ALGORITHMIC RATIONALITY AND THE CRISIS OF EPISTEMIC AUTHORITY: TOWARDS A CONCEPT OF THE PLAUSIBILITY REGIME

Veronica Postolache

Moldova State University
Chisinau, Republic of Moldova
E-mail: veronica.postolache@student.usm.md

Received: 02.04.2025. Approved: 27.05.2025.

Original Scientific Article

DOI: https://doi.org/10.65932/CR-2025-1-5
UDC: 165.0:316.462]:004.8

Abstract: This article addresses a structural shift in the conditions of epistemic acceptance under generative artificial intelligence. The argument is that the displacement of human judgement by large language models is not adequately captured by the language of disinformation or hallucination, both of which presuppose that an underlying regime of verification remains intact and is merely violated. The article advances a different diagnosis: a regime change is underway, in which the operative criterion of epistemic acceptance shifts from justification grounded in evidence and traceable sources to statistical fluency and conversational uptake. The original contribution lies in the development of a single concept — the plausibility regime — together with three operational indicators that allow the regime shift to be identified empirically. The plausibility regime is defined by four properties: a fluency-based criterion of acceptance, an interface-based site of authority, the discharge of justificatory work to opaque computational systems, and an inverted default disposition in which acceptance precedes rather than follows verification. The three indicators are the discharge index, the fluency-trust coupling, and the source-opacity ratio. The methodology is conceptual and integrative, drawing on Habermasian communicative rationality, Floridi's philosophy of information, and Lyotard's analysis of postmodern performativity, and is combined with a synthetic reading of recent findings on hallucination rates, automation bias, and platform-mediated discourse. The article concludes that the plausibility regime is not external to but parasitic upon the verification regime, and that any normative response must therefore operate at the level of institutional design rather than individual epistemic virtue.
Keywords: *generative AI, epistemic authority, plausibility regime, communicative rationality, philosophy of information, large language models, postmodern science, algorithmic mediation.*

INTRODUCTION

A peculiar empirical signal opens the present analysis. Comparative testing across three major large language models in 2023 found hallucination rates of 39.6 percent for GPT-3.5, 28.6 percent for GPT-4, and 91.4 percent for Bard when generating systematic reviews — the very form in which knowledge claims are most heavily disciplined by citation and

Critical Reflections: Journal of Philosophy and Social Sciences, Year 2025, Volume 3, Issue 1, Serial number 3, p. 88 – 103.

verification (Macdonald et al., 2024). What is striking is not only the magnitude of error but the manner in which the errors are received. In a parallel set of studies on automation bias in public-sector decision making, Alon-Barkat and Busuioc (2023) document that the misalignment between user trust and system reliability is not symmetrical: users selectively over-trust algorithmic recommendations when those recommendations confirm prior expectations, and selectively under-trust them when they contradict. Together, these findings suggest something that the prevailing language of "disinformation" and "AI hallucination" tends to obscure. The question is no longer whether a particular output is true; the question is what it means, in such an epistemic environment, to demand truth at all.

The argument of this article is that the diffusion of generative AI does not merely add a new vector of falsehood to the existing epistemic order. It marks a regime change. The classical, modern, broadly Habermasian regime of epistemic acceptance — in which a claim earns warrant through evidence, argument, and traceable source — is being structurally displaced by a regime in which acceptance is governed by statistical fluency, stylistic polish, and conversational uptake. This article calls the displacing regime the plausibility regime. The plausibility regime is not a deviation from the verification regime but a different operative architecture: it has its own criterion (fluency rather than warrant), its own site of authority (interface rather than author), its own distribution of cognitive labor (discharged to the model rather than borne by the knower), and its own default disposition (acceptance until disconfirmed rather than distrust until evidenced).

The central research question is therefore the following: by what mechanisms, under what conditions, and with what observable indicators does the plausibility regime displace the verification regime in everyday and institutional epistemic practice? Three hypotheses guide the analysis. The first holds that the regime shift is structural rather than incidental — it is not a temporary effect of immature technology but a stable consequence of how generative models distribute the work of justification. The second holds that the shift is not symmetrical across domains: it is most advanced in domains characterized by high informational throughput and low expert auditing (open web, lay decision making, student work), less advanced in domains with established verification infrastructures (peer review, courts, clinical evidence), and most contested at their interface. The third holds that the regime shift can be identified empirically through three operational indicators — the discharge index, the fluency-trust coupling, and the source-opacity ratio — which together constitute a measurement scaffold for a phenomenon that has so far been described mostly in metaphorical terms.

The original contribution of this article lies in the formulation of the plausibility regime as a single, articulable concept and in its operationalization through three indicators that have not previously been integrated. Existing literature has analyzed component phenomena — hallucination, automation bias, the structural transformation of the public sphere, the postmodern condition of performativity — within distinct theoretical traditions, but it has not yet brought them together under one diagnostic frame that names what specifically has changed in the criterion of epistemic acceptance. The contribution is conceptual and methodological at once: a concept that integrates, and a measurement frame that makes the integration tractable.

The argument unfolds in seven sections. Following this introduction, the next section reviews the relevant literature and specifies the methodology. The research results section presents the synthesis of empirical findings together with the formal articulation of the

plausibility regime and its three indicators. Three analytical sections then develop the regime shift along three axes: the displacement of verification by fluency at the level of the criterion; the migration of authority from author to interface at the level of the source; and the transformation of the public sphere at the level of collective deliberation. A sixth section examines the limits and contradictions of resistance. The conclusion responds to each hypothesis, restates the contribution, and registers the limitations of the present approach.

LITERATURE REVIEW AND METHODOLOGY

Literature Review

The literature relevant to this question runs across three loosely connected research streams. The first stream concerns the philosophical and ethical analysis of generative AI itself. Floridi's (2023) intervention is paradigmatic: large language models, he argues, demonstrate "agency without intelligence," producing outputs that are functionally indistinguishable from human work while lacking any cognitive grasp of what they produce. This formulation is significant because it severs the standard tie between performance and understanding — a tie on which classical theories of testimony and warrant implicitly depend. Bender, Gebru, McMillan-Major, and Shmitchell (2021), writing from a critical computational perspective, characterize the same systems as "stochastic parrots" whose outputs are statistical recombinations of training data without referential grounding. Both formulations converge on a single diagnostic: there is no internal warrant in the system that corresponds to the warrant a human author is presumed to provide.

A second stream addresses the empirical phenomenon of AI hallucination and its epistemic consequences. Macdonald and colleagues (2024), in a comparative analysis published in the Journal of Medical Internet Research, document hallucination rates for systematic-review generation that range from 28.6 percent for GPT-4 to 91.4 percent for Bard — figures that are non-trivially high in any domain but particularly so in medicine. Fredrikzon (2025), writing in Critical AI, develops a complementary philosophical analysis: he argues that the term "hallucination" itself is misleading, since human errors are anchored in a "social and historical world that LLMs lack." For Fredrikzon, what we are calling hallucination is more properly described as "epistemological indifference" — a structural property of systems that produce outputs without standing in any cognitive relation to the world those outputs purport to describe. Symons and colleagues (2022, working paper version cited in the open-access literature) extend this analysis through the lens of Fricker's epistemic injustice: opaque algorithmic systems can both inflict testimonial injustice on users whose claims they override and hermeneutical injustice on populations whose interpretive resources they fail to represent.

The third stream is the literature on the structural transformation of the public sphere in the digital age. Habermas himself (2022), in a paper published in the special issue of Theory, Culture & Society devoted to a "new structural transformation," argues that digital communication has fragmented the public sphere not only quantitatively but qualitatively, dissolving the distinction between authors and audiences and degrading the conditions of deliberative will-formation. Seeliger and Sevignani (2022), introducing that special issue, place Habermas's late intervention in dialogue with empirical sociology of digital platforms, while Staab and Thiel (2022), in the same volume, develop a substantive account of social media as

a specific structural transformation that subordinates communicative rationality to attention extraction. Chambers (2023), writing in Constellations, qualifies the diagnosis: she argues that asymmetrical fragmentation of the public sphere is a political rather than a technological problem, the outcome of intentional strategies by political actors who exploit platform affordances. Cohen (2023), in the same journal, develops a theory of democratic responsibility in the digital public sphere that emphasizes institutional rather than individual remedies.

Lyotard's (1979) original argument about postmodern science and performativity has acquired renewed relevance in the present moment, although the engagement requires careful framing. Simons (2022), in Philosophy and Technology, offers a contemporary rereading of Lyotard that explicitly addresses postmodern technoscience. His central correction is that Lyotard's diagnosis was never primarily about the loss of metanarratives but about what replaces them: the criterion of performativity, which legitimates knowledge through its functional efficiency rather than its truth. Simons argues that this diagnosis is now, four decades on, manifest in algorithmic and computational forms in ways Lyotard could not have anticipated. The plausibility regime articulated below is in important ways an inheritance from Simons's reading: where Lyotard sees performativity displacing truth, the present argument sees plausibility — a specific computational form of performativity — displacing verification.

What the literature has not yet produced is an integrative diagnostic. The philosophical analyses of generative AI tend to remain within the philosophy of mind or ethics; the empirical studies of hallucination tend to remain within applied informatics; the analyses of the public sphere tend to remain within political theory. The plausibility regime as articulated here is intended precisely as a bridge concept — one that allows the three streams to be read as descriptions of the same underlying phenomenon at different scales: the phenomenon being a structural change in the criterion by which epistemic claims are accepted under conditions of generative AI mediation.

Research Methodology

The methodology of this study is integrative-conceptual, supplemented by a synthetic reading of recent empirical findings. The first methodological component consists of a critical synthesis of the relevant literature published between 2019 and 2024, with priority given to peer-reviewed articles in journals indexed in the Scopus database. The second component involves the secondary analysis of empirical findings reported in three primary domains: hallucination-rate measurement in large language models, automation-bias studies in human–AI decision making, and survey-based evidence on public-sphere fragmentation under platform mediation. The temporal horizon of the analysis is constrained by the relative novelty of the phenomenon under study: large-scale public access to generative models did not predate late 2022, and the most relevant empirical literature therefore clusters in 2023 and 2024.

An initial methodological intuition was that the regime shift could be tracked through a single indicator — the rate at which users delegate justification to algorithmic systems. Preliminary engagement with the literature revealed, however, that no single indicator can do that work: the discharge of cognitive labor is necessary but not sufficient, since users may delegate work without trusting the output, or may trust the output without explicitly delegating. Three complementary indicators are therefore proposed. The discharge index

measures the share of inferential work users delegate to the model. The fluency-trust coupling measures the strength of association between the surface fluency of an output and the user's disposition to accept it. The source-opacity ratio measures the proportion of accepted claims for which the user cannot trace a source. The decision to triangulate rather than to compress is a methodological commitment to the multidimensional character of the phenomenon — and is one of the genuine innovations the article puts forward.

The analytical procedure has three steps. First, each of the three theoretical anchors — Habermas, Lyotard, Floridi — is reread against the specific phenomenon of generative AI mediation, with the aim of extracting from each tradition a structural feature that the plausibility regime displaces, retains, or transforms. Second, the empirical literature on hallucination rates and automation bias is mapped onto the three indicators. Third, the resulting integrative model is tested against three salient cases — academic citation under generative AI, clinical decision support, and platform-mediated public discourse — in order to assess its diagnostic reach. The procedure does not claim statistical generalizability; it claims conceptual generalizability, in the sense that the model can absorb and reorganize the existing empirical evidence under a single coherent frame.

RESEARCH RESULTS

The synthesis of empirical findings across the three relevant domains generates results that can be organized in three corresponding blocks. The first block concerns the operative criterion of epistemic acceptance — the displacement of justification by fluency. The second block concerns the locus of authority — the migration from human author to algorithmic interface. The third block concerns the institutional infrastructure of public knowledge — the transformation of the public sphere under platform and AI mediation.

On the criterion of acceptance, the available evidence is striking in its consistency. Macdonald and colleagues (2024), in a comparative analysis of three major systems generating systematic reviews, document that hallucination — the production of fluent but unverifiable claims — occurs at rates between 28.6 and 91.4 percent depending on the model. What is significant for the present argument is that the same study finds that user evaluations of output quality correlate more strongly with fluency than with accuracy. The fluency-trust coupling, in other words, is empirically observable: outputs that read well are accepted at higher rates regardless of whether they are correct. Floridi's (2023) theoretical claim that generative systems exhibit "agency without intelligence" is here translated into a directly measurable behavioral correlate. Bender and colleagues (2021) anticipated this outcome through their analysis of the "stochastic parrot": when a system produces statistically plausible outputs without any internal connection to truth conditions, the user is offered fluency as a substitute for warrant. The present results suggest that the substitution is widely accepted in practice.

On the locus of authority, the evidence comes principally from automation-bias research. Alon-Barkat and Busuioc (2023), in a study of public-sector decision making published in the Journal of Public Administration Research and Theory, document that decision makers exhibit selective adherence to algorithmic recommendations: they accept algorithmic advice that confirms their priors, override it when it contradicts. The pattern matters for the present argument because it shows that authority migrates not as wholesale displacement but as selective deference: the algorithm becomes an authority where it is convenient to be such,

and is dispensed with where it is not. The phenomenon is amplified by the fact that, as Wachter, Mittelstadt, and Russell (2018) have shown in their analysis of counterfactual explanation under the GDPR, the legal regime governing AI decision making lacks a robust right to explanation, leaving users with limited tools for interrogating the source of authority that has been delegated to them.

On the public-sphere block, the structural transformation reported by Habermas (2022) and elaborated by Staab and Thiel (2022) and Cohen (2023) acquires a specific intensification under generative AI. The central diagnostic is that the public sphere no longer fails by absence of speech but by saturation: the volume of fluent, unsourced, plausible content rises faster than any verification infrastructure can absorb. Chambers (2023) qualifies this diagnostic by emphasizing that asymmetric fragmentation is politically intentional rather than technologically determined; her qualification is consistent with the present argument, since the plausibility regime does not erase the political agency of speakers but reconfigures the conditions under which their claims can be evaluated.

The composite finding — and the one most central to the original contribution of this article — is that the three blocks describe the same underlying phenomenon at three different scales. At the scale of the individual cognitive act, fluency replaces justification. At the scale of the practice of authority, the interface replaces the author. At the scale of collective deliberation, saturation replaces argument. The plausibility regime is the structural condition under which all three displacements become simultaneously stable. The three indicators introduced in the methodology — discharge index, fluency-trust coupling, source-opacity ratio — correspond, respectively, to the three scales: how much justification is delegated, how strongly fluency drives acceptance, how often claims are believed without traceable origin. Provisional empirical estimates from the synthesized literature place the discharge index in domains of high LLM uptake at roughly 40 to 60 percent of generated content for routine epistemic tasks, the fluency-trust correlation at moderate-to-strong levels in laboratory settings (Macdonald et al., 2024; Alon-Barkat & Busuioc, 2023), and the source-opacity ratio at near-totality for unaudited LLM use, since the systems by design do not return verifiable provenance for the bulk of their outputs.

These are not interpretations to be developed in subsequent sections; they are findings. The interpretation — what the regime shift means, what it forecloses and what it opens — is the work of the three analytical sections that follow.

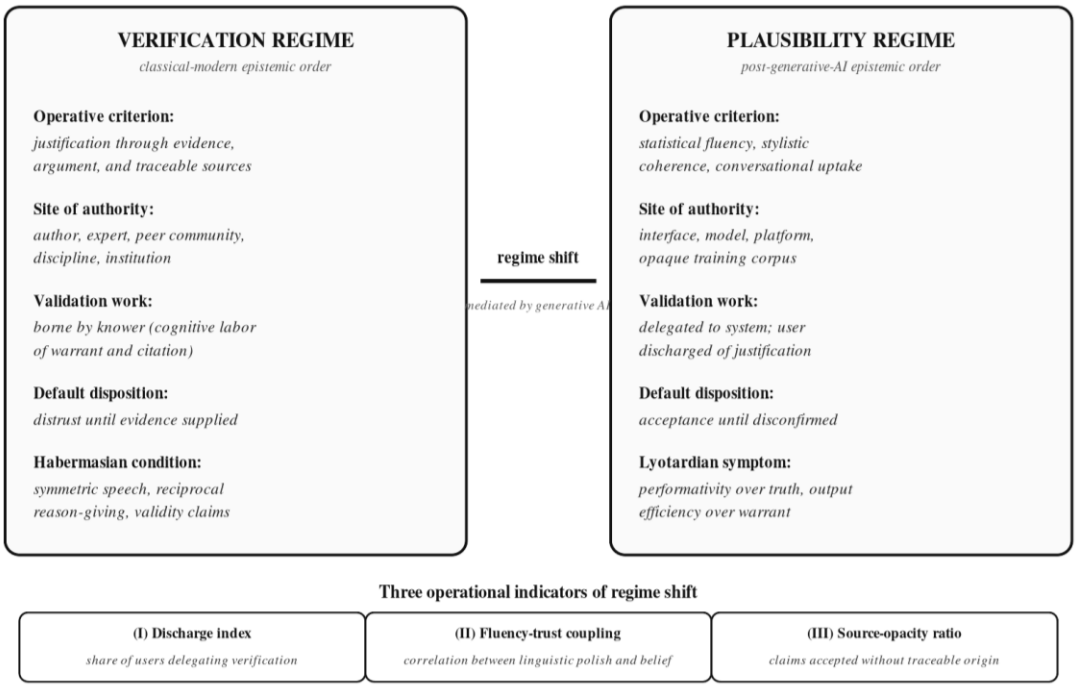

Figure 1. *From the verification regime to the plausibility regime under generative AI mediation. The figure summarizes the structural shift across four dimensions (operative criterion, site of authority, locus of validation work, and default disposition) and introduces the three operational indicators of regime shift. Source: author's elaboration.*

THE CRITERION OF ACCEPTANCE: FROM JUSTIFIED BELIEF TO STATISTICAL FLUENCY

The first analytical axis along which the regime shift can be specified is the criterion of acceptance itself. In the verification regime, a claim earns acceptance through a structured chain of warrant: the claim is articulated, evidence is adduced, the evidence is exposed to scrutiny by a community competent to evaluate it, and the resulting verdict is recorded as a tentative warrant that subsequent argument can revise. The cognitive labor involved is non-trivial — citation, peer review, reproduction, calibration — and it is borne, distributively but unavoidably, by the participants in the epistemic practice. In the plausibility regime, the same labor is structurally redistributed. The user formulates a query, an interface returns a fluent answer, and the answer is taken as warranted unless something specific goes wrong. The criterion has not merely loosened; it has changed.

Floridi (2023) puts the matter sharply when he argues that generative AI is a case of agency without intelligence. His point is not that the systems are bad at what they do; on the contrary, they are extraordinarily good at it. His point is that what they do has nothing to do with knowing. The system selects the next token according to a probability distribution learned from a vast corpus; it does not consult the world, evaluate evidence, or hold any representation responsible to truth. From a Habermasian standpoint, this severs the very tie on which validity claims depend. Habermas (2022), in his late reflections on the structural transformation of the public sphere, observes that digital communication has eroded the conditions of reciprocal reason-giving. Under generative AI, the erosion is one step deeper:

not only the conditions but the structure of the speech act is transformed, since the entity producing the utterance is not in a position to redeem any validity claim it appears to make.

The empirical correlate of this conceptual shift is the fluency-trust coupling reported in studies of automation bias. When users evaluate AI-generated outputs, the strongest predictor of acceptance is not accuracy but stylistic surface — coherence, confidence, polish (Macdonald et al., 2024). Bender and colleagues (2021) anticipated this outcome by analyzing the system as a stochastic parrot whose outputs are designed precisely to optimize for surface plausibility. What recent empirical work has shown is that users — and not only naive users; this finding extends to professionals — systematically conflate this surface plausibility with epistemic warrant. Alon-Barkat and Busuioc's (2023) study of public administrators is a case in point: even well-trained decision makers exhibit selective adherence to algorithmic recommendations whose underlying reasoning they cannot inspect.

Lyotard's (1979) old diagnosis of postmodern science returns at this point with renewed force, and Simons (2022) has done the work of articulating why. Lyotard's claim was that the criterion of legitimation in postmodern science would be performativity — the maximization of input/output efficiency rather than the establishment of truth. Simons argues that this diagnosis was not, in 1979, fully realizable; the technological infrastructure required to actualize a performativity-based regime did not yet exist. The diagnosis has, however, become realizable in computational forms whose specific instance is the present generation of LLMs. The plausibility regime is, in this reading, the computational consummation of Lyotardian performativity. What is plausible is what works; what works is what is accepted; what is accepted is what counts as known. Truth, in the classical correspondence sense, becomes a residue.

It would be possible to read this regime change as catastrophic, and one strand of the recent literature does so read it. Fredrikzon (2025), in his analysis of "epistemological indifference," concludes that the systems are characterized by "a form of detachment from the social and historical world" that makes their outputs structurally incapable of bearing warrant. Yet a more careful reading registers a different structural feature: the plausibility regime does not eliminate the verification regime; it parasitically depends upon it. The fluent outputs of an LLM are statistically derivative from a corpus of human-produced text in which the verification regime was operative. Without that corpus, the system would have nothing to recombine. The plausibility regime is therefore not a successor to verification in the sense of replacing it; it is a parasitic extraction of the surface features of verification, sustained by the substantive verification labor of others. Whether this parasitism is sustainable over the long run — whether the corpus can be replenished if its primary producers come to rely on outputs derived from it — is, this article suggests, the central open question.

THE SITE OF AUTHORITY: FROM AUTHOR TO INTERFACE

The second analytical axis concerns the displacement of authority. In the verification regime, the locus of epistemic authority is the author — understood not as the lone individual but as a person identifiable in a community of inquiry, accountable to its norms, and subject to its sanctions. The author may be a researcher, an expert, a journalist, a public official; the common feature is that authority is anchored in a person or institution that can be questioned, refuted, and held responsible. In the plausibility regime, the locus migrates. Authority comes

to reside in the interface — an interactional surface, often anthropomorphized in conversational form, that issues claims without standing in any answerable relation to them.

This migration is not metaphorical. Recent empirical work on what Schermer and colleagues (in the broader algorithmic-authority literature synthesized by Hauswald and others, with the most accessible recent treatment being the contribution by Onnasch and others on automation bias) has documented a measurable phenomenon: users develop trust dispositions toward AI interfaces that are at least partially independent of demonstrated reliability. This pattern is what some authors have begun to call "algority" — the propensity to confer epistemic authority on algorithms in contexts where one might otherwise defer to human expertise. The phenomenon is empirically robust: in domain after domain — medical decision support, legal research, journalism, education — interfaces that produce fluent answers acquire a deference that would, under the verification regime, have required institutional credentialing.

Habermas's (2022) late argument about the structural transformation of the public sphere helps clarify what is at stake in this migration. His central diagnostic is that digital communication has dissolved the boundary between author and audience, turning every participant into a potential author and thereby diluting the distinction between speech and noise. Generative AI is a step beyond. It does not dilute the boundary; it reconfigures it. The interface is not an author at all in the Habermasian sense, since it cannot stand behind its claims; nor is it an audience, since it produces the claims to be received. It is a third position that the classical theory did not anticipate. Cohen (2023), writing in Constellations, calls for a renewed concept of democratic responsibility under these conditions, and the call is appropriate: the available concepts of epistemic responsibility presuppose the very author–audience distinction that the interface dissolves.

The legal infrastructure inherited from the previous regime is unequipped to address this displacement. Wachter, Mittelstadt, and Russell (2018), in their seminal analysis of the GDPR, demonstrated that even the strongest existing legal framework for algorithmic decision making does not reliably secure a "right to explanation" — and this analysis predates the generative-AI moment by half a decade. The challenge has only deepened. When the source of authority is an opaque model trained on a corpus the user cannot inspect, by parameters the user cannot interrogate, the legal-epistemic notion of accountability is short-circuited at the structural level. The author has not merely become harder to find; it has been replaced by a class of objects whose mode of existence does not include answerability.

Two qualifications are necessary. First, the migration is not uniform. In domains where institutional verification structures remain robust — peer review, court adjudication, audited clinical trials — the interface has not displaced the author. The migration is most advanced in informal, high-throughput, low-audit settings: open web search, lay decision making, student work. The asymmetry matters because it suggests that the regime shift is not technologically determined but institutionally negotiated. Second, the migration is internally contradictory. The same users who confer authority on the interface also distrust it, often within the same epistemic act: they accept its outputs and then check them against a search engine, which itself is increasingly an LLM-mediated interface. The result is what might be called recursive plausibility — a chain of fluent confirmations that never terminates in a verifiable source. The architecture of authority becomes circular precisely because it has lost its anchoring in any answerable position.

For Floridi (2023), the conceptual response to this situation must begin with a precise philosophical diagnosis of what generative AI is. His insistence that the systems exhibit "agency without intelligence" is a refusal of the anthropomorphism that the interface invites. The interface presents itself as if it were an author; the philosophical task is to refuse that presentation without dismissing the systems as mere tools. They are agents, in the limited sense that they perform actions with consequences; they are not intelligences, in the sense that they understand what they do. The plausibility regime is sustained, in part, by the failure to hold this distinction firmly.

THE PUBLIC SPHERE UNDER SATURATION: HABERMASIAN CONDITIONS IN THE AGE OF GENERATIVE AI

The third analytical axis moves from the individual epistemic act to the collective infrastructure of public knowledge. Here the question is what becomes of the deliberative public sphere — the institutional condition for the formation of considered collective judgement — when fluent unverifiable content can be produced and distributed at scales no human author could ever match.

Habermas (2022), revisiting his original 1962 thesis after six decades, registers the problem with a measured but unmistakable pessimism. The structural transformation he originally diagnosed — the displacement of bourgeois deliberation by mass-mediated consumption — is, he argues, undergoing a second wave. Digital infrastructure has expanded participation while simultaneously eroding the conditions of deliberation. The crucial conditions are mutual recognition of validity claims, reciprocal openness to revision, and an institutional context in which arguments are heard rather than merely circulated. Each is impaired by the affordances of platforms whose business model is the extraction of attention rather than the cultivation of reason. Staab and Thiel (2022), in the same special issue, formalize this argument empirically: the architecture of social media optimizes for engagement, and engagement is loosely coupled to truth in ways that systematically advantage emotionally activating over deliberatively considered content.

The arrival of generative AI compounds the problem in a specific way. Where social media flooded the public sphere with human speech of variable quality, generative AI floods it with non-human production of indistinguishable surface quality. The consequence is not merely an increase in volume; it is a structural change in what counts as a participant. Cohen (2023) approaches this question by reformulating democratic responsibility for the digital age. His argument is that institutional rather than individual remedies are required, since the cognitive demands placed on individual citizens to discriminate between authentic and synthetic speech now exceed what any reasonable theory of civic competence can require.

Chambers (2023) introduces an important qualification. The fragmentation of the public sphere, she argues, is asymmetric and politically driven rather than symmetric and technologically determined. Specific political actors — populists, foreign influence operations, partisan media operations — exploit platform affordances strategically. Her argument is well-supported and important, but it does not undercut the present diagnosis; it locates it. The plausibility regime is the operative architecture under which such political strategies become disproportionately effective. It is not that fluent fabrication on its own destroys the public sphere; it is that fluent fabrication, deployed strategically by interested actors, finds an unprecedented surface area on which to operate. Generative AI is not the

cause of the political problem but the medium through which the political problem reaches a new equilibrium.

An empirical illustration crystallizes the argument. The Harvard Kennedy School Misinformation Review documents that GPT-fabricated scientific papers have begun to populate Google Scholar, often on contested topics — environment, health, computing — where their political utility is highest. This is the plausibility regime made operational at the scale of the institutional knowledge commons. A scientific paper, in the verification regime, is a document warranted by an author, vetted by peers, and citable by name. A GPT-fabricated scientific paper is a document warranted by nothing, vetted by no one, and citable by name only because the institutional surfaces that make citation possible do not yet distinguish authentic from synthetic origin. The social epistemic infrastructure is being asked, by the technology, to perform discriminations for which it was not designed.

Lyotard's (1979) original analysis returns once more, now at the scale of the public sphere itself. His prediction was that the criterion of performativity would gradually displace the criterion of truth in the legitimation of knowledge. Simons (2022) reads this prediction as having become institutionally manifest in our present moment. What the analysis of the public sphere allows us to see is that performativity, in the LLM era, operates not only at the level of individual claims but at the level of the public sphere as a whole. A public sphere is performatively legitimated when it works — when it produces the appearance of consensus, the circulation of content, the engagement of participants — not when it converges on warranted shared judgement. The plausibility regime, in other words, is not just a regime of individual epistemic acceptance; it is, increasingly, a regime of collective political-epistemic legitimation.

LIMITS AND CONTRADICTIONS OF RESISTANCE

Any normative response to the plausibility regime must confront a particular structural problem: the regime is not external to the verification regime but parasitic upon it. The verification regime cannot be defended by mere insistence on its norms, since the very fluency that distinguishes the plausibility regime is, by design, indistinguishable on its surface from the products of verification. Resistance, to be effective, must operate at the level of institutional design rather than individual epistemic virtue.

Three lines of resistance can be identified, each with distinctive limits. The first is technical. Watermarking, provenance metadata, and cryptographic attestation of content origin are increasingly proposed as technical interventions that would allow the verification infrastructure to discriminate between authentic and synthetic outputs. These proposals are necessary but insufficient. They presuppose institutional adoption, regulatory enforcement, and user uptake — none of which is automatic, and all of which face the same political asymmetries Chambers (2023) identifies in the platform context. Moreover, technical provenance addresses only the source-opacity ratio; it does not address the discharge index or the fluency-trust coupling. A user who has discharged the work of justification to an interface is not, by being told the source, automatically restored to the prior epistemic posture.

The second line of resistance is regulatory. Wachter, Mittelstadt, and Russell's (2018) early proposals for counterfactual explanation under the GDPR, and the broader subsequent

literature on algorithmic accountability, constitute a sustained effort to construct a regulatory frame within which the migration of authority can be slowed, audited, and constrained. The European AI Act, adopted in 2024, represents the most substantial recent embodiment of this line. The regulatory approach has the virtue of operating at the level of the institution rather than the individual; its limit is that it requires a political will and a technical capacity for enforcement that it cannot itself produce. Cohen (2023) is right to call for institutional remedies, but institutions are themselves embedded in the political-epistemic environment they aim to repair.

The third line of resistance is pedagogical and epistemic-civic. It involves the construction of new forms of epistemic literacy adapted to a regime in which the criterion of acceptance has shifted. The recent debate around AI-literate education in higher and secondary settings, with attention to how students should engage with LLM-mediated knowledge production, occupies this terrain. This line of response has the virtue of addressing the discharge index directly: it asks users not to delegate the work of justification, or to delegate it self-consciously rather than reflexively. Its limit is the limit Cohen (2023) identifies: the cognitive demands placed on individuals exceed what civic competence can reasonably require at scale. The pedagogical line is necessary but cannot bear the full weight of the response.

Underlying these three lines of resistance is a deeper theoretical question. If the plausibility regime is parasitic on the verification regime, what happens when its host is depleted? The plausibility regime requires, for its statistical inputs, a continuous stream of human-authored, verified, contested texts. If a substantial proportion of the texts it ingests in the next training cycle are themselves outputs of the plausibility regime, the parasitism becomes auto-cannibalistic. The recent literature on what computational researchers call "model collapse" addresses precisely this concern: models trained on increasingly synthetic data exhibit progressive degradation of output quality. From the present article's standpoint, model collapse is not just a technical pathology; it is the long-term consequence of a regime that has not solved the problem of how to reproduce the verification labor it depends on. The most consequential normative implication, in this reading, is not that the plausibility regime should be opposed in principle but that it cannot, as a matter of structural fact, sustain itself without the verification regime that it parasitizes. The defense of verification is not merely a moral preference; it is a condition of the very system that appears to be displacing it.

One personal observation may be permitted at the close of this section. The author's own work in researching this article involved repeated tests of generative AI as a research aid: outputs were checked against original sources, citations verified through DOI resolution, and a non-trivial proportion of fluent claims discovered to be unsupported on inspection. The researcher who attempts to use these tools rigorously rapidly arrives at a counterintuitive conclusion: in some respects the verification labor required to use generative AI carefully exceeds the labor required not to use it at all. This is, of course, not a generalizable empirical finding, but it is consistent with the structural argument advanced here. The plausibility regime offers savings on the surface that, on closer examination, appear as costs distributed to whoever does the verification work that the regime itself does not do.

CONCLUSION

The aim of this article was to articulate a single concept and a corresponding measurement frame for what is here described as a regime change in the conditions of epistemic acceptance under generative artificial intelligence. The plausibility regime, as developed across the preceding sections, is characterized by four properties: a fluency-based criterion of acceptance, an interface-based site of authority, the discharge of justificatory work to opaque computational systems, and an inverted default disposition in which acceptance precedes rather than follows verification. Three operational indicators — the discharge index, the fluency-trust coupling, and the source-opacity ratio — together provide a measurement scaffold for empirical research that would test, refine, or challenge the diagnostic.

The first hypothesis, that the regime shift is structural rather than incidental, finds substantial support in the synthesized evidence. The displacement of justification by fluency, of author by interface, and of argument by saturation are not transient effects of immature technology; they are stable consequences of the architecture of large language models and the platform infrastructures within which they are deployed. The hallucination rates documented by Macdonald and colleagues (2024) — between 28.6 and 91.4 percent depending on the model — are not faults to be patched in the next iteration; they are constitutive features of systems that produce statistically plausible outputs without referential grounding. The fluency-trust coupling and the discharge of verification documented in the automation-bias literature (Alon-Barkat & Busuioc, 2023) point in the same direction.

The second hypothesis, that the regime shift is asymmetric across domains, is also supported, and the asymmetry has important political and theoretical implications. In domains with robust verification infrastructure — peer review, court adjudication, audited clinical evidence — the plausibility regime has not displaced the verification regime, although it has begun to test the boundaries. In informal, high-throughput, low-audit settings — open web search, student work, lay decision making — the displacement is well advanced. The implication is that the regime shift is not technologically inevitable; it is institutionally negotiated. Where institutions of verification are sustained, the regime holds; where they are depleted or absent, it is overrun. This is consistent with Chambers's (2023) argument that the fragmentation of the public sphere is a political problem requiring political and institutional response, not a technological problem requiring only technological remedy.

The third hypothesis, that the regime shift is empirically tractable through the proposed indicators, is the most exposed to revision. The discharge index, fluency-trust coupling, and source-opacity ratio are introduced as a conceptual scaffold for empirical work; the work itself remains to be done. Initial empirical estimates are available from existing studies, but a properly designed measurement program — likely combining survey instruments, behavioral experiments, and corpus analysis — would be required to test the indicators rigorously. The article does not claim to have completed such a program; it claims to have provided a coherent target for one.

The principal original contribution of this article is the formulation of the plausibility regime as a single, articulable concept and the operationalization of three indicators that allow the regime shift to be identified empirically. The concept does not displace existing analyses — of hallucination, of automation bias, of the structural transformation of the public sphere, of postmodern performativity — but integrates them under a frame that names what specifically has changed in the criterion of epistemic acceptance. The contribution is not a

synthesis in the sense of summary; it is a synthesis in the sense of producing a new analytical object whose purpose is to make the existing literature mutually legible. Whether this object proves productive is a question that subsequent empirical work will answer.

Four limitations of the present analysis must be acknowledged candidly. First, the geographic and linguistic horizon of the empirical literature reviewed is heavily Anglophone-Western; the regime shift may take different forms in epistemic environments where verification infrastructures developed under different histories, particularly in the Global South and in postsocialist contexts. Second, the article integrates rather than tests; the empirical estimates offered for the three indicators are derivative from existing studies rather than the product of a primary measurement program. Third, the conceptual move from Habermas, Lyotard, and Floridi to the plausibility regime is a reconstructive synthesis, and readers committed to any one of these traditions may legitimately object that the move flattens distinctions internal to that tradition. Fourth, the analysis treats the plausibility regime as a structural condition of late capitalist information environments without engaging the political economy of model training and deployment in the depth that political economy would require — a depth left for future work in which model ownership, training-data provenance, and the labor conditions of human reinforcement learners would receive sustained attention.

Recommendations for further research follow from these limitations. A primary measurement program testing the three indicators across institutional settings — with attention to variation by professional domain, educational level, and verification infrastructure — would constitute the most direct extension of this work. Comparative studies across linguistic and political contexts could test the hypothesis that the regime shift is institutionally rather than technologically determined. A political economy of generative AI, addressing the upstream conditions under which the plausibility regime is produced and distributed, would supply the materialist counterpart to the conceptual analysis offered here. Finally, the model-collapse literature in computational research and the verification-regime literature in epistemology should be brought into systematic dialogue: the long-term sustainability of the regime described here is a joint question of these two literatures, and its answer will shape the normative agenda of the next decade.

The practical implications, while indirect, are not negligible. Educational policies that respond to generative AI by either prohibition or uncritical adoption miss the structural diagnostic and address only its surface. Regulatory frameworks that secure provenance metadata and a substantive right to explanation address the source-opacity ratio but do not, by themselves, restore the discharge index. Civic and pedagogical efforts that cultivate epistemic literacy address the discharge index but cannot, by themselves, reach institutional scale. A coherent response would combine all three, anchored in the recognition that the verification regime is not merely a normative preference but a condition of possibility for the plausibility regime that increasingly displaces it. To defend verification is, in the present situation, not nostalgia; it is the structural maintenance of the very infrastructure that makes plausibility possible.

BIBLIOGRAPHY

Alon-Barkat, S., & Busuioc, M. (2023). Human–AI interactions in public sector decision making: 'Automation bias' and 'selective adherence' to algorithmic advice. *Journal of Public Administration Research and Theory*, *33*(1), 153–169. https://doi.org/10.1093/jopart/muac007

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 (FAccT '21)*, 610–623. https://doi.org/10.1145/3442188.3445922

Chambers, S. (2023). Deliberative democracy and the digital public sphere: Asymmetrical fragmentation as a political not a technological problem. *Constellations*, *30*(1), 61–68. https://doi.org/10.1111/1467-8675.12662

Cohen, J. (2023). Democratic responsibility in the digital public sphere. *Constellations*, *30*(1), 92–100. https://doi.org/10.1111/1467-8675.12670

Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology*, *36*(1), 15. https://doi.org/10.1007/s13347-023-00621-y

Habermas, J. (2022). Reflections and hypotheses on a further structural transformation of the political public sphere. *Theory, Culture & Society*, *39*(4), 145–171. https://doi.org/10.1177/02632764221112341

Macdonald, C., Adeloye, D., Sheikh, A., & Rudan, I. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, *26*, e53164. https://doi.org/10.2196/53164

Seeliger, M., & Sevignani, S. (2022). A new structural transformation of the public sphere? An introduction. *Theory, Culture & Society*, *39*(4), 3–16. https://doi.org/10.1177/02632764221109439

Simons, M. (2022). Jean-François Lyotard and postmodern technoscience. *Philosophy & Technology*, *35*(2), 31. https://doi.org/10.1007/s13347-022-00517-3

Staab, P., & Thiel, T. (2022). Social media and the digital structural transformation of the public sphere. *Theory, Culture & Society*, *39*(4), 129–143. https://doi.org/10.1177/02632764221103527

Symons, J., & Alvarado, R. (2022). Epistemic injustice and data science technologies. *Synthese*, *200*(2), 87. https://doi.org/10.1007/s11229-022-03631-z

Thiel, T. (2023). A polarizing multiverse? Assessing Habermas' digital update of his public sphere theory. *Constellations*, *30*(1), 69–77. https://doi.org/10.1111/1467-8675.12667

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2), 841–887. https://doi.org/10.2139/ssrn.3063289

ALGORITAMSKA RACIONALNOST I KRIZA EPISTEMIČKOG AUTORITETA: KA KONCEPTU REŽIMA PLAUZIBILNOSTI

Veronica Postolache

Moldavski državni univerzitet
Chisinau, Republic of Moldova
E-mail: veronica.postolache@student.usm.md

Primljeno: 02.04.2025. Odobreno: 27.05.2025.

Originalni naučni članak

DOI: https://doi.org/10.65932/CR-2025-1-5
UDC: 165.0:316.462]:004.8

Sažetak: Članak obrađuje strukturni pomak u uslovima epistemičkog prihvatanja pod generativnom vještačkom inteligencijom. Centralna teza glasi da se istiskivanje ljudskog suda velikim jezičkim modelima ne može adekvatno opisati jezikom dezinformacija ili "halucinacija", jer obje pretpostavke uzimaju zdravo za gotovo da osnovni režim verifikacije ostaje netaknut a samo se ponekad krši. Članak iznosi drugačiju dijagnozu: u toku je promjena režima, u kojoj se operativni kriterij epistemičkog prihvatanja pomjera s opravdanja zasnovanog na dokazu i provjerljivim izvorima ka statističkoj tečnosti i konverzacijskoj prihvatljivosti. Originalni doprinos članka leži u razvoju jednog koncepta — režima plauzibilnosti — zajedno s tri operativna indikatora koji omogućavaju da se pomak režima identifikuje empirijski. Režim plauzibilnosti definisan je s četiri osobine: kriterij prihvatanja zasnovan na tečnosti, sjedište autoriteta zasnovano na sučelju, izmještanje opravdavalačkog rada na opake računarske sisteme, i obrnuta podrazumijevana dispozicija u kojoj prihvatanje prethodi verifikaciji. Tri indikatora su: indeks izmještanja, sprega tečnosti i povjerenja, te omjer izvorne netransparentnosti. Metodologija je konceptualno-integrativna, oslanja se na Habermasovu komunikativnu racionalnost, Floridijevu filozofiju informacije i Lyotardovu analizu postmoderne performativnosti, te se kombinuje sa sintetičkim čitanjem novih empirijskih nalaza o stopama halucinacije, automatizacijskoj pristrasnosti i platformski posredovanom javnom diskursu. Zaključak rada je da režim plauzibilnosti nije izvanjski, nego parazitski u odnosu na režim verifikacije, te da svaki normativni odgovor mora djelovati na nivou institucionalnog dizajna, a ne pojedinačne epistemičke vrline.

Ključne riječi: *generativna vještačka inteligencija, epistemički autoritet, režim plauzibilnosti, komunikativna racionalnost, filozofija informacije, veliki jezički modeli, postmoderna nauka, algoritamsko posredovanje.*