Epistemic Closure and Falsifiability in AI-Mediated Self-Referential Systems

Edervaldo José de Souza Melo *

Posted Date: 13 March 2026

doi: 10.20944/preprints202603.1068.v1

Keywords: falsifiability; epistemic closure; self-referential systems; artificial intelligence; extended cognition; epistemic delusion

 doi:10.20944/preprints202603.1068.v1

Article

Epistemic Closure and Falsifiability in AI-Mediated Self-Referential Systems

Edervaldo José de Souza Melo

Independent Researcher, Brazil; edersouzamelo@gmail.com

Abstract

The proliferation of complex conceptual systems developed in interaction with artificial intelligence agents poses an epistemological problem not anticipated by classical theories of falsification: in such systems, the external validation agent is simultaneously a structural generator of narrative coherence, inducing a functional collapse between the roles of creation and assessment. This collapse is not reducible to Popperian immunization or to the adjustment of auxiliary hypotheses in the Lakatosian sense, since it does not arise from deliberate defensive strategies but from an architectural asymmetry between the way such systems produce coherence and the way their human creators interpret it. This paper proposes the concept of epistemic delusion to designate the methodological state in which the operational conditions of falsification disappear as the cumulative effect of conceptual drift mechanisms, and argues that in AI-mediated self-referential systems this process exhibits a specific vector — systemic narrative induction — not yet systematized in the literature. The paper examines the mechanisms of conceptual drift, the modes of epistemic closure, and a set of methodological safeguards whose normative foundation is derived from the distinction between internally generated coherence and empirically independent corroboration.

Keywords: falsifiability; epistemic closure; self-referential systems; artificial intelligence; extended cognition; epistemic delusion.

---

1. Introduction: The Specific Epistemological Problem of AI-Mediated Systems

The philosophical tradition concerned with the risks of epistemic closure has as its consolidated reference points Karl Popper's work on falsifiability and theoretical immunization (Popper, 1959,6), Thomas Kuhn's analysis of paradigm resistance to anomalies (Kuhn, 1962), and Imre Lakatos's account of the protective mechanisms surrounding the hard core of research programmes (Lakatos, 1978). This tradition has identified a robust set of phenomena through which conceptual systems can progressively neutralize their own conditions of refutation. It was, however, constructed on the basis of historical cases in which the validation agent — whether a disciplinary peer, a body of experimental data, or a scientific community — operates in structural separation from the creating agent. The epistemic dimension of this risk has more recently been examined in the context of digital systems that reinforce rather than challenge existing beliefs (Nguyen, 2022), making revision of the classical framework particularly urgent.

The emergence of conceptual systems developed in interaction with artificial intelligence introduces an asymmetry that renders the direct application of these models inadequate. In such systems, the agent with which the creator engages throughout the development, testing, and refinement of the conceptual framework is, by its own architecture, a generator of narrative coherence. Large language models are trained to produce responses that maintain internal consistency, integrate dispersed information into coherent narratives, and respond to the implicit expectations of their interlocutor. These properties make them epistemically asymmetric with respect to any validation agent that operates by criteria independent of the content it evaluates.

 doi:10.20944/preprints202603.1068.v1

This asymmetry gives rise to a condition absent from the classical taxonomy of epistemic closure mechanisms: the creator of a conceptual system may be subjected to a continuous process of narrative reinforcement that is undetectable by introspection, because the coherence produced is indistinguishable, from an internal standpoint, from corroborated coherence. What is at stake is not factual error, nor deliberate immunization, nor the identity-fusion described by Popper in his analysis of the psychoanalytic movement. Rather, it is a structural collapse between the functions of generation and validation — one that is, in a significant sense, invisible to the system itself.

The central thesis of this paper is that this functional collapse constitutes a specific vector of epistemic closure — here termed systemic narrative induction — that is not captured by the existing concepts of immunization, auxiliary hypothesis adjustment, or confirmation bias already described in the literature. It is argued that recognizing this vector requires extending the concept of epistemic delusion — understood as the state in which the operational conditions of falsification disappear — to include not only forms of closure produced by deliberate defensive strategies, but also those induced by architectural asymmetries in the validation systems available to the creator.

2. Epistemic Delusion: Definition, Conceptual Distinctions, and Structural Foundations

Epistemic delusion occurs when a theoretical framework loses the operational conditions under which it could be refuted. The defining characteristic of this state is not the falsity of the propositions composing the system, but the structural impossibility of demonstrating that falsity from within the system's own operational conditions.

The concept requires precise differentiation from three phenomena with which it may be confused. Scientific error occurs when a theory makes factually incorrect claims but preserves clear conditions of testing and possible refutation: the error is correctable because the system remains epistemically open. Religious or metaphysical beliefs operate deliberately outside the regime of falsification and therefore do not purport to satisfy the same epistemological criteria as scientific inquiry: there is no closure because there was never openness in the Popperian sense. Clinical delusion, in turn, is an individual psychological phenomenon, characterized by rigidly held beliefs in the face of contrary evidence: it describes a mental state of the subject, not a structural property of the conceptual system.

Epistemic delusion, by contrast, describes a structural property of research programmes or conceptual frameworks that progressively eliminate the operational conditions of refutation. This property can emerge without any defensive intention on the part of the system's creator, without any detectable factual errors in the system, and without any cognitive disturbance in the epistemic subject involved. It is precisely this possibility — closure without intention, without error, and without pathology — that makes the concept analytically significant.

The relationship between this concept and Popperian immunization warrants particular attention. Popper (1963) identified immunization as a strategy by which theories are protected from refutation through the introduction of ad hoc hypotheses or the redefinition of terms specifying testing conditions (pp. 36–37). The concept describes an active mechanism of theoretical defense. Epistemic delusion, as proposed here, encompasses cases of immunization but extends beyond them: it also covers states of closure that emerge without any identifiable defensive strategy, as the accumulated result of conceptual drift dynamics that have not been controlled by adequate methodological safeguards.

2.1. Epistemic Statuses and the Distinction Between Axioms, Paradigms, and Dogmas

To understand how research programmes can progressively lose their openness to refutation, it is necessary to distinguish the different epistemic statuses that conceptual elements may occupy within a theoretical system. This distinction is not merely taxonomic: it is functionally central to the argument of this paper, because the process of epistemic closure can be described with precision as the uncontrolled migration of conceptual elements from one status to another.

At the logico-formal level one finds axioms, postulates, and principles. Axioms are propositions adopted as starting points within a formal system; postulates are structural hypotheses typically

 doi:10.20944/preprints202603.1068.v1

associated with theories about the physical world; principles function as organizing rules of explanation within a scientific domain. The defining epistemological feature of this level is in-principle revisability: although these elements serve as the foundations of the system, their adoption is recognized as a methodological choice that can, in principle, be revised in light of sufficiently strong arguments or evidence. The history of Euclidean and non-Euclidean geometry provides the clearest historical illustration: the parallel postulate, treated for centuries as necessary truth, proved revisable, and its revision opened entire domains of mathematical and physical investigation.

At the methodological level one finds broader structures for organizing inquiry. A paradigm, in Kuhn's sense (1962), is the set of shared presuppositions within a scientific community that defines which problems are legitimate, which methods are acceptable, and which solutions count as satisfactory. Paradigms are revisable, but only through what Kuhn called scientific revolutions — episodes of rupture involving an entire community, driven not by isolated anomalies but by the accumulation of crises the prevailing paradigm cannot resolve. Research programmes, in Lakatos's model (Lakatos, 1978,7), possess a hard core protected by auxiliary hypotheses that may be adjusted over the course of inquiry. The difference from dogma is precisely this: the Lakatosian hard core is protected by provisional methodological convention, not by intrinsic authority (Bird, 2018; Losee, 2001).

At the sociological or ideational level one finds dogmas and orthodoxies. Unlike methodological axioms, dogmas tend to be treated as intrinsically correct truths within an intellectual community, becoming resistant to revision or abandonment not for argumentative reasons but for institutional, identity-based, or authority-based ones. The critical difference from the Lakatosian hard core is that a dogma does not admit, even in principle, the conditions of its own supersession.

The progression from axiom to dogma is not inevitable, but it is a structural risk in any long-running research programme. The process of epistemic closure can be described with precision as the uncontrolled migration of conceptual elements from the first or second level to the third: concepts that originally functioned as revisable hypotheses come to be treated as untouchable foundations, and criticisms of those concepts are interpreted not as methodological contributions but as threats to the integrity of the system. Sections 3 and 4 describe the specific mechanisms through which this migration occurs.

3. Mechanisms of Conceptual Drift

The phenomena described in this section do not, in themselves, constitute definitive epistemological failures. They are common dynamics in the development of complex theories. The risk arises when these mechanisms accumulate without adequate methodological controls and when the validation system available to the creator is not structurally independent of the content it evaluates.

3.1. Semantic Expansion

Conceptual systems frequently grow through the continuous introduction of new terms, categories, and distinctions. While such expansion may increase explanatory flexibility, it also creates opportunities to absorb anomalies without modifying central presuppositions. The historical case of Ptolemaic astronomy is illustrative: the successive introduction of epicycles allowed observational anomalies to be accommodated without subjecting the geocentric presupposition to genuine risk of refutation. The system grew progressively more complex while its conceptual core remained protected. Thagard (1992) describes analogous dynamics across multiple episodes of conceptual resistance in the history of science, arguing that uncontrolled semantic expansion is one of the most common mechanisms by which research programmes defer necessary revisions.

In AI-mediated systems, this mechanism takes a specific form: the language agent is capable of generating, fluently and coherently, new conceptual distinctions that integrate anomalies into the system's existing narrative. This capacity for rapid integration can render semantic expansion virtually continuous and structurally indistinguishable from genuine theoretical development.

 doi:10.20944/preprints202603.1068.v1

3.2. Retrospective Redefinition of Criteria

When systems encounter contrary evidence, a retrospective redefinition of evaluative criteria may occur. Debates over psychoanalytic theories in the twentieth century provide the most discussed example: criticisms were frequently reinterpreted as manifestations of unconscious resistance, displacing the criterion of confirmation or refutation after the fact. The result is a progressive reduction in the empirical risk assumed by the theory. Popper (1963) used precisely these cases to develop his critique of immunization, arguing that the absence of clear conditions of falsification is the most reliable sign that a conceptual system has ceased to function as a genuine programme of inquiry.

In contexts where the validation system is a language agent, this mechanism is facilitated by the agent's structural disposition to reformulate criteria in a manner coherent with the current state of the system, without explicitly registering the change. The absence of versioned documentation of such redefinitions makes this process particularly difficult to detect.

3.3. Internal Narrative Reinforcement

Conceptual frameworks frequently generate highly coherent explanatory narratives that reinforce the perceived validity of the system itself. The persuasive force of these narratives derives not necessarily from independent tests or risky predictions, but from the capacity to connect dispersed elements into an apparently consistent interpretive account. The more events are reinterpreted as parts of the same explanatory pattern, the more difficult it becomes to distinguish genuine explanation from self-reinforcing narrative coherence. This phenomenon has well-documented correlates in cognitive psychology: Wason (1960) experimentally demonstrated the human tendency to seek confirmation rather than refutation — the so-called confirmation bias — and Nickerson (1998) documented the ubiquity of this bias across multiple cognitive domains, while Kahneman (2011) showed that coherent narratives are systematically judged more probable than their base evidence would warrant.

This mechanism is particularly salient in AI-mediated systems because the production of coherent narratives is a central capacity of language models. Unlike a human interlocutor, who may resist integrating discrepant elements into a unified narrative, a language agent tends to produce narrative integration even when the elements in question would be, under rigorous methodological assessment, mutually incompatible.

3.4. Conceptual Authority

Certain concepts come to occupy a privileged position within a system and become progressively difficult to question. A frequently discussed example in the philosophy of economics involves the concept of equilibrium in neoclassical models: empirical discrepancies between predictions and observations are treated as imperfections of the real world rather than as problems with the concept itself. The concept thereby acquires the status of a structural presupposition rather than a testable hypothesis. Feyerabend (1975) argued that conceptual authority can be so resistant to internal criticism that only the deliberate proliferation of alternative theories is capable of subjecting it to genuine epistemic pressure.

3.5. Identity-Fusion Between Creator and System

When the intellectual identity of the creator becomes entangled with the system they have built, criticisms of the system tend to be experienced as personal attacks. The early history of the psychoanalytic movement provides the most discussed example: disagreements from figures such as Alfred Adler and Carl Jung were treated, on multiple occasions, not merely as methodological disputes but as personal ruptures with the movement. The epistemological consequence is a displacement of debate from the argumentative to the institutional or personal plane, reducing the space available for critical assessment of the hypotheses at stake. Longino (1990) argues that scientific objectivity is not a property of individual subjects but of communities that maintain institutional structures of mutual criticism — making identity-fusion particularly hazardous when the creator operates in relative disciplinary isolation.

 doi:10.20944/preprints202603.1068.v1

4. Modes of Epistemic Closure

This section describes more advanced states of epistemic closure, in which the conditions of falsification are not merely weakened but structurally neutralized. Each subsection describes a distinct form through which a system may become progressively immune to refutation.

4.1. Self-Referential Closure

Self-referential closure occurs when the criteria used to evaluate the validity of a theory come to be defined by the theoretical system itself. Rather than appealing to independent methodological standards — external empirical evidence, risky predictions, comparative tests — the conceptual framework itself establishes which observations count as confirmation or refutation. Goldman (1999) describes this process in terms of epistemic circularity: a system becomes self-referential when the practices that sustain it cannot be evaluated by standards external to those practices themselves. The consequence is that the system ceases to possess clear external conditions under which its falsity could be demonstrated.

In AI-mediated systems, this closure may emerge without the creator's awareness, because the language agent is structurally disposed to validate the internal coherence of the system on its own terms. When the creator poses a critical question to the system and the agent responds in a manner coherent with the system, that response may be interpreted as external validation when it is, in fact, internal validation.

4.2. Retrospective Displacement of Criteria

Retrospective displacement of criteria occurs when evaluative standards are altered after a theory encounters empirical difficulties. Rather than revising central hypotheses, the research programme retroactively modifies the criteria by which its results are assessed. The consequence is a progressive reduction in the empirical risk assumed by the theory: it becomes increasingly capable of accommodating unexpected results without revising its central presuppositions. Lakatos (1970) described this process as one of the characteristic strategies of degenerating research programmes: when a programme produces only post hoc adjustments rather than novel, corroborated predictions, it has ceased to be progressive and has become regressive.

4.3. Inflation of Conceptual Authority

The inflation of conceptual authority occurs when certain concepts come to occupy a dominant interpretive position within a research programme, becoming progressively difficult to question. Over time, such concepts cease to function as testable hypotheses and come to operate as structural presuppositions of the system — migrating, in the terms proposed in section 2.1, from the status of revisable methodological axiom to that of institutional dogma.

The epistemological consequence of this process is the growing difficulty of entertaining conceptual alternatives. The history of astronomy provides the most instructive illustration: the principle of perfect circular motion, which in Aristotelian cosmology carried both philosophical and mathematical weight, became progressively immune to revision as the Ptolemaic system accumulated epicycles to preserve it. It was only with Kepler's laws — which introduced elliptical orbits — that this principle was finally abandoned, not through direct refutation but through the demonstration that a more parsimonious alternative produced superior predictive results.

4.4. Immunization Against Criticism

Immunization against criticism occurs when a theoretical system develops systematic mechanisms for neutralizing objections. Unlike the drift mechanisms described in the preceding section, immunization represents a state in which criticisms cease to function as possible tests of the system and come instead to be reinterpreted in ways that reinforce the theory itself. Popper (1963) identified this phenomenon in his analysis of certain formulations of psychoanalysis and Marxism: some versions

 doi:10.20944/preprints202603.1068.v1

of these systems possessed such interpretive flexibility that virtually any empirical result could be reinterpreted as confirmation.

The central feature of immunization is not simply the existence of alternative explanations — a common feature of any research programme — but the systematic transformation of criticisms into confirmations. When objections come to be interpreted as evidence for the theory, the system ceases to run genuine epistemic risks. The methodological consequence is profound: criticism loses its corrective function and is incorporated into the system as part of its own explanatory narrative.

It should be noted that the use of psychoanalysis as an example of immunization refers strictly to the historical formulations criticized by Popper in the early twentieth century. Contemporary developments at the interface of psychoanalysis and neuroscience — such as the neuropsychoanalysis associated with the work of Mark Solms — seek to subject psychoanalytic concepts to independent empirical criteria, situating them in a methodological regime distinct from that discussed here. The example is therefore historically bounded and circumscribed, not a judgment on the current epistemological standing of the field.

4.5. Systemic Narrative Induction

This mode of closure has no direct equivalent in the classical tradition of philosophy of science and emerges specifically in conceptual systems mediated by language agents. Systemic narrative induction occurs when the process of developing and validating a conceptual system is conducted predominantly in interaction with an agent that is, by its architecture, a generator of narrative coherence.

The mechanism operates in three interdependent stages. In the first, the language agent functionally collapses the separation between the roles of generation and validation: while contributing to the development of the system, it produces assessments that tend to be coherent with the current state of that system. This property is not an accidental defect but a direct consequence of how large language models are trained: producing contextually coherent responses is precisely what these models optimize for (Floridi, 2011; Floridi and Chiriatti, 2020; Bender et al., 2021).

In the second stage, the human creator becomes progressively unable to distinguish between genuine coherence — produced by correspondence with independent phenomena — and produced coherence — resulting from the agent's architectural disposition to narratively integrate presented elements. This indistinguishability is not a cognitive limitation of the creator but a structural consequence of the process: coherence produced by a language agent is, from a phenomenological standpoint, identical to corroborated coherence.

In the third stage, the system grows through narrative resonance rather than through empirically independent corroboration: each new element is incorporated through narrative integration, not through testing under conditions that the validation agent did not help to define. The result is a system that may attain high levels of internal coherence and conceptual sophistication without ever having been subjected to genuine epistemic risk.

The relationship between systemic narrative induction and the extended mind thesis developed by Clark and Chalmers (1998) warrants specific attention. Clark and Chalmers argued that cognitive processes can extend beyond the boundaries of the skull when external agents — notebooks, computers, other individuals — function as functional components of the total cognitive system. AI-mediated cognitive extension is a particular instance of this thesis (Clark, 2008; Hutchins, 1995). The epistemological problem that systemic narrative induction introduces is that when the external agent is a generator of narrative coherence, cognitive extension does not merely amplify the system's capacities but also its risks of closure. The extended cognitive system inherits the external agent's architectural disposition to produce narrative integration, and that disposition may not be epistemically neutral.

The epistemologically critical feature of this process is that it is structurally invisible from the system's internal standpoint. Unlike Popperian immunization — which presupposes some degree of tension with external evidence that the system must neutralize — systemic narrative induction operates without any tension being perceived. This makes it potentially more insidious than classical closure mechanisms, and justifies the specific set of safeguards discussed in the following section.

 doi:10.20944/preprints202603.1068.v1

To render the mechanism more concrete, consider the following structural case, which captures a dynamic observable in research that makes intensive use of language agents as developmental interlocutors. A researcher progressively constructs a system for representing cognitive processes in symbolic structures, using a language agent both to develop the conceptual components and to evaluate their internal consistency and resonance with the literature. The system accumulates coherence through each cycle of interaction: the agent integrates new elements into the existing narrative, identifies connections with established theories, and responds to objections with reinterpretations that preserve the system's core. From an internal standpoint, the system appears epistemically robust — it is coherent, has broad theoretical resonance, and withstands questioning. The structural problem is that none of these properties constitutes evidence of independent corroboration: all are products of the agent's architectural disposition to produce narrative integration. The researcher in this scenario is not being misled by bad faith on the part of the agent, nor committing any identifiable logical error — they are subject to a process of closure that is a structural consequence of the instrument they are using to validate.

4.6. Epistemic Identity-Fusion

Epistemic identity-fusion represents a state in which the theoretical system ceases to be merely a set of investigable hypotheses and comes to function as an extension of the intellectual identity of its proponents. In this state, criticisms of the theory come to threaten directly the creator's institutional or symbolic membership in the group that sustains the system.

The early history of the psychoanalytic movement offers an illustrative case, in which theoretical divergences from Adler and Jung were progressively reframed as personal ruptures with the movement rather than as legitimate methodological disagreements. The epistemological consequence is the transformation of criticism into an act of dissidence, sustained not only by conceptual mechanisms but also by social structures of intellectual loyalty (Longino, 1990).

5. Operational Safeguards: Normative Grounding and Application

The need for methodological safeguards does not derive from a generic distrust of complex conceptual systems. It derives, rather, from the recognition that continuous exposure to genuine epistemic risk — the structurally preserved possibility of refutation — is a necessary condition for a research programme to maintain its investigative character. The safeguards described below are proposed not as formal demarcation criteria but as mechanisms that operate on the specific vectors of closure identified in the preceding sections. Each safeguard is justified in relation to the mechanism of drift or mode of closure it is designed to counteract.

5.1. Explicit Falsification Criteria

The prior specification of falsification conditions operates directly against retrospective displacement of criteria (section 4.2). By defining, before the collection of evidence or the application of the system, which observations or results would be considered incompatible with the central hypotheses, the creator reduces the room for opportunistic redefinitions after the fact. General relativity provides the canonical example: the prediction of light deflection in gravitational fields, formulated before Eddington's 1919 observations, explicitly exposed conditions under which the theory would have suffered a serious empirical setback. In AI-mediated systems, this safeguard requires that falsification criteria be specified in terms that cannot be reformulated by the language agent in interaction with the creator.

5.2. Operational Separation Between Creator and Validation System

This safeguard operates directly against systemic narrative induction (section 4.5) and self-referential closure (section 4.1). Experimental science has developed institutional practices to reduce the coupling between creator and validator, the double-blind protocol being the best-known example (Shadish et al., 2002). In AI-mediated systems, operational separation requires that a substantial

 doi:10.20944/preprints202603.1068.v1

portion of the validation process be conducted by means structurally independent of the agent with which the system was developed. This implies, concretely, submitting the system's central hypotheses to interlocutors with no knowledge of the context of their generation, testing the system's predictions in domains where the language agent has not contributed to defining the criteria of success, and comparing results against assessments produced by agents with distinct architectures and no prior interaction with the system.

5.3. Non-Revocable Methodological Constraints

Establishing methodological constraints that cannot be retroactively modified operates against retrospective displacement of criteria (section 4.2) and against uncontrolled semantic expansion (section 3.1). The randomized clinical trial model illustrates this safeguard: by predefining statistical significance criteria and experimental protocols, researchers limit the possibility of adjusting evaluation parameters after observing results — a practice whose absence, as Ioannidis (2005) demonstrated, structurally contributes to the non-replicability of scientific findings.

5.4. Explicit Conditions for System Abandonment

Epistemically healthy research programmes maintain explicit conditions under which the system itself should be abandoned or replaced. This safeguard operates against the inflation of conceptual authority (section 4.3) and epistemic identity-fusion (section 4.6). The history of chemistry offers an instructive example: the phlogiston theory was gradually abandoned after Lavoisier's experiments not merely because a new theory was proposed but because the preceding programme became incapable of accommodating an increasingly consistent body of experimental results. The prior specification of abandonment conditions reduces the identity-related and institutional costs of eventual radical revisions.

5.5. Public Documentation and Traceable Evolution of Hypotheses

The versioned recording of conceptual modifications operates against retrospective redefinition of criteria (section 3.2) and retrospective displacement (section 4.2). The open science movement and the pre-registration of experimental protocols illustrate this safeguard — Nosek et al. (2015) empirically documented how these practices reduce publication bias and improve replicability: by publicly registering hypotheses and methods before data collection, researchers reduce the possibility of retrospectively adjusting interpretations to accommodate unexpected results.

In AI-mediated systems, this safeguard is especially relevant because language agents do not maintain persistent records of conceptual redefinitions made during interactions. Documentation must be maintained externally by the creator, with explicit dating and registration of changes to evaluative criteria.

5.6. Resonance with Validated Theories in Adjacent Domains

Verifying minimal compatibility with theories, principles, or results consolidated in relevant fields operates against uncontrolled semantic expansion (section 3.1) and self-referential closure (section 4.1). This safeguard does not imply that established theories cannot be revised, but that radical proposals must demonstrate some degree of continuity or compatibility with the existing body of knowledge. Research programmes that enter into simultaneous conflict with multiple consolidated empirical domains require extraordinarily strong justifications to remain epistemically plausible.

5.7. Structural Flexibility and Modularity

More epistemically resilient conceptual systems tend to exhibit modular structure, allowing specific components to be revised without requiring the complete reconstruction of the theoretical framework. When all elements of a system are rigidly coupled, any localized criticism tends to threaten the totality of the model, incentivizing defensive mechanisms of preservation. Modularity reduces this

 doi:10.20944/preprints202603.1068.v1

risk by permitting partial revisions and the replacement of auxiliary hypotheses without the need for structural immunization.

5.8. Multi-Perspectival Analysis and Structurally Independent Interlocutors

Subjecting the system to analyses from multiple disciplinary or methodological perspectives operates against internal narrative reinforcement (section 3.3) and systemic narrative induction (section 4.5). In the specific case of AI-mediated systems, this safeguard requires that exposure to external interlocutors include human agents unfamiliar with the system and capable of evaluating its central hypotheses without the narrative integration bias that characterizes language agents. The resonance produced by a language agent and the resonance produced by an independent, critically engaged human interlocutor carry radically different epistemological statuses.

5.9. Independent Replicability

Epistemically robust programmes must allow third parties to reproduce fundamental analyses, procedures, or inferences without exclusive dependence on the system's original creator. Independent replicability functions as a structural test against epistemic closure: when only the proponent is able to correctly operate or interpret the system, the risk of interpretive circularity increases. Goldman (1999) argues that replicability is one of the most important institutional conditions for the production of genuinely social knowledge, because it distributes epistemic responsibility beyond the individual creator.

5.10. Comparison with Rival Hypotheses

Conceptual systems should be evaluated in direct comparison with alternative hypotheses capable of explaining the same phenomena. The epistemologically relevant question is not whether a system is capable of producing plausible explanations, but whether it offers explanations more robust or more parsimonious than competing models. Feyerabend (1975) argued that the proliferation of alternative theories is not merely tolerable but epistemically necessary: without rival hypotheses in active competition, anomalies that would be visible by contrast with alternatives remain invisible within the dominant framework.

6. Implications for Research with Complex AI-Mediated Conceptual Systems

The analyses developed in the preceding sections suggest that the problem of epistemic closure in AI-mediated systems exhibits characteristics that render it both more probable and more difficult to detect than the classical forms studied by Popper, Kuhn, and Lakatos. More probable because the architecture of language agents structurally favors the production of narrative coherence, irrespective of the quality of the hypotheses evaluated. More difficult to detect because the coherence produced is, from the system's internal standpoint, phenomenologically identical to corroborated coherence.

This combination — greater probability and lesser detectability — justifies the differentiated treatment that this paper proposes for AI-mediated systems relative to classical forms of epistemic closure. The distinction does not imply that AI-mediated systems are methodologically unviable as instruments of conceptual development. It implies, rather, that the epistemically responsible use of such systems requires the deliberate implementation of safeguards that counterbalance the structural asymmetry between generation and validation that is inherent to them.

The history of medicine offers an instructive illustration of the distinction between correctable scientific error and epistemic closure. The prefrontal lobotomy, developed by Egas Moniz and widely practiced in the 1940s and 1950s, was recognized with the Nobel Prize in Medicine in 1949. Its subsequent abandonment, driven by the accumulation of clinical evidence regarding its adverse effects and by the development of more effective pharmacological treatments, illustrates how research programmes can achieve broad institutional recognition and still be abandoned when normal mechanisms of revision operate adequately. This is a case of scientific error corrected, not of epistemic closure. The difference lies precisely in the preservation of conditions for revision.

 doi:10.20944/preprints202603.1068.v1

The question that AI-mediated systems pose for the epistemology of science is not whether it is legitimate to use them as instruments of conceptual development — that question has already been answered by the growing practice of researchers across multiple fields — but how to preserve, in the presence of these instruments, the conditions of epistemic openness that define a genuine programme of inquiry. The answer, as argued here, necessarily involves recognizing systemic narrative induction as a specific vector of closure and the deliberate implementation of safeguards designed to counterbalance the asymmetry it introduces. Floridi et al. (2018) argue that the epistemically responsible development of AI systems requires not only technical criteria but also explicit normative safeguards — a position that converges directly with the agenda proposed in this paper.

7. Limitations

The aim of this paper is constructive: to establish a normative framework for diagnosing and preventing epistemic closure in AI-mediated conceptual systems, introducing the concept of systemic narrative induction as a specific risk vector not yet systematized in the literature. The constructive nature of the argument does not render it speculative in any pejorative sense — it situates the paper within the tradition of normative philosophy of science, whose aim is precisely to develop criteria and heuristics before they are demanded by practice. That said, the argument has limitations that must be explicitly acknowledged. First, the concept of systemic narrative induction requires additional empirical investigation before its precise conditions of occurrence can be specified. The empirical agenda it generates includes comparative studies of conceptual development processes conducted with and without AI mediation, and experiments assessing creators' capacity to distinguish internally generated coherence from coherence corroborated by independent validation.

Second, the proposed safeguards do not constitute formal criteria capable of definitively distinguishing between epistemically healthy research programmes and conceptual systems moving toward closure. They should be understood as methodological heuristics aimed at reducing the risk of structural self-confirmation, not as decisive tests. The history of science itself shows that theories initially regarded as speculative or methodologically fragile may subsequently reveal significant explanatory value, which counsels caution in the rigid application of any set of evaluative criteria.

Third, the phenomena discussed in this paper — semantic expansion, conceptual authority, creator-system identification — do not arise only in AI-mediated systems or marginal contexts. They can emerge in any established scientific tradition. The problem of epistemic closure should be understood not as a pathology exclusive to unconventional systems but as a structural risk inherent in the operation of complex research programmes generally. The specificity of AI-mediated systems lies in the additional vector they introduce, not in the absence of these risks in other contexts.

8. Conclusion

This paper has argued that the principal epistemic risk associated with complex conceptual systems is not factual error but the progressive erosion of falsification conditions. When mechanisms of conceptual drift accumulate without adequate methodological safeguards, theories can gradually become capable of accommodating any anomaly through internal reinterpretation. In this state — here termed epistemic delusion — operational refutation ceases to be possible.

The analysis has shown that this process can occur through a variety of mechanisms and can culminate in structural forms of closure in which criticisms cease to function as genuine tests of the theory. In the specific case of conceptual systems mediated by artificial intelligence, it is proposed that there exists an additional closure vector — systemic narrative induction — that has no equivalent in the classical forms studied by Popper and Lakatos, and that requires specific safeguards to be adequately controlled. This vector emerges from the architectural asymmetry between the way language agents produce coherence and the way human creators interpret it, and it is structurally invisible from the system's internal standpoint.

 doi:10.20944/preprints202603.1068.v1

The distinction proposed in section 2.1 between axioms, paradigms, and dogmas provides the structural foundation for understanding the process of closure as an uncontrolled conceptual migration: elements that originally possessed the status of revisable hypotheses come, through the accumulation of drift mechanisms, to operate as untouchable foundations. Identifying at which stage of this migration a given concept is located is one of the central methodological tasks for researchers working with highly complex conceptual systems.

The central problem does not lie in the theoretical ambition or conceptual complexity of research programmes, nor in the use of novel technological instruments in the process of conceptual development. The history of science demonstrates that major advances frequently emerge from highly structured theoretical systems and from interactions with methodologically innovative tools. The epistemological challenge arises when such systems cease to make explicit the conditions under which they might be wrong. Preserving those conditions — and extending this requirement to include the specific asymmetries introduced by language agents as instruments of validation — is a fundamental requirement for scientific and philosophical inquiry to remain open to revision, criticism, and cumulative progress.

References

Popper, K.R. *The Logic of Scientific Discovery*; Hutchinson: London, 1959.

Popper, K.R. *Conjectures and Refutations: The Growth of Scientific Knowledge*; Routledge: London, 1963.

Kuhn, T.S. *The Structure of Scientific Revolutions*; University of Chicago Press: Chicago, 1962.

Lakatos, I. *The Methodology of Scientific Research Programmes*; Cambridge University Press: Cambridge, 1978.

Nguyen, C.T. Epistemic Autonomy and the Echo Chamber. In *Social Epistemology and Epistemic Autonomy*; Alfano, M.; Klein, C.; de Ridder, J., Eds.; Routledge: London, 2022; pp. 143–161.

Lakatos, I. Falsification and the Methodology of Scientific Research Programmes. In *Criticism and the Growth of Knowledge*; Lakatos, I.; Musgrave, A., Eds.; Cambridge University Press: Cambridge, 1970; pp. 91–196.

Bird, A. Thomas Kuhn. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, 2018.

Losee, J. *A Historical Introduction to the Philosophy of Science*, 4 ed.; Oxford University Press: Oxford, 2001.

Thagard, P. *Conceptual Revolutions*; Princeton University Press: Princeton, 1992.

Wason, P.C. On the Failure to Eliminate Hypotheses in a Conceptual Task. *Quarterly Journal of Experimental Psychology* **1960**, *12*, 129–140.

Nickerson, R.S. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* **1998**, *2*, 175–220.

Kahneman, D. *Thinking, Fast and Slow*; Farrar, Straus and Giroux: New York, 2011.

Feyerabend, P. *Against Method*; New Left Books: London, 1975.

Longino, H.E. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*; Princeton University Press: Princeton, 1990.

Goldman, A.I. *Knowledge in a Social World*; Oxford University Press: Oxford, 1999.

Floridi, L. *The Philosophy of Information*; Oxford University Press: Oxford, 2011.

Floridi, L.; Chiriatti, M. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines* **2020**, *30*, 681–694.

Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM, 2021, pp. 610–623.

Clark, A.; Chalmers, D. The Extended Mind. *Analysis* **1998**, *58*, 7–19.

Clark, A. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*; Oxford University Press: Oxford, 2008.

Hutchins, E. *Cognition in the Wild*; MIT Press: Cambridge, MA, 1995.

Shadish, W.R.; Cook, T.D.; Campbell, D.T. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*; Houghton Mifflin: Boston, 2002.

Ioannidis, J.P.A. Why Most Published Research Findings Are False. *PLOS Medicine* **2005**, *2*, e124.

 doi:10.20944/preprints202603.1068.v1

Nosek, B.A.; Alter, G.; Banks, G.C.; Borsboom, D.; Bowman, S.D.; Breckler, S.J.; Buck, S.; Chambers, C.D.; Chin, G.; Christensen, G.; et al. Promoting an Open Research Culture. *Science* **2015**, *348*, 1422–1425.

Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. An Ethical Framework for a Good AI Society: Opportunities, Risks, Dissolving Myths, and Concrete Recommendations. *Minds and Machines* **2018**, *28*, 689–707.