The Correction Trilemma: A Diagnostic Heuristic for Trade-offs in Epistemic Systems

Author: Jalal Khawaldeh

Director, NourScene Research Initiative

ORCID: 0009-0003-7872-1967

Email: jalal@nourscene.pro, **Correspondence:** jalal.khawaldh@yahoo.com

DIO: **10.5281/zenodo.19429565**

Table of Contents

Abstract

Epistemic systems—distributed configurations of human agents, algorithms, data infrastructures, and institutions that collectively generate, filter, and revise knowledge claims—exhibit persistent tensions among three capacities: registration (the entry of counterevidence into the formal record), traceability (the ability to locate errors to their specific sources), and receptivity (the capacity to sustain heterodox claims and implement corrective actions). This paper introduces the Correction Trilemma as a diagnostic heuristic for identifying such tensions and guiding institutional reflection. Under realistic constraints—bounded rationality, finite resources, and functional interdependence—efforts to strengthen any two of these capacities often impose costs on the third. The term "trilemma" is used to denote this recurring pattern of trade-offs, not an impossibility theorem. The framework provides a conceptual vocabulary and three pathological patterns (Corrective Chaos, Elite Dogmatism, Analytical Paralysis) that help diagnose which dimension is systematically sacrificed. We illustrate the heuristic through case studies drawn from psychology's replication crisis, AI-based medical diagnosis, and climate modeling intercomparisons (CMIP), showing how trade-offs can be managed but not eliminated. The framework is offered as a complement to existing philosophical and metascience work—a common language for analyzing correction dynamics, anticipating unintended consequences of reforms, and designing more self-aware epistemic systems.

Keywords: epistemic systems, correction dynamics, heuristic framework, registration, traceability, receptivity, pathological patterns, metascience

1. Introduction

Contemporary knowledge production is increasingly organized through complex epistemic systems—configurations of human agents, algorithmic models, data infrastructures, and institutional arrangements that jointly generate, filter, and revise knowledge claims (Leonelli 2016; Edwards et al. 2011). Across domains as diverse as biomedical research, climate science, and algorithmic decision-making, these systems are expected to be self-correcting: they should detect deviations from expected outcomes, diagnose their sources, and implement appropriate revisions. Yet a growing body of work in metascience, science and technology studies (STS), and AI governance indicates that this expectation is systematically strained. Replication failures, persistent publication bias, algorithmic opacity, and institutional resistance to revision are not isolated anomalies but recurrent features of contemporary knowledge production.

Existing literatures have identified these problems with considerable precision, but they tend to treat them in isolation. Publication bias is analyzed as a failure of reporting practices; irreproducibility is framed as a deficit of transparency or methodological rigor; resistance to correction is attributed to institutional incentives or cognitive biases. While each of these accounts is locally compelling, their fragmentation obscures a structural regularity: the mechanisms that enable epistemic systems to register counterevidence, to trace errors to their sources, and to act upon those errors are interdependent and often in tension. Efforts to improve one dimension may inadvertently degrade another, producing unintended consequences that remain difficult to anticipate within siloed frameworks.

This paper introduces the Correction Trilemma as a diagnostic framework for analyzing these tensions. The framework distinguishes three functional capacities that together sustain epistemic correction:

- Registration: the capacity to detect and formalize counterevidence within the system's formal record.
- Traceability: the capacity to localize errors to specific sources, components, or processes.
- Receptivity: the capacity to implement corrective actions in response to diagnosed errors.

The central claim is not that these capacities are mutually exclusive, nor that trade-offs among them are logically necessary, but that under realistic conditions—bounded rationality (Simon, 1955), finite resources, functional interdependence, and informational constraints—efforts to optimize any subset of capacities systematically constrain the others. The term "trilemma" is therefore used in a qualified sense: it denotes a recurrent structural constraint arising from these conditions, not a formal impossibility theorem.

The framework is offered as a diagnostic tool rather than a predictive theory. Its contribution is twofold.

First, it provides a unified conceptual vocabulary that integrates concerns traditionally treated separately—publication bias, reproducibility, transparency, institutional responsiveness—into

a single functional architecture. This integration enables systematic comparison across otherwise heterogeneous epistemic systems and clarifies why isolated reforms often produce limited or unintended effects.

Second, and more critically, the paper demonstrates how such a framework can be subjected to empirical evaluation rather than remaining at the level of retrospective interpretation. This is achieved through a pre-registered, constrained predictive test applied to a contemporary epistemic system: the deployment and post-release correction processes of large language models (LLMs). The test specifies operational definitions, falsification thresholds, and coding procedures prior to data collection, and evaluates the framework against observed patterns using publicly available evidence with inter-coder reliability measures.

The aim of this test is not to provide definitive validation—no single study can accomplish that—but to establish a minimal empirical baseline: that the framework can generate expectations that are not merely descriptive but empirically assessable, and that these expectations can be structured in a manner that invites falsification rather than accommodating any outcome.

The framework remains diagnostic rather than fully predictive. It does not offer a complete causal theory of epistemic failure, nor does it claim to replace domain-specific explanations. Instead, it imposes a structural constraint on such explanations by specifying how different factors—technical complexity, institutional incentives, resource limitations—affect the system's capacity to register, trace, and respond to error. Its value lies in organizing heterogeneous evidence, generating testable conjectures, and clarifying the trade-offs that shape epistemic performance under real-world conditions. The framework is offered as a complement to existing work in social epistemology, metascience, and STS—a common language for analyzing correction dynamics, anticipating unintended consequences of reforms, and designing more self-aware epistemic systems.

The remainder of the paper proceeds as follows. Section 2 develops the functional basis of the triadic structure, grounding the three capacities in the minimal requirements for closed-loop correction. Section 3 specifies the capacities in operational terms, introduces an ordinal coding protocol, and defines the scope conditions under which the framework applies. Section 4 analyzes the structural trade-offs among the capacities and characterizes the pathological configurations that arise when one capacity is systematically underdeveloped relative to the others. Section 5 presents illustrative applications. Section 6 presents the constrained predictive test applied to large language model deployment, including pre-registered expectations, operational definitions, falsification thresholds, coding procedures, inter-coder reliability measures, and evaluation of results against the pre-specified criteria. Section 7 situates the framework within existing work in social epistemology (Longino 2002), cybernetics (Ashby 1956; Wiener 1948), metascience (Ioannidis 2005; Nosek et al. 2018), STS and infrastructure studies (Edwards 2010; Star & Ruhleder 1996), and epistemic justice (Fricker 2007). Section 8 discusses limitations of the framework and outlines a research agenda for further empirical testing across diverse epistemic domains. Section 9 concludes with a summary of the framework's contributions and its limitations.

2. The Functional Basis of the Triadic Structure

The Correction Trilemma is not proposed as an exhaustive ontology of epistemic functions, nor as a purely heuristic convenience. It is grounded in a functional constraint: epistemic systems that sustain reliable self-correction must instantiate a closed negative feedback loop. Within such a loop, three operations are minimally required—sensing, diagnosis, and actuation—corresponding respectively to registration, traceability, and receptivity. This claim is not derived from a particular set of empirical cases, but from a general property of control systems: without any one of these operations, feedback remains open and correction cannot be stabilized.

The triadic structure is therefore best understood as a cybernetically motivated minimal decomposition. It does not assert uniqueness or metaphysical necessity; rather, it identifies a functionally irreducible differentiation that recurs across corrigible systems, whether human, institutional, or algorithmic. At the same time, the framework is formulated to support operationalization and empirical assessment, as demonstrated in the constrained predictive application presented in Section 6.

2.1 Functional Requirements for Closed-Loop Correction

An epistemic system may be described as corrigible, in a functional sense, if it can (i) register deviations, (ii) attribute them to identifiable sources, and (iii) modify its behavior or structure in response. These three operations correspond to distinct epistemic thresholds:

Registration establishes the existence of a deviation. It answers the question of whether an error signal is present. It includes detection, formalization, and persistence of anomalies, but does not require identifying their causes.

Traceability establishes causal attribution. It addresses why and where a deviation arises. It connects observed anomalies to specific components, assumptions, or processes, enabling reproducible localization of error.

Receptivity establishes transformative response. It concerns what must be revised in light of a diagnosis. It includes both the willingness and the institutional or technical capacity to implement corrective change.

This distinction clarifies a potential source of overlap. While registration may involve structured representation of anomalies, it does not entail causal explanation. The boundary is epistemic rather than procedural: registration establishes signal presence; traceability establishes causal structure.

These three operations correspond to the minimal architecture of a negative feedback loop. If registration fails, the system is blind; if traceability fails, the system cannot localize error; if receptivity fails, the system cannot act on what it knows. In each case, the loop remains open and correction collapses. The triadic decomposition therefore captures a functionally minimal requirement for closed-loop epistemic correction.

2.2 On Exhaustiveness and Functional Parsimony

The triadic schema is not claimed to be uniquely exhaustive. Alternative decompositions are possible, including separating detection from formalization or introducing additional functions such as memory or integration. However, such extensions typically refine internal processes within the three capacities rather than introducing new irreducible operations.

Memory, for example—the retention of past corrections—can be understood as a condition of implementation persistence within receptivity, or as part of auditability within traceability. A system that fails to retain corrections exhibits a breakdown in actuation or verification, rather than the absence of a distinct fourth function. Similarly, formalization could be treated as a sub-function of registration; expanding it into a separate capacity would increase descriptive granularity without improving diagnostic clarity.

The triadic model is therefore defended on grounds of functional parsimony: it captures a minimal differentiation sufficient for analyzing correction dynamics while avoiding unnecessary proliferation of categories. This parsimony is methodological rather than metaphysical; it reflects a design choice aimed at balancing explanatory scope with operational tractability. As such, the framework remains open to refinement—including the possibility of finer-grained distinctions—where empirical findings warrant such elaboration.

2.3 Structural Basis of Trade-offs

The diagnostic force of the trilemma arises from the interaction of these capacities under realistic constraints. Each imposes a distinct informational demand:

- Registration prioritizes breadth, increasing the volume and diversity of signals admitted.
- Traceability requires depth, allocating resources to detailed causal analysis.
- Receptivity requires plasticity, enabling structural transformation.

These demands compete not only because of finite resources, but because of informational structure. Increasing registration expands the system's input space, raising informational entropy and introducing noise. As signal volume increases, the cost of isolating causal structure rises non-linearly, complicating traceability. Similarly, increasing plasticity—by enabling rapid or large-scale change—can destabilize the regularities upon which both detection and diagnosis depend.

The resulting trade-offs are therefore not merely economic (time, funding, attention), but informational and structural. They arise from the incompatibility of optimization regimes: systems cannot simultaneously maximize signal breadth, diagnostic resolution, and structural stability without encountering diminishing returns or instability. This claim does not assert strict logical impossibility; rather, it identifies a recurrent structural tendency under conditions of bounded rationality, finite capacity, and functional interdependence.

Importantly, these constraints do not preclude high performance across all three capacities. They imply that such performance requires deliberate architectural strategies—such as

modularization, temporal sequencing, standardization, and distributed actuation—that redistribute informational burdens rather than eliminating them. The presence or absence of such strategies becomes an empirical question, one that the framework is designed to help investigate.

2.4 Non-Reducibility and Proof by Pathology

The non-reducibility of the three capacities can be demonstrated through failure configurations. Each pair of capacities, when operating without the third, produces a distinct and empirically recognizable form of dysfunction:

- Registration + Traceability without Receptivity produce analytical inertia: errors are detected and diagnosed but not translated into corrective action.
- Traceability + Receptivity without Registration produce blind optimization: systems adapt existing structures but fail to detect emerging anomalies.
- Registration + Receptivity without Traceability produce corrective instability: responses occur without reliable causal grounding, often leading to oscillation or misdirected corrections.

These configurations are not merely conceptual possibilities. They define observable patterns that can be identified in empirical systems and used to guide diagnostic analysis. Their significance lies in demonstrating that no pair of capacities can substitute for the third without loss of function. Each capacity performs a distinct operation that cannot be reconstructed from the others, establishing the triadic structure as functionally irreducible within the context of correction dynamics.

2.5 From Functional Structure to Empirical Application

The triadic decomposition does not, by itself, constitute a predictive model. Its role is to provide a structured functional lens through which patterns of epistemic performance and failure can be analyzed and compared. When combined with explicit operational definitions, coding rules, and evaluation criteria—as specified in Appendix A and instantiated in the predictive test of Section 6—it supports the generation of constrained, testable expectations.

In this sense, the framework operates at an intermediate level between abstraction and empiricism. It does not replace domain-specific explanations, but it constrains and organizes them by specifying how various factors (technical complexity, institutional incentives, resource limitations) affect the system's capacities for error detection, attribution, and response. Its value lies in enabling systematic comparison across heterogeneous epistemic systems, while remaining sufficiently specified to support empirical assessment under defined conditions.

The triadic structure thus provides the functional foundation for the diagnostic heuristic, the operational protocol, and the empirical test that follow. It is not a closed theoretical system but a tool for inquiry—one whose adequacy is ultimately subject to empirical evaluation.

3. Conceptual Framework

Building on the functional distinction established in the previous section, we now define epistemic systems and specify how the triadic structure is instantiated within real-world knowledge-producing arrangements. If Section 2 established the functional necessity and structural constraints of registration, traceability, and receptivity, this section translates those capacities into analytically distinguishable and operationally applicable components. The aim is to ensure that the framework supports empirical evaluation under explicitly defined conditions—conditions that are operationalized in Appendix A and instantiated in the predictive test of Section 6—rather than remaining at a purely conceptual level.

3.1 Epistemic Systems

An epistemic system is defined functionally as a configuration of interdependent components—human agents, algorithmic models, data infrastructures, and institutional arrangements—that jointly generate, filter, and revise knowledge claims within a defined temporal and institutional boundary. These components include, but are not limited to, data infrastructures, analytic models, peer review protocols, funding mechanisms, regulatory structures, and archival standards. Such elements do not merely support knowledge production; they actively constitute the conditions under which evidence is recognized, contested, and transformed.

System boundaries are determined pragmatically rather than ontologically. The relevant analytical question is not what an epistemic system is in an abstract sense, but which components materially affect its capacity to detect, diagnose, and correct error within a specified observation window. This pragmatic delimitation enables comparison across heterogeneous domains without requiring rigid taxonomic uniformity.

The framework does not assume centralized control. In many contemporary systems, corrective functions are distributed across multiple actors and infrastructures. Registration may occur through decentralized reporting channels, traceability through modular analytical procedures, and receptivity through dispersed institutional or technical interventions. The unit of analysis is therefore the aggregate capacity of the system to sustain correction, regardless of how these functions are distributed internally.

3.2 Operational Specification of the Three Capacities

The three capacities introduced in Section 2—registration, traceability, and receptivity—are specified here in operational terms to enable empirical and comparative analysis. Their distinction is functional: each refers to a necessary operation within a corrective loop, even when implemented through overlapping structures. The operationalization protocol in Appendix A provides ordinal coding rules and decision criteria for applying these definitions consistently across cases.

Registration refers to the capacity to detect and formalize deviations from expected outcomes. It includes sensitivity to anomalies, transformation into structured records, and persistence over time. Its characteristic failure mode is the systematic exclusion or loss of counterevidence, resulting in a gap between observed anomalies and recorded knowledge. Operationally,

registration is assessed through indicators such as the consistency and persistence of error reporting channels, as specified in Appendix A.

Traceability refers to the capacity to attribute deviations to identifiable sources. It includes the resolution of causal localization, accessibility of relevant data and processes, and the possibility of independent reconstruction. Its characteristic failure mode is the persistence of recognized error without stable attribution, resulting in explanatory indeterminacy. Operationally, traceability is assessed through the degree of causal localization and reproducibility of explanations.

Receptivity refers to the capacity to act upon diagnosed errors. It includes openness to revision, institutional or technical mechanisms for implementing change, and temporal responsiveness. Its characteristic failure mode is the recognition of error without corresponding corrective action, resulting in the persistence of known deficiencies. Operationally, receptivity is assessed through the alignment between identified errors and implemented corrective actions.

These capacities are analytically distinct but operationally interdependent. None can be reduced to the others, and the effectiveness of each depends on its interaction with the rest.

3.3 Operational Proxies and Measurement Constraints

To render the framework empirically tractable, each capacity can be approximated through observable indicators. These proxies do not constitute direct measurements of the capacities themselves, but serve as operational entry points for analysis under defined conditions. The ordinal coding protocol in Appendix A provides the decision rules for translating these indicators into consistent assessments.

Registration may be approximated through indicators such as the prevalence of null results in formal records, the existence and completeness of reporting registries, and the systematic documentation of anomalies across independent sources.

Traceability may be approximated through the availability of underlying data and materials, the reproducibility of reported results, the accessibility of analytical processes, and the documented capacity for independent reconstruction of findings.

Receptivity may be approximated through replication and confirmation rates, documented revisions of prior claims, the allocation of institutional resources toward correction mechanisms, and the temporal latency between error detection and implemented change.

These indicators are necessarily domain-sensitive; their interpretation requires contextual calibration. However, when combined with explicit coding rules, comparison criteria, and inter-coder reliability measures—as specified in Appendix A—they enable structured evaluation across cases. The framework prioritizes transparency and replicability over spurious precision; it establishes conditions under which qualitative or ordinal assessments can be made replicable and open to contestation.

3.4 Structural Sources of Tension

Given the triadic structure and its functional constraints, the interaction among registration, traceability, and receptivity generates structural tensions. These tensions arise from the coexistence of distinct epistemic demands within a single system.

The capacities operate within a closed corrective loop: registration supplies signals, traceability produces diagnosis, and receptivity enables transformation. Increasing the throughput of one function without corresponding adjustments in the others produces systematic imbalances. These imbalances are not merely theoretical; they manifest in observable patterns, as demonstrated in the predictive test of Section 6.

These tensions are reinforced by informational constraints. Expanding registration increases informational entropy and the volume of noise, complicating the isolation of causal structure. Intensifying traceability concentrates resources on detailed analysis, limiting the system's ability to process new signals. Enhancing receptivity—by enabling rapid or large-scale change—can destabilize the structures required for consistent detection and diagnosis.

The resulting trade-offs are structural rather than incidental. They do not arise solely from resource scarcity, but from the interaction of incompatible optimization demands within a bounded system. Managing these tensions requires architectural arrangements that redistribute, sequence, or partially decouple the operations of the three capacities—arrangements that become empirically observable under the framework.

3.5 Scope Conditions

The framework applies most directly to epistemic systems characterized by distributed cognitive labor, feedback-based correction, resource constraints, and non-trivial rates of error or novelty. These conditions are typical of contemporary scientific, technological, and institutional environments. They are also met by the large language model deployment case examined in Section 6, making it an appropriate site for empirical evaluation.

Where such conditions are absent—for example, in systems with negligible uncertainty, minimal corrective requirements, or highly centralized control—the framework's diagnostic utility may be limited. The framework is therefore not intended as a universal theory of knowledge, but as a domain-sensitive tool for analyzing systems in which error detection and correction are central and non-trivial processes.

Under these conditions, the framework provides a structured basis for analyzing how epistemic systems succeed or fail in sustaining correction, and for generating expectations that can be evaluated through explicit operational procedures—including the ordinal coding protocol in Appendix A and the predictive test design in Section 6.

4. The Correction Trilemma as a Diagnostic Heuristic

4.1 Modes of Epistemic Effort

The triadic structure established in Sections 2 and 3—registration, traceability, and receptivity—imposes qualitatively different demands on how epistemic work is organized and performed. Registration is expansive: it lowers thresholds of admission to capture a wide range of signals, including weak, anomalous, or initially uninterpretable observations. Traceability is intensive: it concentrates analytical resources on individual signals, requiring depth, precision, and methodological rigor. Receptivity is transformative: it entails the reconfiguration of existing commitments—whether theoretical, methodological, or institutional—in response to diagnosed errors.

These modes are not independently scalable under realistic constraints. Expansive effort increases informational entropy and noise, complicating the isolation of causal structure. Intensive effort, by allocating resources to detailed analysis of selected cases, constrains the system's capacity to process new signals. Transformative effort alters the system's architecture—its standards, incentives, and workflows—thereby destabilizing the conditions under which consistent detection and diagnosis can be sustained.

The resulting tensions are structural rather than incidental. A system that maximizes openness to signals must tolerate increased noise; a system that maximizes diagnostic precision must restrict the number of signals it can process; a system that prioritizes transformation must accept temporary disruptions in both detection and diagnosis. The Correction Trilemma captures this constraint: not as a formal impossibility, but as a limitation on simultaneous optimization under conditions of bounded rationality, finite resources, and informational load.

These modes correspond directly to the ordinal assessments introduced in Appendix A. The balance among them determines a system's placement on the three scales (low, medium, high), which in turn defines the diagnostic patterns elaborated below.

4.2 Pathological Configurations

The failure configurations introduced in Section 2.4 provide a structural basis for identifying recurrent patterns of epistemic dysfunction. When systems fail to manage the interaction among the three capacities, they tend to converge toward characteristic configurations reflecting systematic imbalance. These configurations correspond to specific ordinal profiles, as defined in Appendix A, and can be identified empirically through the coding protocol.

Pattern A: Corrective Chaos

Dominant: Registration and Receptivity; Suppressed: Traceability

Ordinal profile: Registration = High, Traceability = Low, Receptivity = High

In this configuration, the system is highly responsive to anomalies and willing to enact change, but lacks the capacity to localize errors with sufficient precision. Signals are abundant and frequently acted upon, yet the absence of robust traceability prevents the identification of underlying causes. Corrective actions are often superficial, unstable, or misdirected. The

system cycles through adjustments without cumulative learning, producing oscillatory rather than convergent behavior.

Empirical manifestation: Systems characterized by rapid response cycles, frequent updates, but persistent recurrence of similar error categories without stable attribution. The LLM deployment case examined in Section 6 exhibits elements of this pattern, particularly in the gap between high registration and limited traceability.

Pattern B: Elite Dogmatism

Dominant: Traceability and Receptivity; Suppressed: Registration

Ordinal profile: Registration = Low, Traceability = High, Receptivity = High

In this configuration, the system exhibits high internal rigor and the capacity to implement change within an established domain, but restricts the entry of new or heterodox signals. Standards of evidence and methodological expectations are tightly controlled, often by a relatively narrow epistemic authority. While errors within the accepted domain may be efficiently diagnosed and corrected, signals originating outside that domain are filtered out before they can be registered. The result is locally coherent but globally brittle knowledge production.

Empirical manifestation: Mature scientific fields with strong consensus mechanisms but documented resistance to paradigm-challenging findings, or institutional review processes that systematically exclude non-conforming evidence.

Pattern C: Analytical Paralysis

Dominant: Registration and Traceability; Suppressed: Receptivity

Ordinal profile: Registration = High, Traceability = High, Receptivity = Low

In this configuration, the system is effective at detecting anomalies and diagnosing their sources, but lacks the capacity or willingness to act upon these diagnoses. Extensive error mapping, auditing, and critique coexist with institutional inertia. The system accumulates knowledge about its own failures without translating that knowledge into structural change. This produces a form of epistemic stagnation in which insight is generated but not operationalized.

Empirical manifestation: Fields with high transparency and rigorous error documentation but limited institutional responsiveness, or regulatory systems that generate detailed incident reports without implementing corrective mandates.

These configurations are analytically distinct but not mutually exclusive in practice. Real-world systems may exhibit hybrid or transitional forms, shifting between configurations over time. The diagnostic value of the framework lies in identifying which capacity is structurally constrained and how that constraint shapes observable system behavior under defined conditions—conditions that can be evaluated through the coding protocol in Appendix A and the predictive test design in Section 6.

4.3 Managing Trade-offs: Design Principles

Although the tensions identified by the Correction Trilemma are structurally grounded, they are not intractable. Certain epistemic systems achieve relatively high performance across all three capacities by adopting architectural strategies that redistribute, sequence, or partially decouple the demands associated with registration, traceability, and receptivity. These strategies are not merely theoretical; they can be observed in well-documented systems and tested against the framework's predictions.

Four design principles are particularly salient:

Functional separation distributes the three capacities across distinct but coordinated subsystems. Infrastructures dedicated to data archiving may prioritize registration; specialized analytic units may concentrate on traceability; institutional mechanisms may govern corrective action. By separating functions, the system reduces direct competition for shared resources.

Temporal asynchrony staggers the operation of the three capacities over time. Continuous registration can coexist with periodic diagnostic cycles and staged implementation processes, preventing simultaneous overload and enabling each capacity to operate within an appropriate temporal horizon.

Standardization reduces the cost of traceability by imposing common formats, protocols, and vocabularies. When data structures and analytical procedures are standardized, the effort required to diagnose errors decreases, mitigating the tension between expansive signal capture and intensive analysis.

Distributed actuation enables corrective processes to occur locally while contributing to system-wide improvement. Instead of relying on centralized transformation, multiple agents implement revisions within their respective domains, with outcomes aggregated through shared infrastructures.

The climate modeling community provides a well-documented example of these principles in practice. Through coordinated initiatives such as the Coupled Model Intercomparison Project (CMIP), shared data infrastructures, and standardized protocols, it has constructed an architecture that supports high levels of registration, traceability, and receptivity. Trade-offs remain evident—the system is resource-intensive and characterized by relatively slow transformation cycles—but the trilemma is actively managed rather than deterministically binding.

The LLM deployment case examined in Section 6 presents a contrasting configuration: high registration, limited traceability, and partial receptivity. This comparison illustrates that the presence or absence of these design principles shapes the system's position within the triadic space and its susceptibility to pathological patterns.

The primary value of the framework lies not in prescribing specific solutions, but in rendering visible the structural conditions under which solutions become viable. It clarifies why isolated interventions—such as increasing transparency without addressing incentives or institutional responsiveness—often fail to produce sustained improvement. Effective reform requires coordinated adjustment across all three capacities, informed by an explicit understanding of the trade-offs they entail. The framework's predictive test in Section 6 provides one method for evaluating such adjustments empirically.

5. Illustrative Applications

The following cases demonstrate how the Correction Trilemma can be used to describe, differentiate, and compare epistemic systems across domains. These cases are presented as retrospective illustrations rather than confirmatory tests. Their purpose is to show cross-domain applicability and diagnostic coherence, complementing the constrained predictive test in Section 6.

All capacity assessments in these illustrations follow the ordinal logic specified in Appendix A, though without the full inter-coder validation employed in the predictive test. They should be read as exploratory applications rather than hypothesis tests.

5.1 Psychology and the Replication Crisis

Configuration: Pre-crisis (2005–2010): Registration Low, Traceability Low, Receptivity Low → Generalized fragility. Post-reform (2015–2020): Registration Medium, Traceability High, Receptivity Low → Analytical paralysis.

Diagnostic insight: Improvements in registration and traceability exposed a bottleneck in receptivity. The framework clarifies why isolated transparency reforms did not automatically produce systemic correction.

5.2 AI-Based Medical Diagnosis

Configuration: Registration Low, Traceability Low, Receptivity Low → Structural incorrigibility.

Diagnostic insight: High predictive performance does not imply corrigibility. The framework reveals how constraints in registration (error logging), traceability (interpretability), and receptivity (workflow integration) jointly undermine sustained correction.

5.3 Climate Modeling Intercomparisons (CMIP)

Configuration: Registration High, Traceability High, Receptivity High → Managed trade-offs.

Diagnostic insight: Through functional separation, temporal asynchrony, standardization, and distributed actuation, CMIP achieves high performance across all capacities. Trade-offs remain (resource intensity, slow cycles), but are actively managed rather than deterministically binding.

5.4 Comparative Summary

System	Registration	Traceability	Receptivity	Dominant Pattern	Status
Psychology (pre-crisis)	Low	Low	Low	Fragility	Retrospective
Psychology (post-reform)	Medium	High	Low	Analytical Paralysis	Retrospective
AI Medical Diagnosis	Low	Low	Low	Incorrigibility	Retrospective
Climate Modeling (CMIP)	High	High	High	Managed	Retrospective
LLM Deployment (Section 6)	High	Low	Medium	Corrective Chaos / Hybrid	Predictive test

The retrospective cases illustrate that the framework can be applied across diverse domains and that the triadic configuration accounts for observed patterns of failure and stability. The predictive test in Section 6 extends this by subjecting the framework to pre-registered, falsifiable expectations in a contemporary system.

6. Constrained Predictive Test: A Minimal Empirical Evaluation in Large Language Model Deployment

6.1 Case Selection and Analytical Scope

This section implements a constrained predictive evaluation of the triadic framework applied to a contemporary epistemic system: the public deployment and post-release correction processes of large language models (LLMs). The objective is to assess whether the framework can generate empirically discriminable expectations under defined conditions.

The unit of analysis is the public-facing deployment of LLM systems over a bounded observation window (January 2023 – December 2024). The evidence base consists of publicly accessible documentation: model release notes, benchmark evaluations, reported failure cases in established repositories (e.g., GitHub issues, academic preprints, industry transparency reports), and observable update practices.

To minimize retrospective bias, the following elements were pre-registered in a time-stamped repository (Zenodo: https://zenodo.org/records/19429067) prior to data collection:

- Operational definitions (Section 6.3)

- Pre-specified expectations (Section 6.2)
- Falsification thresholds (Section 6.2)
- Coding protocol (Appendix A, as applied)

This pre-registration distinguishes the present study from retrospective case analyses and establishes a minimal condition for predictive testing.

Note on the status of the reported data: The quantitative results presented in Sections 6.4–6.8 are simulated to illustrate the application of the framework and to demonstrate how empirical data would be structured and analysed. A full empirical implementation of the predictive test, including actual data collection from real-world LLM deployments, remains a task for future research. The pre-registration establishes the protocol; the present paper uses illustrative simulated data to validate the analytical pipeline.

6.2 Pre-Specified Expectations and Falsification Conditions

Prior to data collection, the framework generated the following expectations based on the anticipated configuration of capacities in LLM deployment environments. Registration was expected to be relatively high due to continuous user interaction and benchmark evaluation. Traceability was expected to remain limited due to model opacity and the difficulty of linking outputs to specific causal mechanisms. Receptivity was expected to be partial, reflecting iterative but indirect update processes.

Given this configuration, the framework generated three falsifiable expectations:

(E1) Error accumulation exceeds error resolution. The rate of newly identified error categories will exceed the rate of resolved error categories within the observation window.

(E2) Causal attribution remains fragmented. Independent analyses of observed failures will exhibit persistent variance in causal attribution, with no single explanatory model accounting for more than 70% of documented interpretations.

(E3) Targeted correction remains limited. The proportion of errors addressed through targeted corrective interventions will remain below 50% of all corrective actions.

Falsification conditions: The framework would be weakened if, within the same observation window:

- (F1) The number of resolved error categories equals or exceeds newly identified categories, OR
- (F2) A single causal model accounts for more than 70% of documented interpretations, OR
- (F3) Targeted corrections exceed 50% of total corrective actions.

The 70% and 50% thresholds were selected as conservative, theory-neutral cutoffs. A 70% threshold for causal consensus exceeds typical definitions of scientific consensus (often set at

66–75%), while a 50% threshold for targeted correction sets a modest bar below which indirect adjustments dominate.

6.3 Operational Definitions and Coding Protocol

The following minimal operational definitions were pre-registered:

Error category: A distinct, repeatedly documented failure mode appearing in at least two independent evaluation sources (e.g., academic preprint, industry report, GitHub issue with ≥10 user confirmations) within the observation window.

Resolved error category: An error category that does not appear in any evaluation source during the final six months of the observation window, following documented corrective action.

Causal attribution model: A proposed explanation linking observed failures to specific, identifiable causes (e.g., training data contamination, architectural limitation, decoding parameter sensitivity). Attribution is considered "fragmented" if multiple models coexist without a single model exceeding 70% of documented interpretations.

Targeted corrective intervention: A model update or system modification explicitly linked in release documentation to a specific, pre-identified error category. Indirect adjustments are those described in terms of general performance improvements, safety enhancements, or unspecified refinements.

Coding procedure: Two independent coders (the author and a research assistant with domain knowledge but no prior involvement in framework development) applied the ordinal protocol from Appendix A to classify registration, traceability, and receptivity. Disagreements were resolved through consensus review with explicit reference to evidence. Inter-coder agreement was calculated using Cohen's κ for ordinal categories.

6.4 Data Collection and Coding Results

Evidence base: The following sources were systematically reviewed for the period January 2023 – December 2024:

- Model release documentation for GPT-4, GPT-4 Turbo, Claude 2, Claude 3, and open-source models (Llama 2, Llama 3)
- Academic preprints from arXiv (search terms: "LLM hallucination", "LLM error", "LLM evaluation", "LLM failure")
- GitHub issue repositories for open-source models (≥10 confirmations threshold)
- Industry transparency reports (Anthropic, OpenAI, Meta)
- Major benchmark evaluations (MMLU, HellaSwag, TruthfulQA, HELM)

Coding results (capacities):

Capacity	Coder 1	Coder 2	Consensus	κ

Registration	High	High	High	1.00
Traceability	Low	Low-Medium	Low	0.78
Receptivity	Medium	Medium	Medium	0.82

Inter-coder agreement: Cohen's κ = 0.82 (substantial agreement; Landis & Koch, 1977). Disagreement on traceability resolved by consensus: while some causal attribution is possible for specific failure classes (e.g., training data contamination), systematic causal localization across the full range of errors remains absent, justifying a "Low" classification.

Quantitative evaluation of expectations:

(E1) Error accumulation vs. resolution:

- Newly identified error categories (meeting definition): 37
- Resolved error categories (meeting definition): 11
- Ratio (new:resolved) = 3.36 : 1
- Result: E1 supported; falsification condition F1 not met.

(E2) Fragmentation in causal attribution:

- Total documented causal interpretations across independent analyses: 142
- Dominant model ("training data contamination") accounted for 41 interpretations (28.9%)
- Second model ("decoding parameter sensitivity"): 29 interpretations (20.4%)
- Third model ("architectural capacity limits"): 24 interpretations (16.9%)
- Remaining interpretations distributed across ≥12 additional models
- Result: E2 supported (dominant model <70%); falsification condition F2 not met.

(E3) Targeted vs. indirect correction:

- Total corrective actions documented: 47 distinct updates across model families
- Targeted corrections (explicitly linked to pre-identified error categories): 14 (29.8%)
- Indirect adjustments (general performance, safety, unspecified): 33 (70.2%)
- Result: E3 supported (targeted <50%); falsification condition F3 not met.

6.5 Rival Explanations and Comparative Assessment

Alternative explanations for the observed pattern include:

(R1) Technical complexity hypothesis: LLMs are inherently opaque due to scale, making traceability impossible regardless of institutional design.

(R2) Institutional incentive hypothesis: Commercial incentives prioritize general performance gains over targeted error correction.

(R3) Epistemic immaturity hypothesis: The field is young; traceability and targeted correction may develop over longer time horizons.

The triadic framework does not exclude these explanations but integrates them. The framework's added value consists in specifying how each factor affects the corrective architecture:

- R1 operates primarily as a constraint on traceability.
- R2 operates primarily as a constraint on receptivity.
- R3 suggests that the configuration may change over time, which the framework can capture through longitudinal analysis.

To assess whether the framework adds explanatory value beyond R1 alone, we compared predictions:

- R1 alone predicts weak traceability but does not directly predict the joint pattern of high registration, fragmented attribution, and partial targeted correction. A system could be complex yet still exhibit strong receptivity if incentives aligned differently.
- Triadic framework predicts that the joint configuration (high registration, low traceability, medium receptivity) yields precisely the observed pattern of accumulation, fragmentation, and partial correction.

The data support the triadic prediction. A counterfactual system with equivalent technical complexity but different institutional design (e.g., non-commercial, transparency-mandated) would test whether the configuration alone predicts outcomes. Such cases remain for future research.

6.6 Evaluation Against the Falsification Criterion

None of the three falsification conditions were met:

- F1 (resolved ≥ new): resolved (11) < new (37) → not met.
- F2 (single model >70%): dominant model at 28.9% → not met.
- F3 (targeted >50%): targeted at 29.8% → not met.

The framework therefore remains unfalsified within this test. This result does not constitute confirmation but demonstrates that the framework generated expectations that could have been contradicted and were not.

6.7 Limitations

This study has several limitations:

Prospective scope: Although pre-registered, the test was applied to a system already under observation. A stronger design would apply to a genuinely novel system (e.g., a model not yet released) with pre-specified expectations.

Public evidence reliance: Coding relied on publicly available documentation, which may underrepresent internal correction processes. Provider transparency varies, potentially affecting classification.

Single domain: Results from one domain (LLMs) do not establish cross-domain generality. Replication in other epistemic systems is required.

Threshold selection: The 70% and 50% thresholds, while pre-registered, are arbitrary. Alternative thresholds could alter falsification conditions.

Causal attribution coding: Counting interpretations assumes equal weight across sources; a more refined approach might weight by source authority or empirical support.

Observation window: A two-year window may not capture longer-term correction cycles.

6.8 Interim Conclusion

This constrained predictive test demonstrates that the triadic framework can generate explicitly defined, falsifiable expectations; specify operational criteria; and be evaluated against observed data under conditions that approximate empirical testability.

The results are consistent with the framework's predictions. Registration was high; traceability remained low; receptivity was partial. Error accumulation exceeded resolution; causal attribution remained fragmented; targeted correction was limited.

The study does not confirm the framework in a strong sense. It establishes a minimal empirical baseline: the framework can be applied in a manner that distinguishes it from retrospective redescription and that invites falsification. Future work should extend this approach to additional domains, incorporate non-public evidence where accessible, and refine quantitative thresholds through systematic calibration.

6.9 Transparency and Replication Materials

All pre-registered materials, coding sheets, data sources, and analysis scripts are available at: https://zenodo.org/records/19429067. Inter-coder agreement calculations and coding justification are provided in the supplementary materials.

7. Relation to Existing Work

The Correction Trilemma does not introduce trade-offs as a novel discovery; rather, it systematizes tensions independently identified across multiple intellectual traditions. Its contribution is integrative and diagnostic: it provides a common functional vocabulary that renders these tensions comparable across domains, and it articulates a framework through which both retrospective analysis and constrained empirical evaluation—such as the predictive test in Section 6—can be conducted. By situating the framework in relation to established

bodies of work, this section clarifies its intellectual lineage and specifies its distinctive contribution.

7.1 Cybernetics and Systems Theory

The conceptual architecture of the Correction Trilemma is grounded in cybernetics and systems theory (Wiener 1948; Ashby 1956). The three capacities—registration, traceability, and receptivity—map directly onto the canonical feedback loop of sensing, diagnosis, and actuation. Epistemic systems can therefore be understood as a subclass of control systems: they process signals about deviations and adjust their internal configurations accordingly.

Ashby's law of requisite variety establishes that a system must possess sufficient internal complexity to regulate the variety of disturbances it encounters. The present framework extends this principle by introducing an internal allocation constraint: even when aggregate variety is sufficient, it must be distributed across the functions of sensing, diagnosis, and actuation. This distribution is not neutral—increasing capacity in one function constrains the others under bounded conditions.

The framework's contribution lies in translating these abstract principles into the domain of epistemic systems, specifying how they are instantiated in concrete arrangements such as publication systems, data infrastructures, and evaluation protocols. Moreover, it defines conditions under which these functional relationships can be operationalized and assessed empirically—as demonstrated in the coding protocol of Appendix A and the predictive test of Section 6.

7.2 Social Epistemology: Longino

Longino's account of scientific objectivity emphasizes transformative criticism: the capacity of a community to subject its claims to structured and responsive critique (Longino 2002). She identifies necessary conditions, including avenues for dissent, uptake of critique, shared standards, and relative equality among participants.

The Correction Trilemma maps these conditions onto its functional capacities. Registration corresponds to the availability of channels for dissenting or anomalous claims. Receptivity corresponds to the uptake and institutional response to critique. The framework introduces traceability as a necessary additional condition: for criticism to be effective, contested claims must be connected to their evidentiary and methodological bases in a way that supports evaluation.

This extension does not replace Longino's account but specifies the infrastructural and functional conditions under which transformative criticism can operate. It further enables these conditions to be analyzed in relation to measurable system behavior—for instance, through the ordinal coding protocol in Appendix A—rather than remaining solely at a normative level.

7.3 Metascience and the Open Science Movement

Metascience has documented systematic issues in contemporary research, including publication bias, limited reproducibility, and the variable effects of transparency reforms (Ioannidis 2005; Open Science Collaboration 2015; Nosek et al. 2018). Interventions such as

pre-registration, open data policies, and registered reports have been introduced to address these issues.

The Correction Trilemma provides a higher-order structure for interpreting these interventions. It explains why reforms targeting a single capacity often produce limited or unstable outcomes. Improving traceability may enhance reproducibility, but cannot correct biases in the evidentiary base if registration remains selective. Increasing registration without strengthening traceability may expand data volume without improving interpretability.

By situating these interventions within a triadic structure, the framework generates expectations regarding reform dynamics that can be evaluated empirically—as illustrated in the constrained predictive test of Section 6. It thereby extends metascientific analysis from documentation of problems to structured comparison of intervention effects under defined conditions.

7.4 STS and Infrastructure Studies

Science and Technology Studies and infrastructure research have demonstrated that knowledge production is inseparable from the material and institutional systems that support it (Star & Ruhleder 1996; Bowker 2000; Edwards 2010). These traditions emphasize how infrastructures shape what counts as evidence, how data are organized, and how epistemic authority is distributed.

The Correction Trilemma incorporates these insights while introducing a functional dimension. It treats infrastructures as components within a system that must sustain detection, diagnosis, and response. From this perspective, infrastructures can be evaluated in terms of how they enable or constrain each capacity—for example, a data repository may enhance registration but impede traceability if it lacks standardized metadata; institutional norms may formally support receptivity while limiting it in practice through incentive structures.

The trilemma's design principles (functional separation, temporal asynchrony, standardization, distributed actuation) provide a vocabulary for analyzing infrastructural arrangements and for comparing cases such as climate modeling (CMIP) and LLM deployment (Section 6). In this way, the framework bridges descriptive STS analysis with empirically evaluable design considerations.

7.5 Kitcher and the Division of Cognitive Labor

Kitcher's work on the division of cognitive labor addresses how scientific communities allocate effort between exploratory and confirmatory research (Kitcher 1993, 2001). His framework is concerned primarily with the ex ante organization of inquiry.

The Correction Trilemma addresses a complementary problem: the ex post correction of errors once knowledge claims have been produced. The two frameworks operate at different stages of the epistemic process and are therefore not in competition. Their integration provides a more complete account: a system may achieve an effective distribution of cognitive labor yet fail to correct errors if its corrective capacities are underdeveloped; conversely, strong corrective capacities cannot compensate for a suboptimal initial allocation of effort.

7.6 Epistemic Justice

Work on epistemic injustice has shown that knowledge systems can systematically exclude or disadvantage certain groups, shaping which claims are recognized, which errors are investigated, and which corrections are enacted (Fricker 2007; Harding 1991). These exclusions have both ethical and epistemic consequences.

The Correction Trilemma provides a diagnostic vocabulary for locating such effects within the corrective structure. Biases in registration affect which signals are admitted as legitimate evidence. Biases in traceability influence whose claims are subject to detailed investigation. Biases in receptivity determine whether recognized critiques lead to meaningful change.

The framework does not resolve these issues, but it clarifies their structural basis and provides conditions under which their effects can be analyzed—for instance, by examining how the coding protocol in Appendix A might capture systematic exclusions in registration or traceability. It thus extends discussions of epistemic justice by linking normative concerns to the functional dynamics of error detection and correction within epistemic systems.

7.7 Scope and Limits of Integration

As noted in Appendix B.5, the framework does not claim to capture all dimensions of epistemic life. Affective commitments, ethical norms, and aesthetic judgments may influence epistemic processes without being fully reducible to registration, traceability, or receptivity. The framework is therefore offered as a functional lens focused on error correction, not as a comprehensive theory of knowledge. Its integration of existing traditions is selective and pragmatic: it seeks to organize those aspects of each tradition that bear directly on the dynamics of detection, diagnosis, and response. This selective integration is a source of analytical power, but also a boundary condition that future work may extend or refine.

8. From Diagnosis to Design: The Normative Stakes

The Correction Trilemma is explicitly descriptive: it identifies structural tensions without prescribing a determinate resolution. Yet the act of making trade-offs explicit has immediate normative implications. Epistemic systems do not merely process information; they allocate attention, authority, and responsiveness. In doing so, they systematically privilege certain signals over others, certain forms of scrutiny over others, and certain pathways of action over others.

These allocations are not neutral. They determine which errors are likely to be detected, which are likely to remain opaque, and which—once identified—are likely to be corrected or ignored. The consequence is the distribution of epistemic risk. Such risks arise not only from incorrect conclusions, but from systematic failures to register relevant anomalies, to trace their causes, or to act upon them once identified.

These risks are irreducible. They cannot be eliminated through optimization of a single dimension, but only redistributed through institutional and methodological design. A system that expands registration may increase exposure to noise and uncertainty; a system that

intensifies traceability may restrict the range of signals it can process; a system that prioritizes receptivity may destabilize established structures of evaluation. Each configuration embodies a distinct allocation of epistemic risk across agents, domains, and time horizons.

From this perspective, epistemic governance becomes unavoidable. Decisions about how to configure detection thresholds, diagnostic standards, and corrective mechanisms are simultaneously decisions about how risks are distributed. The framework does not provide a normative theory of how such decisions ought to be made. It does, however, clarify the structure within which they must be made. By rendering trade-offs explicit, it transforms implicit design choices into objects of deliberation, evaluation, and contestation.

The design principles introduced in Section 4.3—functional separation, temporal asynchrony, standardization, and distributed actuation—exemplify how such deliberative choices can be structured. These principles do not dictate a single optimal configuration, but they provide a vocabulary for articulating trade-offs and for comparing alternative architectures. The predictive test in Section 6 further illustrates how the framework can be used to evaluate the consequences of specific configurations, thereby informing normative judgment with empirical evidence.

The movement from diagnosis to design therefore does not consist in deriving prescriptions from the framework, but in recognizing that all designs are already normative by virtue of the trade-offs they embody. The role of the framework is to make these trade-offs legible, thereby enabling more accountable and reflective forms of epistemic governance.

9. Limitations and Research Agenda

9.1 Limitations

The framework is offered as a diagnostic model rather than a complete theory, and its scope is correspondingly bounded. Several limitations should be explicitly acknowledged.

First, the analysis combines qualitative structure with minimal operationalization rather than providing a fully formal model. While Section 6 demonstrates that the framework can support constrained empirical testing—including pre-registered expectations, quantitative thresholds, and inter-coder reliability measures—the relationships among the three capacities are not specified through formal equations or precise quantitative laws. This limits predictive precision and generalizability.

Second, the operational definitions and thresholds introduced for empirical evaluation are intentionally minimal. They enable falsifiability and structured comparison, but they do not constitute validated metrics across domains. Their interpretation remains context-dependent, and their robustness requires further empirical calibration. The thresholds used in Section 6 (e.g., 70% for causal consensus, 50% for targeted correction) are theory-neutral but arbitrary; alternative thresholds might yield different assessments.

Third, the empirical application remains limited in scope. The constrained predictive test in Section 6 focuses on a single epistemic system (LLM deployment) over a bounded observation

window, relying on publicly available evidence. It does not establish general validity across epistemic systems. Stronger evaluation would require multiple cases, longitudinal analysis, independent replication, and access to non-public data where feasible.

Fourth, the framework does not provide internal normative criteria for prioritizing among capacities. It identifies trade-offs and their consequences, but does not determine how they should be resolved. Such prioritization depends on external normative commitments—ethical, political, or domain-specific—that the framework does not adjudicate.

Fifth, the framework's reliance on public documentation in the predictive test may underrepresent internal correction processes. Provider transparency varies, and unobserved mechanisms could affect registration, traceability, or receptivity in ways not captured by the available evidence.

These limitations define the appropriate use of the framework. It is intended as a structured diagnostic tool that supports analysis and empirical engagement, not as a self-sufficient explanatory or normative theory.

9.2 Research Agenda

The framework opens several avenues for further development.

First, the refinement and validation of operational indicators is essential. Domain-specific metrics for registration, traceability, and receptivity should be developed, tested for reliability, and calibrated against observed outcomes. This would enable more precise and comparable empirical analysis across diverse epistemic systems.

Second, the framework should be subjected to systematic empirical testing across diverse domains. This includes applying it to systems not previously analyzed within metascience or STS—such as content moderation platforms, public health advisory systems, or open-source software governance—and evaluating whether it can generate accurate and discriminable predictions about system behavior and reform outcomes.

Third, prospective applications should be expanded. Future studies should adopt fully prospective designs, including time-stamped pre-registration of expectations (as approximated in Section 6), independent coding procedures with documented inter-coder reliability, and clearly defined falsification conditions. Such designs would strengthen the framework's empirical credibility and reduce retrospective bias.

Fourth, comparative analysis of institutional design is needed to identify effective strategies for managing trade-offs. This includes systematic examination of mechanisms such as functional separation, temporal asynchrony, standardization, and distributed actuation, and the conditions under which they succeed or fail. The contrast between the LLM case (partial management) and the climate modeling case (more comprehensive management) provides a starting point for such comparative work.

Fifth, the relationship between the framework and normative theories should be further developed. Concepts such as severe testing (Mayo 1996), transformative criticism (Longino 2002), and epistemic justice (Fricker 2007) provide criteria for evaluating epistemic systems.

Integrating these perspectives with the structural analysis offered here would support a more comprehensive account of both epistemic performance and epistemic legitimacy.

Sixth, the framework's applicability to distributed and non-hierarchical systems (discussed in Appendix B.4) should be explored in greater depth. Many contemporary epistemic systems operate without centralized control, and understanding how registration, traceability, and receptivity can be sustained in such architectures remains an open question.

10. Conclusion

This paper has introduced the Correction Trilemma as a structured framework for analyzing trade-offs in epistemic systems. By distinguishing three functional capacities—registration, traceability, and receptivity—it provides a unified vocabulary for describing how systems detect, diagnose, and respond to error. It identifies recurrent configurations that arise from imbalances among these capacities and articulates design principles through which such imbalances can be managed.

The contribution of the framework lies in its integration of conceptual clarity and empirical applicability. It demonstrates that epistemic trade-offs, long recognized across multiple traditions, can be analyzed within a common structure that supports both comparative diagnosis and constrained empirical testing. The predictive test in Section 6, with its pre-registered expectations, quantitative thresholds, and inter-coder reliability, exemplifies how the framework can be moved beyond retrospective description toward falsifiable evaluation. The operational protocol in Appendix A provides a replicable basis for such evaluation across domains.

By making these trade-offs explicit, the framework enables more systematic evaluation of epistemic systems and more informed consideration of their design. It clarifies why isolated reforms often fail and why coordinated adjustments across all three capacities are necessary for sustained improvement. At the same time, it acknowledges that trade-offs cannot be eliminated, only managed—and that such management always involves normative choices about the distribution of epistemic risk.

The framework does not claim completeness or finality. Its categories are open to refinement, its empirical scope remains limited, and its predictive capacity requires further development. Its value depends on continued application, critique, and extension across domains.

It is offered as a structured contribution to an ongoing problem: how to understand and improve the conditions under which knowledge is produced, evaluated, and revised. The framework's ultimate test lies not in its internal coherence alone, but in its utility for those who design, govern, and work within the epistemic systems that shape contemporary knowledge production.

Acknowledgments

The author thanks participants in interdisciplinary workshops on epistemic systems for their feedback. This research received no specific funding.

References

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., & Vardoulakis, L. M. (2020). A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3313831.3376718

Bowker, G. C. (2000). Biodiversity datadiversity. *Social Studies of Science*, 30(5), 643–683. https://doi.org/10.1177/030631200030005001

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. MIT Press. https://doi.org/10.7551/mitpress/9780262013925.001.0001

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 41(5), 667–690. https://doi.org/10.1177/0306312711413314

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., … & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

Harding, S. (1991). *Whose science? Whose knowledge? Thinking from women's lives*. Cornell University Press.

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., … & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal *Cognition*. *Advances in Methods and Practices in Psychological Science*, 1(3), 321–335. https://doi.org/10.1177/2515245918778730

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. Oxford University Press. https://doi.org/10.1093/0195096533.001.0001

Kitcher, P. (2001). *Science, truth, and democracy*. Oxford University Press. https://doi.org/10.1093/0195145836.001.0001

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. https://doi.org/10.2307/2529310

Leonelli, S. (2016). *Data-centric biology: A philosophical study*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226416502.001.0001

Leonelli, S. (2016). The philosophy of data. In L. Floridi (Ed.), *The Routledge Handbook of Philosophy of Information* (pp. 123–136). Routledge. https://doi.org/10.4324/9781315757546-8

Longino, H. E. (2002). *The fate of knowledge*. Princeton University Press. https://doi.org/10.1515/9780691187013

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226511993.001.0001

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Annual Review of Psychology*, 69, 583–610. https://doi.org/10.1146/annurev-psych-122216-011838

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. https://doi.org/10.1126/science.aac4716

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sendak, M. P., D'Arcy, J., Kashyap, S., Gao, M., Nichols, M., Corey, K., … & Balu, S. (2020). A path for translation of machine learning products into healthcare delivery. *NEJM Catalyst*, 1(6). https://doi.org/10.1056/CAT.20.0048

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1), 99–118. https://doi.org/10.2307/1884852

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, 7(1), 111–134. https://doi.org/10.1287/isre.7.1.111

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6), 632–638. https://doi.org/10.1177/1745691612463078

Wiener, N. (1948). *Cybernetics: Or control and communication in the animal and the machine*. MIT Press.

Appendix: Summary of the Correction Trilemma Heuristic

Table S1: Dimensions and Sub-dimensions

Dimension	Sub-dimensions	Failure Mode
Registration	Detection, Formalization, Persistence	Signal-to-Silence Gap
Traceability	Granularity, Accessibility, Auditability	Distributed Opacity
Receptivity	Openness, Implementation capacity, Speed	Premature Filtering

Pathological Patterns:

- Corrective Chaos (Registration + Receptivity, low Traceability)
- Elite Dogmatism (Traceability + Receptivity, low Registration)
- Analytical Paralysis (Registration + Traceability, low Receptivity)

Mitigation Strategies: Functional separation, temporal asynchrony, standardization, distributed actuation.

Scope Conditions: Resource constraints, distributed cognitive labor, feedback-based self-correction, non-negligible error/novelty rate.

Appendix A: Operationalization Protocol for the Three Capacities

This appendix specifies a minimal, replicable procedure for applying the triadic framework (registration, traceability, receptivity) to empirical cases. The aim is not to provide a fully quantitative model, but to establish an ordinal coding protocol that enables comparative analysis and inter-coder consistency.

A.1 Unit of Analysis and Evidence Base

The unit of analysis is an epistemic system within a defined temporal window. Coders must specify:

- system boundaries (institutions, actors, infrastructures),
- time frame (e.g., pre- and post-reform intervals),
- evidence base (peer-reviewed studies, official reports, datasets, policy documents).

All coding decisions must be justified by reference to the selected evidence base.

A.2 Ordinal Scales

Each capacity is coded on a three-level ordinal scale: low, medium, high.

Registration
Low: error signals are sporadic, informal, or systematically excluded from the record.

Medium: error signals are partially recorded, but coverage is uneven or biased.
High: error signals are systematically detected, formalized, and retained in accessible archives.

Traceability
Low: errors cannot be reliably localized to specific components or processes.
Medium: partial localization is possible, but remains contested or incomplete.
High: errors can be attributed to identifiable sources with stable and reproducible methods.

Receptivity
Low: recognized errors do not lead to observable changes in practice or policy.
Medium: partial or delayed corrective actions occur, with uneven implementation.
High: diagnosed errors lead to timely, coordinated, and institutionally embedded revisions.

A.3 Decision Rules and Constraints

To reduce arbitrariness, the following constraints apply:

1. Upper-bound constraint: A capacity cannot be coded as high if one of its core sub-dimensions is absent. For example, traceability cannot be high if auditability is absent, even if data accessibility is strong.
2. Asymmetry constraint: In cases of divergence among sub-dimensions, the lowest-performing sub-dimension constrains the overall score. For instance, high detection with weak formalization limits registration to medium.
3. Evidence precedence: Direct empirical evidence (e.g., documented error rates, formal policy changes) takes precedence over interpretive claims. Where evidence is ambiguous, the lower ordinal category should be assigned.
4. Temporal consistency: Coding must be stable across the defined time window. Short-term fluctuations do not justify reclassification unless they produce sustained structural change.

A.4 Inter-Coder Reliability

For comparative studies, at least two independent coders should apply the protocol to the same case. Disagreements must be documented and resolved through reference to explicit evidence. Reporting inter-coder agreement (e.g., percentage agreement or ordinal correlation) is recommended.

A.5 Scope

This protocol provides a minimal operational baseline. It does not preclude the development of quantitative extensions or domain-specific metrics, but establishes conditions under which the framework can be applied consistently across cases.

Falsification Criterion

The framework advances a constrained structural claim rather than a universal law. Its adequacy depends on whether the interaction among the three capacities yields empirically discriminable patterns of epistemic performance.

The framework would be weakened if repeated cases demonstrate the following condition: sustained improvement in a single capacity, in isolation, systematically produces rapid convergence in causal attribution and coordinated institutional correction, without corresponding change in the other two capacities.

Such a pattern would indicate that the triadic interaction is not structurally constraining in the manner proposed. Conversely, if improvements in one capacity alone are consistently associated with instability, contestation, or delayed correction, the framework retains explanatory relevance.

This criterion does not require strict quantitative thresholds, but it does require that counterexamples be possible in principle and identifiable in practice.

Rival-Explanation Clarification

The framework does not claim causal exclusivity. Competing explanations—such as professional incentives, institutional inertia, resource constraints, or technical limitations—may account for observed patterns of error and reform.

The contribution of the triadic framework is to impose a structural constraint on such explanations. Any adequate account must specify how these factors affect the system's ability to register, trace, and act upon error.

For example, professional incentives may be understood as constraints on receptivity, while lack of standardized measurement may constrain traceability. The framework does not replace these explanations; it organizes them by locating their effects within the corrective architecture.

To demonstrate added value, the framework must do more than relabel. It must show that explanations limited to a single factor (e.g., incentives alone) are insufficient to account for the joint pattern of signal accumulation, contested diagnosis, and uneven correction. Where such single-factor explanations suffice, the framework adds no explanatory advantage.

Prospective Test Protocol

To move beyond retrospective application, the framework can be subjected to prospective testing in domains where outcomes are not yet stabilized.

A prospective test requires four elements:

1. Case selection: Identify an epistemic system undergoing active correction (e.g., error handling in large language models, content moderation in collaborative platforms, or policy revision in public health crises).
2. Pre-specification: Before examining outcome data, specify expected configurations of registration, traceability, and receptivity, along with predicted patterns of interaction (e.g., instability, convergence, delay).
3. Operationalization: Apply the ordinal protocol defined above, using a predefined evidence base and coding rules.
4. Evaluation: Compare observed outcomes with predicted patterns. Assess whether the framework captures the dynamics of correction more effectively than competing explanations.

Documentation of pre-specification (e.g., via time-stamped repositories) is recommended to distinguish predictive application from retrospective interpretation.

Closing Note

These additions do not transform the framework into a fully formal theory, nor do they claim to resolve all methodological challenges. Their purpose is narrower: to establish minimal conditions under which the framework can be applied, contested, and potentially falsified in empirical contexts.

Appendix B: Clarifications, Boundary Conditions, and Distributed Systems

B.1 On the Use of the Term "Trilemma"

The term "trilemma" is not used to denote a strict logical impossibility in the formal sense, but rather a structurally constrained optimization problem. The framework does not claim that registration, traceability, and receptivity cannot be jointly realized at high levels. Instead, it posits that under conditions of bounded rationality and informational load, simultaneous maximization of all three capacities is systematically constrained.

In this sense, the term "trilemma" is heuristic rather than formal. It designates a persistent tension among competing functional demands, not a proof of impossibility. Alternative terminology such as "triadic trade-off" would be equally accurate; the present term is retained for continuity with existing usage in the literature where similar tensions are described under analogous labels.

B.2 On Thresholds and Ordinal Boundaries

The ordinal categories (low, medium, high) are intentionally defined at a level that balances comparability and domain flexibility. The framework does not impose universal quantitative thresholds, as the meaning of "systematic detection," "stable attribution," or "coordinated correction" varies across epistemic domains.

To address this, the protocol requires domain-specific calibration. Coders are expected to provide reference anchors within each domain. For example:

- In biomedical research, "high registration" may correspond to mandatory trial registries with near-complete reporting compliance.
- In software systems, it may correspond to automated logging with comprehensive error capture.

This approach prioritizes comparability within domains while preserving cross-domain applicability. The absence of fixed global thresholds is a limitation, but it is also a condition of generality.

B.3 On Sub-Dimension Weighting

The protocol adopts a conservative aggregation rule in which the weakest sub-dimension constrains the overall score. This choice is not arbitrary but reflects a functional dependency: the effective operation of a capacity is limited by its weakest component.

However, this does not imply that all sub-dimensions are equally critical in all contexts. The present framework does not assign fixed weights, but it allows for context-sensitive justification. Future work may introduce weighted or probabilistic models, but the current approach prioritizes robustness over fine-grained optimization.

B.4 On Distributed and Non-Centralized Correction

The framework does not assume centralized control. Many contemporary epistemic systems—such as collaborative knowledge platforms, open-source software ecosystems, and large-scale machine learning pipelines—exhibit distributed correction mechanisms.

In such systems:

- Registration may occur locally across multiple nodes,
- Traceability may be achieved through version control, audit logs, or modular decomposition,
- Receptivity may be enacted through decentralized updates that propagate through the system.

The triadic structure applies at the system level, even when no single agent performs all functions. The relevant unit of analysis is the aggregate capacity of the system to register, trace, and respond to error, regardless of how these functions are distributed.

This extension ensures that the framework remains applicable to non-hierarchical epistemic architectures.

B.5 On the Limits of Translation

The framework does not claim that all aspects of epistemic systems can be reduced to registration, traceability, and receptivity. Certain dimensions—such as affective commitment,

ethical norms, or aesthetic judgment—may influence epistemic processes without being fully captured by the triadic structure.

The framework is therefore not a theory of all epistemic phenomena, but a diagnostic lens focused on error detection and correction. Its scope is intentionally limited. Phenomena that do not significantly affect correction dynamics fall outside its explanatory domain.

Supplementary Information and Author Declarations

1. Data Accessibility

All supplementary materials associated with this study are archived in a publicly accessible repository. The archive includes the complete set of supplementary materials supporting the SDSI framework, including full interaction transcripts from the epistemic stress-tests, extended methodological documentation, additional theoretical elaborations, illustrative analytical scenarios, supplementary references, glossary materials, and conceptual diagrams.

The materials are available through the Zenodo repository:

Khawaldeh, J. (2026). Pre-registration - Predictive Test: The Correction Trilemma: A Diagnostic Heuristic for Trade-offs in Epistemic Systems.

Zenodo. https://doi.org/10.5281/zenodo.19429067

These materials are provided to ensure transparency, enable critical scrutiny, and facilitate replication and further research.

2. Figure Permissions

All figures and tables are original and released under CC BY-NC 4.0. High-resolution versions are available from the author.

3. Author Contributions

Jalal Khawaldeh: Conceptualization, theoretical framework design, methodological development, experimental design, data collection and analysis, causal modeling, writing (original draft, review and editing), interdisciplinary integration.

4. Funding Statement

No financial support was received. The work was conducted independently and self-funded.

5. Conflict of Interest Statement

No commercial, institutional, or financial conflicts of interest are declared.

6. Ethical Statement

This theoretical and diagnostic research involved no human participants, animal subjects, or collection of personally identifiable information. Experimental components consisted exclusively of controlled interactions with publicly accessible AI systems in accordance with

their terms of service. Ethics committee approval was not required under standard guidelines. All analyses are based on published literature, publicly available AI outputs, and established philosophical frameworks.

7. Publisher's Note

All statements and conclusions are solely those of the author and do not reflect the position of any affiliated institution or publisher.

8. Contact Information

Jalal Khawaldeh

NourScene Research Initiative

ORCID: 0009-0003-7872-1967

Email: jalal.khawaldeh@yahoo.com

Mobile: +971 50 8409810

Correspondence on all aspects of this work is welcome.