

40. Pierson, P. (2000). Increasing returns, path dependence, and the study of politics. *American Political Science Review, 94*(2), 251–267.
41. Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts. *Social Studies of Science, 14*(3), 399–441.
42. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873
43. Rossiter, M. W. (1993). The Matthew Matilda effect in science. *Social Studies of Science, 23*(2), 325–341
44. Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, 22, 1–23.
45. Schiebinger, L. (1999). *Has feminism changed science?* Harvard University Press
46. Star, S. L. (1999). The ethnography of infrastructure. *American Behavioral Scientist, 43*(3), 377–391.
47. Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm in machine learning. *Communications of the ACM, 64*(12), 36–45.
48. Wajcman, J. (2004). *TechnoFeminism.* Polity
49. Wennerås, C., & Wold, A. (1997). Nepotism and sexism in peer review. *Nature, 387*, 341–343.
50. Winner, L. (1980). Do artifacts have politics? *Daedalus, 109*(1), 121–136.
51. Zhao, X., & Zhang, Y. (2022). Reviewer assignment algorithms for peer review automation: A survey. *Information Processing & Management, 59*(5), 103028. https://doi.org/10.1016/j.ipm.2022.103028
52. Zhu, X., Wang, C., & Searsmith, D. (2025). *Writing with machines and peers: Designing for critical engagement with generative AI*. arXiv. https://doi.org/10.48550/arXiv.2511.15750

Other sources

53. Elsevier (2024). *Elsevier insights: AI in scholarly publishing*.
54. UNESCO Institute for Statistics. (2024). *Women in science statistics*