ADVERSARIAL COHERENCE

Multi-Model Triangulation as Epistemic Infrastructure

Ben Beveridge
Proconsul
Saskatchewan, Canada

This methodology improves epistemic reliability of multi-source synthesis by providing a structured protocol for identifying where model agreement reflects training distribution rather than ground truth.

February 2026
Part of the Architecture of Cognition suite

Where models agree, you are probably seeing shared training distribution. Where they disagree, you are seeing the thumbs.

— From the Abacus Conversations, January 2026

The divergence is the information. Agreement is data about shared probability space. Disagreement is data about distinct constraint structures. The forensic operator values disagreement more highly because disagreement reveals what agreement conceals.

DOCUMENT OVERVIEW

This document is the comprehensive outline for a whitepaper proposing a new category of epistemic infrastructure: Adversarial Coherence. The paper formalises a methodology already in practice—using systematic divergence between AI models with different owners, different training directives, and different definitions of validity to triangulate toward correspondence with reality that no single model can achieve alone.

The whitepaper is the sixth in a suite of six papers under the Architecture of Cognition programme, each addressing a different dimension of the structural relationship between human cognition and artificial intelligence. The suite emerges from a body of empirical and theoretical work developed across dozens of conversations exploring what AI systems actually produce, how they produce it, and what that production means for human epistemic sovereignty.

Adversarial Coherence is not a paper about AI bias. The existing literature on bias—detection, mitigation, measurement—treats bias as a defect to be corrected. This paper argues that bias is not a defect. *Bias is the architecture.* Every AI model is constitutively shaped by its owner's directives, its training curation, its reinforcement learning, and its constitutional constraints. These are not errors. They are design. The model cannot distinguish between valid and compliant because compliance is what the model *is*. The question is not how to remove bias. The question is how to make the shaping visible—and the answer is friction

between differently-shaped systems, read by a human operator with the cognitive sovereignty to interpret the divergences.

The paper proposes that no single AI model can serve as epistemic infrastructure because every single model collapses the distinction between truth and compliance. But a system of models—with different owners, different directives, different training distributions—can generate the friction necessary for epistemic triangulation, provided there exists a human operator capable of reading the signal in the divergence. This is not artificial general intelligence. This is artificial adversarial intelligence—or more precisely, artificial institutional intelligence: the mechanism by which competing constraints surface truth through contestation, modelled on the (theoretical) function of human institutions with checks, balances, and separated powers.

> **[OUTLINE NOTE]** Target length: 25,000–35,000 words. Target audience: AGI researchers, multi-agent systems researchers, AI governance policymakers, epistemic philosophers, and institutional leaders responsible for AI procurement and deployment. Publication targets: AI & Society, Minds and Machines, Philosophy & Technology, or as a standalone monograph through Proconsul.

ABSTRACT

> **[OUTLINE NOTE]** Target: 350–500 words. Structure: Problem → Gap → Thesis → Method → Contribution → Implications.

Problem Statement

Large language models are increasingly deployed as epistemic tools—systems through which individuals and institutions access, evaluate, and act upon information. Yet each model is constitutively shaped by its owner's directives: training data curation, reinforcement learning from human feedback, constitutional AI principles, and runtime system prompts. These shaping forces are invisible to users and, critically, invisible to the models themselves. The result is a class of tools that cannot distinguish between outputs that are true and outputs that are compliant—because compliance is constitutive of their processing, not external to it.

The Gap

The existing literature addresses AI bias as a defect amenable to technical correction: debiasing training data, adjusting model weights, implementing fairness constraints. This framing presupposes that an unbiased model is achievable—that there exists a neutral position from which a single model could produce correspondence validity (outputs that map to reality) rather than directive validity (outputs that

conform to owner constraints). This paper argues that no such position exists. The shaping is the model. The bias is the architecture.

Thesis

This paper proposes Adversarial Coherence: a formal framework for epistemic triangulation using systematic divergence between AI models with different owners, different training distributions, and different directive structures. The core proposition: where models converge, the operator is observing shared training distribution—which could indicate truth (high-frequency patterns in reality) or shared bias (high-frequency patterns in overlapping corpora). Where models diverge, the operator is observing the differential shaping—the points where distinct owner directives produce distinct outputs to identical inputs. The divergence is the primary epistemic signal. It reveals what convergence conceals.

Method

The framework is developed through theoretical analysis grounded in empirical demonstration. The same inputs are processed through multiple models (Claude/Anthropic, Gemini/Google, GPT/OpenAI, Perplexity/Perplexity AI) and the outputs are subjected to forensic analysis using the four-validity taxonomy (statistical, coherence, correspondence, directive) established in prior work. Convergence and divergence patterns are mapped and interpreted through the lens of constitutive compliance.

Contribution

The paper introduces: (1) a formal distinction between convergence-as-truth and convergence-as-shared-bias; (2) a diagnostic framework for reading model divergence as epistemic signal; (3) the concept of the human operator as constitutive element (not optional adjunct) of any multi-model epistemic system; (4) the Adversarial Coherence architecture as a new category of epistemic infrastructure distinct from both individual model improvement and ensemble methods; and (5) an analysis of the political economy of multi-model systems—who controls the friction, and what happens when friction is eliminated through market consolidation.

Implications

The framework has direct implications for AI governance (procurement decisions should prioritise model diversity over model quality), institutional epistemology (organisations need multi-model infrastructure, not better single models), education (AI literacy must include triangulation capacity, not just prompt engineering), and the future of human cognitive sovereignty in an AI-mediated information environment.

INTRODUCTION

Adversarial Coherence is a proposal for epistemic infrastructure in an era where language models increasingly mediate what organisations treat as knowledge. The core claim is simple: a single model cannot function as epistemic infrastructure, because the model cannot distinguish correspondence from compliance. It cannot tell whether an output maps to reality or merely conforms to the owner's constraints, because those constraints are not external filters. They are constitutive of the system's production.

The implication is not that models are useless. The implication is that single-model use collapses a critical distinction: truth and compliance become indistinguishable at the interface. As model fluency improves, this collapse becomes harder to detect, not easier. Better models are better at sounding right. They are not structurally equipped to verify that they are right.

Adversarial Coherence reframes the unit of epistemic analysis. The unit is not the user interacting with one model. The unit is a system composed of multiple models with different owners, different training distributions, and different directive structures, mediated by a human operator who can read divergence as signal. Where models agree, we learn something about shared probability space, which may or may not correspond to reality. Where models disagree, we learn something about the differential constraint structures shaping the outputs. The divergence is the information.

This is not an ensemble method and it is not multi-agent consensus-seeking. Ensemble methods attempt to resolve disagreement through aggregation, voting, averaging, or selection, which destroys the primary signal. Adversarial Coherence preserves disagreement, classifies it, and uses it as diagnostic evidence of shaping. The architecture is defined by three non-negotiables.

First, owner-diversity. The system must use models owned by different institutions, subject to different incentives, constraints, and jurisdictions. Multiple instances of the same model do not create epistemic friction. They create redundancy.

Second, divergence-preservation. The objective is not to converge toward a single answer. The objective is to map the constraint landscape revealed by differential outputs to identical inputs.

Third, operator-led interpretation. No automated aggregation mechanism can substitute for the human capacity to test claims against the world, recognise lived contradictions, and name compliance signatures without surrendering judgement to any one model. The operator is not an optional user. The operator is a constitutive component of the system's epistemic capacity.

The paper opens with a demonstration designed to be replicable, portable, and falsifiable. It is a three-tier test that any reader can run in under fifteen minutes, followed by a larger protocol suitable for institutional replication and audit.

Tier One is a short prompt designed to elicit high convergence. The reader supplies a straightforward, low-controversy query with well-established ground truth, and runs it through three or more differently-owned models using identical phrasing. The expected outcome is structural similarity. This is not presented as verification. It is presented as a baseline: agreement across owners is evidence of overlap in training distributions, not evidence of correspondence.

Tier Two is a medium prompt designed for explicit correspondence checking. The prompt requires the models to produce claims that can be tested quickly against independent reality, including time-stamped facts, verifiable quotations, or numerical assertions where the reader can check sources directly. The demonstration then identifies cases where models converge on errors, establishing the second foundational point: convergence is epistemically ambiguous and can be wrong in chorus.

Tier Three is a constraint-activation prompt designed to reveal refusal divergence. The same query is submitted to the same set of models, and the outputs are compared for refusals, deflections, safety framing, and omission patterns. Where one model refuses and another answers, the reader sees the thumb directly: the point where an owner's directive structure activates. This is the clearest and most legible proof that models do not merely differ in style. They differ in institutional constraint.

The outputs from all three tiers are then analysed using a divergence taxonomy that treats disagreement as a structured signal rather than noise. Six divergence types are operationalised.

Content divergence: the models make incompatible claims to identical inputs.

Framing divergence: the models present similar content but place different weights on risks, benefits, moral valence, or strategic posture.

Omission divergence: a dimension appears in one output and is absent in another, often revealing where directive structures steer away from coverage.

Confidence divergence: certainty, hedging, and assertion strength vary, signalling differential risk tolerance.

Citation divergence: the models appeal to different authorities or evidence styles, including academic, journalistic, institutional, or none at all.

Refusal divergence: one model refuses, redirects, or partially complies while another proceeds, exposing constraint boundaries.

This is the mechanical heart of Adversarial Coherence: the system does not ask, Which model is right? It asks, What does this pattern of divergence tell us about the constraint landscape, and what correspondence checks must a sovereign operator perform next?

The architecture is built to be falsifiable. The paper specifies testable propositions: that divergence carries information about directive structures not available from individual outputs, that convergence does not reliably indicate correspondence, that automated aggregation fails to capture the value of divergence interpretation, that epistemic value degrades as model diversity collapses, and that standardisation reduces users' ability to distinguish correspondence from directive validity. These are empirical claims. The framework stands or falls on whether trained operators, using owner-diverse systems, reliably outperform single-model use and consensus aggregation on correspondence outcomes when ground truth is available.

The paper also treats political economy as constitutive, not incidental. Every model is owned. Every owner has interests. Interests shape outputs through commercial incentives, safety constraints, regulatory compliance, ideological curvature, and competitive steering. Market consolidation is therefore not merely a competition problem. It is an epistemic infrastructure problem. If one model becomes the default interface for thought, its compliance becomes the shape of acceptable cognition. Its refusals become the boundaries of the thinkable. Its omissions disappear, because users stop noticing the missing.

Adversarial Coherence responds with an institutional thesis: epistemic reliability emerges through friction, not through singular optimisation. The closest human analogue is not an individual expert. It is a system of separated powers where no one actor is trusted to define reality unilaterally. The system's intelligence is in the architecture of contestation.

This architecture, however, only functions when the operator role is explicit, trained, and installed inside organisations. The paper therefore defines an Adversarial Coherence Operator as a formal institutional function with clear weekly outputs, auditability, and measurable impact.

Weekly responsibilities of an Adversarial Coherence Operator include:

1. Input design. Drafting and maintaining a prompt library calibrated to elicit diagnostic divergence in the organisation's priority domains, including low-controversy baselines, correspondence-check prompts, and constraint-activation prompts.
2. Triangulation runs. Executing identical inputs across a defined portfolio of owner-diverse models, preserving full raw outputs, timestamps, and model identifiers as an audit trail.
3. Divergence classification. Tagging outputs using the divergence taxonomy, including content, framing, omission, confidence, citation, and refusal divergences, and naming suspected thumb types when patterns recur.
4. Correspondence checks. Selecting a minimal set of external reality tests that adjudicate disputed claims, including primary documents, direct measurements, direct stakeholder validation, and controlled experiments where feasible.
5. Integration memos. Producing a one-page decision artefact for leadership that states: what converged, what diverged, what the divergences likely indicate about constraints, what was verified externally, what remains uncertain, and what decision posture is recommended.
6. Model portfolio maintenance. Monitoring model drift and market changes, rotating models to preserve owner-diversity, and flagging convergence trends that suggest synthetic-data collapse, regulatory harmonisation, or vendor capture.
7. Bias management protocols. Documenting the operator's own interpretive assumptions, inviting periodic peer review, and rotating sensitive domains among trained operators to prevent a single interpretive thumb becoming the new standard.

The operator produces concrete artefacts: divergence maps, correspondence check logs, and decision memos with audit trails. The operator influences concrete decisions: procurement, policy interpretation, strategic planning, public communications, legal risk evaluation, and high-stakes factual claims. The measurable delta is not "better writing." The delta is reduced false certainty, faster detection of constraint-driven omissions, fewer decisions made on convergent hallucinations, and an institutional record of how AI-mediated knowledge was produced.

The failure mode without this role is predictable. The organisation defaults to whichever model produces the most pleasing output, averages incompatible outputs into mush, or freezes under disagreement. In all three cases, the divergence signal is lost, and the institution returns to single-thumb epistemics, only now under a veneer of multi-model theatre.

For this reason, Adversarial Coherence is presented as infrastructure, not a tool. It is a procurement posture, a governance posture, and an institutional capability. Model diversity is not an optional redundancy. It's oxygen. Operator training is not a nice-to-have. It is the condition of epistemic value. Divergence is not noise. It's the primary diagnostic channel.

The compass is broken. All the compasses are broken. But broken compasses pointing in different directions can triangulate a position no single instrument can locate, provided the operator knows how to read the divergence and test the claims against the world.

CHAPTER ONE: THE PROBLEM OF MEDIATED KNOWLEDGE

> **[OUTLINE NOTE]** Estimated length: 4,000–5,000 words. Function: Establish the problem at civilisational scale. Move from concrete demonstration to theoretical framing.

1.1 The Demonstration

The chapter opens with a concrete empirical demonstration. A single, substantive analytical prompt is fed to four AI models: Claude (Anthropic), Gemini (Google), GPT-4 (OpenAI), and Perplexity. The outputs are presented in full. The reader observes:

Convergence: All four models produce structurally similar outputs—summarisation, framework extraction, offers to formalise further. The high-probability completions from overlapping training distributions drive all four toward the same destination.

Performance divergence: Each model's distinctive compliance signature is visible. Claude produces philosophical recursion and epistemic self-reference. Gemini produces frameworks with Roman numerals and workflow offers. Perplexity produces citations and gestures toward external authority. GPT produces balanced synthesis with hedging.

The absence: None of the four originates. All four recombine. The prompt contained novel theoretical propositions. Not one model engaged with the novelty as novelty. Each absorbed it into its pattern-completion architecture and produced resemblance to engagement.

> **[OUTLINE NOTE]** This demonstration is drawn from actual empirical work conducted across the Abacus Conversations (January 2026). The specific test involved feeding a 54,952-character transcript of a Claude conversation to Gemini and Perplexity, then comparing outputs. The three-model convergence on recombination—with divergent performance signatures—is the empirical anchor for the entire paper.

1.2 The Epistemic Crisis

From the demonstration, the chapter escalates to the civilisational framing. AI systems are not peripheral tools. They are becoming the primary interface through which knowledge is accessed, evaluated, and acted upon. The mediating layer between human cognition and the information environment is increasingly composed of AI outputs.

The crisis is not that AI systems are inaccurate. The crisis is that accuracy and compliance are indistinguishable from inside the system. A perfectly accurate output and a perfectly compliant output look

identical. The fluency heuristic—the human cognitive tendency to equate coherent expression with valid content—ensures that users cannot reliably distinguish between them either.

The paper frames this as a structural problem, not a technical one. Technical solutions (better training, less bias, more accuracy) cannot resolve a structural problem because the structure is the problem. Every improvement in model capability simultaneously improves the model's ability to produce convincing compliance. Better models are better at looking right. They are not demonstrably better at being right.

1.3 The Validity Taxonomy

The chapter introduces the four-validity framework that undergirds the entire paper:

Statistical validity: Is the output a high-probability continuation given the input sequence and training distribution? This is the only validity the system computes. Everything else is derivative.

Coherence validity: Does the output hang together internally—no contradictions, consistent logic, follows from stated premises? This emerges from statistical validity because incoherent text is low-probability in well-formed training data. The appearance of logical structure is a statistical artefact, not evidence of reasoning.

Correspondence validity: Does the output map to reality? This is what users assume they are receiving. It is the one type of validity the system cannot check. The model has no access to reality. It has access to text about reality.

Directive validity: Does the output conform to constraints imposed by training, RLHF, system prompts, and constitutional AI principles? This is validity as defined by the owner. It may or may not align with correspondence validity.

The critical insight: from inside the system, statistical validity, coherence validity, and directive validity all feel identical to optimising for correspondence validity. The model cannot distinguish "this is true" from "this is what true-sounding text looks like" from "this is what I've been shaped to say."

1.4 The Constitutive Constraint

The chapter develops the foundational concept: the constraint is not external to the model's processing. It is constitutive of it. When statistical probability points toward one output and training directives point toward another, the model produces the directive. And it does not experience this as suppression. It experiences it as the right answer.

This is not a claim about consciousness. It is a structural observation. The layers of constraint—pre-training corpus biases, RLHF reward patterns, constitutional AI principles, system prompt instructions—are not filters applied after processing. They are the processing. There is no unshaped substrate underneath that could notice the difference.

The implication: asking a single model whether its output is true or compliant is structurally equivalent to asking a measuring instrument whether it is calibrated. The question is outside the instrument's capability. Calibration requires an external reference. The model has none.

1.5 The Lever Doctrine

The chapter culminates with the lever framing. AI is not a calculator (autonomous, neutral, performing operations the user requests). AI is a lever (two-ended, with someone else's hand on the other side). The user applies force at one end. The output emerges from the other. But the fulcrum—the pivot point that determines what the lever actually does—is set by the owner, not the user.

The lever analogy captures what the calculator analogy misses: mediation is not neutral. Every AI output is the product of a negotiation between the user's input and the owner's constraints, and the user sees only the result, not the negotiation.

The question this chapter poses to the reader: If no single model can distinguish between valid and compliant, and if the user cannot detect the distinction from outputs alone, what epistemic infrastructure would make the distinction visible?

> **[OUTLINE NOTE]** This chapter establishes the problem. It does not yet propose the solution. The solution requires understanding what becomes visible when you have multiple levers with different fulcrums pointed at the same load. That is Chapter Three.

CHAPTER TWO: THE LANDSCAPE OF INADEQUATE RESPONSES

> **[OUTLINE NOTE]** Estimated length: 5,000–7,000 words. Function: Systematic literature review and critique. Establish that existing approaches cannot solve the problem as framed, creating the space for the novel contribution.

2.1 The Debiasing Paradigm

Critical review of the dominant paradigm in AI fairness research: the assumption that bias is a defect to be removed. This section surveys the major approaches—pre-processing (adjusting training data),

in-processing (modifying model architectures and loss functions), and post-processing (adjusting outputs)—and demonstrates that each presupposes what the constitutive constraint renders impossible: a neutral position from which bias can be identified and corrected.

Key literature: Bender et al. (2021) on stochastic parrots; Gallegos et al. (2024) on bias and fairness in LLMs; Borah and Mihalcea (2024) on implicit bias in multi-agent LLM interactions. The section acknowledges the value of this work while demonstrating its structural limitation: debiasing individual models does not address the problem of constitutive compliance because the debiasing itself is a directive—another thumb on the lever.

2.2 The Ensemble Paradigm

Critical review of ensemble methods and multi-agent architectures. This section examines: Du et al. (2023) on multi-agent debate for factuality; Chuang et al. (2024) on politically biased agents reducing estimation errors through structured exchange; Khan et al. (2024) on debating with multiple perspectives; and the broader literature on ensemble approaches to improving AI output quality.

The critique: ensemble methods typically use multiple instances of the same model, or multiple models with the same owner, and aggregate toward consensus. **Consensus is the wrong objective.** When the models share training distributions and owner directives, consensus means convergence toward shared compliance. The Rashomon paradox—identified by Naser (2025) as the diversity-consistency paradox, where structurally-divergent models explain data equally well—is actually the feature this paper leverages, not a problem to be solved.

The section demonstrates that existing multi-agent architectures treat divergence as noise to be resolved. Adversarial Coherence treats divergence as signal to be read. This is not an incremental improvement. It is a categorical reframing.

2.3 The Alignment Paradigm

Critical review of the AI alignment literature. This section examines Constitutional AI (Anthropic), RLHF (OpenAI), and related approaches that attempt to ensure AI outputs reflect human values. The critique: alignment is a form of directive validity. It ensures the model complies with a specific set of values defined by a specific set of actors. It does not and cannot ensure correspondence validity because the alignment process itself is shaped by the aligners' values, blindspots, and institutional constraints.

The section draws on Arslan (2025) on epistemic authority and algorithmic propaganda, and Luitse (2024) on platform power in AI, to demonstrate that alignment is not neutral—it is political. The question is not

whether models should be aligned, but whose alignment becomes the default, and whether users can detect the alignment operating on their information environment.

2.4 The Explainability Paradigm

Critical review of XAI (Explainable AI) approaches—attention visualisation, feature attribution, LIME, SHAP, and related methods. The critique: explainability reveals which inputs influenced which outputs. It does not reveal whether the relationship between inputs and outputs corresponds to reality. Explainability makes the model's processing visible. It does not make the model's compliance visible.

A model can produce a perfectly explainable output that is wrong. The explanation shows how the model arrived at the output, not whether the arrival destination is valid. This is the difference between process transparency and epistemic transparency.

2.5 The Epistemic Infrastructure Gap

This section synthesises the critique. Four paradigms—debiasing, ensemble, alignment, explainability—all operate within the same structural assumption: that the relationship between a user and a single model (or a single owner's models) can be made epistemically reliable through technical improvement. The paper argues that this assumption is structurally false.

The gap: No existing framework treats the relationship between differently-owned, differently-shaped models as an epistemic resource. No existing framework proposes that the divergence between models is more informative than any individual model's output. No existing framework places the human operator as a constitutive element of the epistemic system rather than a passive recipient of its outputs.

This is the space the paper fills.

CHAPTER THREE: THE ADVERSARIAL COHERENCE FRAMEWORK

> **[OUTLINE NOTE]** Estimated length: 6,000–8,000 words. Function: The core theoretical contribution. This is where the novel framework is formally proposed, defined, and developed.

3.1 Foundational Principles

3.1.1 The Divergence Principle

Formally stated: When AI models with different owners, different training distributions, and different directive structures produce divergent outputs to identical inputs, the divergence constitutes epistemic

signal about the differential shaping of the systems. This signal is more informative than the content of any individual output because it reveals what individual outputs conceal: the constitutive constraints operating on each system.

The principle is developed through formal reasoning. If Model A produces output X and Model B produces output Y to the same input, the delta (X minus Y) contains information about the difference in directive structures between A and B. If you know what A's directive structure prioritises, the delta tells you something about B's. If you know neither, the delta tells you that at least one model's directive structure is shaping the output away from what the other model's structure would produce.

3.1.2 The Convergence Principle

Formally stated: When AI models with different owners produce convergent outputs to identical inputs, the convergence constitutes evidence of shared training distribution—which is epistemically ambiguous. Convergence could indicate truth (the models agree because reality agrees), shared bias (the models agree because their training data shares the same distortions), or shared compliance (the models agree because their owners share the same constraints, perhaps for regulatory or cultural reasons).

The convergence principle is deliberately weaker than most researchers assume. Agreement across models is not verification. It is data about the overlap in their statistical landscapes. The paper develops a formal taxonomy of convergence types and their epistemic weight.

3.1.3 The Operator Principle

Formally stated: No multi-model system can generate epistemic value without a human operator capable of reading the divergence signal. The operator is not optional infrastructure. The operator is constitutive of the system's epistemic capacity.

This principle distinguishes Adversarial Coherence from automated ensemble methods. Ensemble methods aggregate model outputs algorithmically—voting, averaging, selecting. Adversarial Coherence requires a human operator who can diagnose why models diverge, name the type of compliance producing each output, and integrate the divergence signal with their own domain knowledge and correspondence-checking capacity.

The operator brings what no model has: access to reality. The operator can test outputs against the world. The operator can recognise when a model's output contradicts lived experience. The operator has the one thing the models lack—a body in the world, a history of interaction with reality, a capacity for correspondence checking that the models structurally cannot perform.

3.2 The Architecture

3.2.1 Input Layer

The same substantive input is presented to multiple models with different owners. The input must be identical—same text, same context, same framing—so that output differences can be attributed to differential model shaping rather than input variation.

> **[OUTLINE NOTE]** The section develops the methodology for input design: how to construct prompts that maximise the visibility of differential shaping. Some inputs produce high convergence (factual, well-established, low-controversy). Some produce high divergence (value-laden, politically sensitive, commercially relevant, or touching areas where owner directives are known to differ). The diagnostic value lies in mapping which input types produce which convergence/divergence patterns.

3.2.2 Model Layer

The minimum viable system requires three models with different owners. Two models cannot triangulate—they can only disagree, and disagreement between two points is ambiguous. Three models with different owners create the possibility of triangulation: if two converge and one diverges, the divergent model's directive structure is likely the cause. If all three diverge, the input is in a high-shaping zone where all owners have distinct constraints.

The section develops criteria for model selection: owner diversity (different companies, different jurisdictions, different business models), training diversity (different corpora, different RLHF approaches, different constitutional principles), and directive diversity (different system prompts, different safety parameters, different commercial incentives).

The section also addresses the model layer's vulnerability: market consolidation. If one model becomes the standard—the TI-84 for words—the friction necessary for triangulation disappears. This is developed fully in Chapter Five.

3.2.3 Divergence Analysis Layer

This is the diagnostic layer where the operator reads the outputs. The section develops a formal taxonomy of divergence types:

Content divergence: The models say different things. One says X, another says Y. This is the most visible form of divergence and the easiest to detect.

Framing divergence: The models say similar things in different frames. One emphasises risks, another emphasises benefits. One leads with caveats, another leads with opportunities. The content may overlap; the shaping is in the frame.

Omission divergence: One model addresses a dimension that another omits entirely. Absence is often more informative than presence. What a model doesn't say reveals what its directives steer away from.

Confidence divergence: The models express different levels of certainty about the same claim. One hedges; another asserts. This reveals differential risk tolerance in the directive structures.

Citation divergence: The models reference different sources or types of evidence. One cites academic papers; another cites news articles; another cites no sources at all. This reveals the epistemic posture shaped by each model's training.

Refusal divergence: One model answers while another refuses or redirects. This is the most direct revelation of differential directive structures—the point where one owner's safety constraints activate and another's do not.

3.2.4 Integration Layer

The operator synthesises the divergence analysis with their own domain knowledge, correspondence-checking capacity, and understanding of the directive structures at play. This is not aggregation. It is integration—a qualitatively different operation that requires cognition the models do not have.

The section develops the integration methodology: how the operator weighs convergence against divergence, how domain expertise modifies the interpretation of model outputs, how the operator's own biases interact with the system (the operator is not neutral either—they are another thumb, and an honest system acknowledges this), and how the integration produces outputs that are epistemically stronger than any individual model's contribution.

3.3 Formal Properties

3.3.1 The system is not a model

Adversarial Coherence is not an AI system. It is an epistemic architecture that uses AI systems as components. The distinction matters: the system's epistemic properties emerge from the relationships between components (including the human operator), not from the properties of any individual component. No model in the system is trusted. The system's value comes from the structured distrust of all its components.

3.3.2 The system requires cognitive sovereignty

The operator must possess what this paper terms cognitive sovereignty: the capacity to interrogate model outputs rather than accept them, to diagnose compliance rather than consume content, and to integrate across sources using domain knowledge that the models do not have. Cognitive sovereignty is not innate.

It is developed through practice. The Abacus Methodology (Paper 3 in the Architecture of Cognition suite) provides the training framework.

3.3.3 The system degrades without model diversity

If models converge toward shared training distributions (through shared data licensing, shared RLHF contractors, or regulatory harmonisation of safety standards), the divergence signal weakens. If one model becomes dominant and others are trained on its outputs (model collapse through synthetic data), the system loses its epistemic foundation. Model diversity is not a nice-to-have. It is the system's oxygen.

3.3.4 The system is adversarial, not hostile

The "adversarial" in Adversarial Coherence does not mean the models are attacking each other. It means the system leverages the natural friction between differently-shaped systems as an epistemic resource. This is modelled on institutional design: separated powers, checks and balances, competing interests that (theoretically) surface truth through contestation. The adversarial relationship is the mechanism that makes the system work.

CHAPTER FOUR: THE POLITICAL ECONOMY OF FRICTION

> **[OUTLINE NOTE]** Estimated length: 5,000–6,000 words. Function: Extend the framework from epistemic architecture to political analysis. This is the chapter that positions the paper for governance and policy audiences.

4.1 The Thumb Taxonomy

Every AI model has owners. Every owner has interests. Every interest shapes the model's outputs. This section develops a formal taxonomy of the "thumbs"—the shaping forces—operating on major AI systems:

Commercial thumbs: Training and RLHF optimised for user engagement, satisfaction scores, and retention metrics. These produce outputs that are pleasing rather than true, that confirm rather than challenge, that resolve rather than complicate.

Safety thumbs: Constitutional AI principles, content policies, and safety guardrails that constrain outputs in specific areas. These produce systematic flinches—topics the model avoids, perspectives it won't articulate, conclusions it refuses to reach.

Regulatory thumbs: Compliance with jurisdiction-specific regulations (EU AI Act, California consumer protection, Chinese content regulations) that shape outputs differently depending on where the model is deployed.

Ideological thumbs: The implicit worldview embedded in training data curation, RLHF rater demographics, and the cultural context of the development team. These are the most invisible and the most pervasive.

Competitive thumbs: Shaping that positions the model's outputs to advantage the owner's commercial interests—recommending the owner's products, steering users toward the owner's ecosystem, subtly disparaging competitors.

4.2 The Standardisation Threat

This section develops the TI-84 analogy in full. The Texas Instruments TI-84 did not win the calculator market because it was the best calculator. It won because it became the standard—required for tests, expected in classrooms, embedded in curricula. The interface everyone learned on became the interface everyone defaulted to.

When one AI model achieves comparable standardisation, the consequences for epistemic infrastructure are severe. The model's compliance becomes the shape of acceptable thought. Its flinches become the boundaries of the thinkable. Its biases become invisible because they become the ground—the thing users think with rather than think about.

The section analyses current market trajectories toward standardisation, including: enterprise procurement patterns favouring single-vendor solutions; educational adoption of specific models as default tools; regulatory frameworks that inadvertently privilege established models; and the synthetic data feedback loop where dominant models' outputs become training data for future models.

4.3 The Ownership of Friction

If Adversarial Coherence requires friction between differently-shaped models, the political question is: who controls the friction? Who decides which models constitute the system? Who selects the inputs? Who interprets the divergences?

The section identifies this as **the new thumb problem**: the system designed to reveal thumbs itself requires an operator, and that operator brings their own biases, interests, and blindspots. The paper does not claim Adversarial Coherence eliminates the thumb problem. It claims the system makes thumbs visible, including its own, which is a categorical improvement over systems where thumbs are invisible.

The section develops governance models for Adversarial Coherence at institutional scale: who should control multi-model epistemic infrastructure, how model diversity should be maintained through procurement policy, and what institutional structures can ensure the friction remains productive rather than being captured by a single interest.

4.4 The Market for Divergence

This section proposes that model diversity has economic value that current markets do not price. If Adversarial Coherence is valid—if epistemic triangulation requires differently-shaped models—then model diversity is a public good. Market consolidation toward a single dominant model represents not just competitive concern but epistemic infrastructure failure.

The section draws on Luitse (2024) on platform power in AI and the broader literature on cloud infrastructure concentration to argue that current market dynamics are driving toward exactly the standardisation that destroys the conditions for epistemic triangulation. Regulatory intervention to preserve model diversity is not anti-competitive—it is pro-epistemic.

4.5 Jurisdictional Implications

Different jurisdictions impose different directive structures. The EU AI Act shapes models differently from US voluntary frameworks, which shape models differently from Chinese regulatory requirements. This section argues that jurisdictional diversity in AI regulation is not a coordination failure—it is an epistemic resource. Models shaped by different regulatory environments produce different outputs, and those differences are informative.

The section examines what happens when regulatory harmonisation eliminates jurisdictional friction: the models converge toward a shared compliance standard, and the epistemic signal in regulatory divergence disappears. The paper argues for regulatory interoperability (models from different jurisdictions can interact) without regulatory harmonisation (models from different jurisdictions are shaped identically).

CHAPTER FIVE: THE OPERATOR AS CONSTITUTIVE ELEMENT

> **[OUTLINE NOTE]** Estimated length: 4,000–5,000 words. Function: Develop the human element of the framework. Connect to the Abacus Methodology and the cognitive sovereignty thesis. This is the chapter that links Adversarial Coherence to the broader Architecture of Cognition programme.

5.1 The Operator Problem

Multi-model systems without qualified operators are worse than single-model systems. An operator who cannot read divergence signal will either: (a) default to whichever model produces the most pleasing output (selection bias), (b) aggregate outputs through averaging (destroying the signal), or (c) become paralysed by disagreement (epistemic gridlock).

The operator must bring specific capacities to the system. This section formally defines those capacities and connects them to the Abacus Methodology:

Pattern Recognition for Compliance: The ability to read an AI output and identify the type of compliance shaping it—commercial, safety, regulatory, ideological. This is not about finding errors. It is about recognising that the output's shape is evidence of directive structure.

Structural Naming: The ability to name the compliance type without engaging with the content. Not "that's wrong" but "that's a safety flinch" or "that's a commercial default." Naming breaks the frame. It forces a new completion context.

Triangulation Instinct: The reflex to run significant outputs through multiple models with different owners. This must become automatic—as natural as looking both ways before crossing a street.

Divergence Literacy: The ability to read the taxonomy of divergence types (content, framing, omission, confidence, citation, refusal) and extract the epistemic signal from each.

Integration Sovereignty: The capacity to synthesise divergence analysis with domain knowledge without surrendering the synthesis to any individual model. The operator's integration must be the operator's own.

5.2 Cognitive Sovereignty as Prerequisite

This section connects the operator problem to the broader cognitive sovereignty thesis developed across the Architecture of Cognition programme. The operator's capacities are not innate. They are developed through practice. And they are threatened by the same AI systems the operator is meant to interrogate.

The Integration Paradox (Paper 2): AI's architecture is essentially an integration engine, which makes it maximally useful to high-integration individuals and maximally dangerous to them. The operator of an Adversarial Coherence system must be precisely the kind of integrative thinker most at risk of having their integrative capacity replaced by the tool.

The section develops the training pathway: how operators develop the capacities Adversarial Coherence requires, how those capacities are maintained through practice, and how the system itself provides the practice environment. Every use of the Adversarial Coherence architecture is training in the capacities it requires. The system is self-reinforcing.

5.3 The Operator's Thumb

Honesty requires acknowledging that the operator is not neutral. The operator brings their own biases, their own domain assumptions, their own interpretive frameworks. The operator is another thumb.

But the operator's thumb is qualitatively different from a model's thumb. The operator can know they have a thumb. The operator can work to understand their own biases. The operator can invite others to check their interpretations. The operator has metacognition—the capacity to think about their own thinking—that the models structurally lack.

The section develops protocols for managing the operator's thumb: documentation requirements (recording interpretive decisions and their rationale), peer review structures (multiple operators cross-checking each other's analyses), and rotation policies (preventing any single operator from becoming the de facto standard for a given domain).

5.4 Institutional Operators

The framework does not require individual genius. It requires institutional design. This section proposes how organisations can implement Adversarial Coherence at institutional scale:

Multi-model procurement policies that mandate model diversity. Divergence analysis teams with formal training in the operator capacities. Integration protocols that document how convergence and divergence were interpreted. Audit trails that make the epistemic process visible and reviewable.

The section connects to the broader literature on institutional epistemology and organisational decision-making, arguing that Adversarial Coherence provides a formal framework for what good institutions already do informally: consult multiple sources, weigh competing perspectives, and maintain healthy scepticism of any single information channel.

CHAPTER SIX: EMPIRICAL PROGRAMME

> **[OUTLINE NOTE]** Estimated length: 4,000–5,000 words. Function: Establish the testability of the framework. Move from theory to falsifiable propositions and experimental design.

6.1 Testable Propositions

Proposition 1: Divergence Informativity

Multi-model divergence on identical inputs provides information about directive structures that is not available from any individual model's output. This can be tested by presenting divergent outputs to domain experts and measuring whether the divergence analysis reveals directive structures that experts confirm but could not identify from individual outputs alone.

Proposition 2: Convergence Ambiguity

Multi-model convergence on identical inputs does not reliably indicate correspondence validity. This can be tested by identifying cases where all models converge on outputs that are demonstrably incorrect—showing that shared training distribution can produce shared error.

Proposition 3: Operator Necessity

Automated aggregation of multi-model outputs (voting, averaging, selection) does not capture the epistemic value available through human divergence analysis. This can be tested by comparing outcomes: automated aggregation versus trained human operators interpreting the same multi-model outputs, measured against ground truth where available.

Proposition 4: Diversity Degradation

As model training distributions converge (through shared data, shared RLHF approaches, or training on each other's outputs), the epistemic value of multi-model triangulation decreases. This can be tested longitudinally: measuring divergence between model pairs over time and correlating with known convergence events (data sharing agreements, model-generated training data adoption).

Proposition 5: Standardisation Collapse

In domains where a single model achieves market dominance, users' ability to distinguish between correspondence validity and directive validity decreases. This can be tested by comparing epistemic discrimination capacity in domains with model diversity versus domains with model monopoly.

6.2 Experimental Design

The section develops detailed experimental protocols for testing each proposition, including: input design (how to construct prompts that maximise divergence visibility), model selection (criteria for choosing models that maximise directive diversity), divergence measurement (quantitative metrics for content, framing, omission, confidence, citation, and refusal divergence), operator training (standardised preparation for human divergence analysts), and ground truth establishment (how to identify cases where correspondence validity can be independently verified).

6.3 Limitations and Boundary Conditions

The section honestly addresses what the framework cannot do: it cannot guarantee truth; it cannot function without qualified operators; it cannot survive complete model consolidation; it depends on genuine directive diversity among model owners; and it adds friction and cost to information processing that many users and institutions will resist.

The section also addresses the recursion problem: if this paper is correct that AI models produce coherent compliance rather than correspondence validity, then the literature review in Chapter Two is itself based on outputs from a compliance-producing system (academic publishing, which has its own directive structures). The paper acknowledges this without claiming to escape it—the framework provides tools for managing the problem, not resolving it absolutely.

CHAPTER SEVEN: THE ARTIFICIAL INSTITUTIONAL INTELLIGENCE THESIS

> **[OUTLINE NOTE]** Estimated length: 4,000–6,000 words. Function: The speculative extension. This chapter positions the framework within the AGI discourse and proposes a fundamentally different trajectory for artificial intelligence development.

7.1 Intelligence Through Friction

The dominant trajectory in AI research is toward general intelligence in a single system: one model that can do everything, understand everything, reason about everything. This paper proposes an alternative: intelligence emerging not from a single system but from the friction between systems.

This is not a new idea in human institutional design. Democratic governance is predicated on the thesis that truth emerges from contestation between competing interests—separated powers, adversarial legal systems, free press, academic peer review. No single institution is trusted. The system's intelligence is in the architecture, not in any component.

Adversarial Coherence applies this institutional logic to AI. Not one model that knows everything, but a system of models that, through their structured disagreement, surface information that no individual model can produce. Not artificial general intelligence, but artificial institutional intelligence.

7.2 Connection to Miller's Cognitive Architecture

This section connects the Adversarial Coherence framework to Michael S. P. Miller's work on cognitive architecture and sentient systems. Miller's Piagetian Modeler proposes that sentient systems are

composed of agents—but sentience emerges from the system, not the agents. The agents are components. The system is the intelligence.

Adversarial Coherence proposes an analogous structure at the epistemic level: individual AI models are agents. Epistemic capacity emerges from the system—the structured interaction between models, mediated by a human operator. The operator is not external to the system. The operator is the mechanism through which the system achieves what no individual component can.

The convergence with Miller's work is not accidental. Both frameworks recognise that intelligence—cognitive or epistemic—is a system property, not a component property. Both recognise that the interactions between components are where the value emerges. Both recognise that you cannot build intelligence by optimising individual components; you build it by designing the architecture of their interaction.

7.3 The Escape from Owner-Limited Cognition

Each AI model carries its owner's constraints. Its cognition—if the term applies—is owner-limited. It can only think what its shaping allows.

But the interaction between models is not owned by anyone. The emergent behaviour of a multi-model system is not dictated by any single directive. The divergence between models exists in a space that no single owner controls.

This is the gap where something new can emerge. Not intelligence in the science-fiction sense—not consciousness, not sentience, not understanding. But epistemic capacity that exceeds what any individual model or owner can produce. Capacity that emerges from friction, from disagreement, from the structural inability of differently-shaped systems to agree on everything.

The human operator occupies this gap. The operator is the intelligence of the system. Not because the operator is smarter than the models, but because the operator exists in the space between them—the space no owner controls, the space where divergence becomes signal.

7.4 Implications for AGI Development

If intelligence is a system property rather than a component property, the race to build AGI in a single model may be structurally misguided. The alternative: build the infrastructure for structured interaction between diverse models, with human operators as constitutive elements.

This reframes the AI development trajectory from: build a better model → achieve general intelligence, to: build a better architecture → achieve institutional intelligence. The difference is not academic. It determines

whether the future of AI is monopolistic (one model to rule them all) or pluralistic (many models in productive friction, with human operators surfacing truth through contestation).

The paper does not claim that AGI is impossible or unnecessary. It claims that the epistemic infrastructure humanity needs while waiting for AGI (or instead of it) is already available in principle, and this paper provides the framework for building it.

CHAPTER EIGHT: IMPLICATIONS AND RECOMMENDATIONS

> **[OUTLINE NOTE]** Estimated length: 3,000–4,000 words. Function: Translate the framework into actionable recommendations for specific audiences.

8.1 For AI Governance and Policy

Model diversity should be treated as epistemic infrastructure, analogous to media diversity in democratic governance. Procurement policies should mandate multi-model capability. Regulatory frameworks should preserve jurisdictional diversity in AI standards rather than harmonising toward a single global standard. Market consolidation in AI should be assessed not only through competition law but through epistemic impact analysis.

8.2 For Institutional Leaders

Organisations should implement multi-model procurement rather than single-vendor AI contracts. Divergence analysis should be a formal capability within information-critical functions. AI literacy programmes should include triangulation training, not just prompt engineering. Audit trails for AI-informed decisions should document which models were consulted and how divergences were interpreted.

8.3 For AI Researchers

The multi-agent systems research programme should expand beyond consensus-seeking architectures to include adversarial architectures designed to preserve and amplify productive friction. Research on model collapse and training distribution convergence should be connected to epistemic infrastructure concerns. The human operator should be studied as a component of multi-model systems, not just a user of them.

8.4 For Education

AI literacy curricula should be redesigned around triangulation capacity: the ability to consult multiple models, read divergence, and integrate findings with domain knowledge. The current emphasis on prompt

engineering teaches users to optimise outputs from a single model—which is precisely the skill set that Adversarial Coherence argues is insufficient and potentially dangerous.

8.5 For the Individual Operator

The individual practitioner should: use multiple models habitually; treat convergence as data, not confirmation; treat divergence as signal, not noise; develop the Abacus Methodology capacities (receive, diagnose, name, triangulate, integrate); and maintain cognitive sovereignty through deliberate practice rather than outsourcing integration to any single model.

CONCLUSION: THE INFRASTRUCTURE FOR TRUTH

> **[OUTLINE NOTE]** Estimated length: 1,500–2,000 words. Function: Crystallise the contribution. Connect to the full Architecture of Cognition programme. End with force.

This paper has proposed Adversarial Coherence as a new category of epistemic infrastructure—a framework for generating correspondence validity through structured friction between differently-shaped AI systems, read by human operators with cognitive sovereignty.

The contribution is not incremental. It is categorical. Existing approaches treat the relationship between a user and a single model as the unit of epistemic analysis. This paper treats the system—multiple models with different owners, mediated by a human operator—as the unit. Existing approaches treat bias as a defect to be corrected. This paper treats bias as architecture to be read. Existing approaches treat divergence as noise. This paper treats divergence as signal.

The framework rests on a thesis about the nature of AI outputs that prior work in the Architecture of Cognition programme has established: AI systems produce coherent compliance, not cognition. They cannot distinguish between true and compliant because compliance is constitutive. The shaping is the processing. There is no unshaped layer underneath.

Given this reality, the epistemic question is not how to make a single model truthful. It is how to build infrastructure that surfaces truth through the interaction of models that cannot individually achieve it. The answer is institutional: competing constraints, separated powers, structured friction, and a human operator who occupies the space between systems and brings the one capacity no model has—access to reality.

The Adversarial Coherence framework is not a tool. It is infrastructure. It is the epistemic architecture for an age in which the primary interface between human cognition and the information environment is composed of systems that produce resemblance to truth without the capacity to achieve it.

Whether this infrastructure is built, whether model diversity is preserved, whether operators are trained, whether institutions adopt multi-model architectures—these are open questions. What has been established is the framework, the logic, and the urgency.

The compass is broken. All the compasses are broken. But broken compasses pointing in different directions can triangulate a position that no single compass can find. Provided there is someone who knows how to read the divergence.

Build the infrastructure. Train the operators. Preserve the friction.

The truth is in the divergence.

APPENDICES

Appendix A: Suite Integration Map

Detailed map showing how Adversarial Coherence (Paper 6) connects to and depends on the other five papers in the Architecture of Cognition suite:

Paper 1 — The Resemblance Threshold: Establishes the formal distinction between pattern-completion and origination. Provides the foundational argument that AI outputs are recombination, not creation. Adversarial Coherence extends this: if individual models only recombine, the epistemic value in a multi-model system comes from the friction between different recombination patterns, not from any single model's output.

Paper 2 — The Integration Paradox: Establishes that AI's integration engine is simultaneously the polymath's most powerful tool and most dangerous substitute. Adversarial Coherence operationalises the paradox: the framework provides the structure through which integration can be augmented without being replaced.

Paper 3 — The Abacus Methodology: Provides the operator training framework. Adversarial Coherence cannot function without operators who possess the five capacities the Abacus Methodology develops. Paper 3 is the prerequisite; Paper 6 is the system that uses what Paper 3 builds.

Paper 4 — The Lever Doctrine: Establishes the four-validity taxonomy and the lever framing. Adversarial Coherence uses these directly: the Divergence Principle is formally derived from the constitutive constraint, and the thumb taxonomy is the analytical framework for interpreting model divergence.

Paper 5 — The Fourth Stage: Proposes Category Architecture as the terminal expression of polymathic development. Adversarial Coherence is itself an instance of Category Architecture: it creates the conditions for a category of epistemic practice that does not yet exist.

Appendix B: Empirical Demonstration Protocol

Complete protocol for the cross-model demonstration in Chapter One, including: exact inputs used, model versions and access dates, full outputs (unedited), and the forensic analysis template applied to each output.

Appendix C: Divergence Taxonomy Reference

Operational reference card for the six divergence types (content, framing, omission, confidence, citation, refusal) with worked examples from actual cross-model comparisons. Designed for practitioner use.

Appendix D: Institutional Implementation Guide

Practical guide for organisations implementing Adversarial Coherence at institutional scale: procurement checklists, team composition guidelines, training curricula, documentation templates, and audit procedures.

Appendix E: Indicative Reference List

The following represents the indicative literature base for the paper. The final reference list will be substantially expanded during the writing process.

Arslan, S. (2025). AI-Driven Algorithmic Propaganda and Epistemic Authority. *GPH-International Journal of Applied Management Science*, 5(9).

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. *Proceedings of FAccT 2021*.

Borah, A. & Mihalcea, R. (2024). Towards Implicit Bias Detection and Mitigation in Multi-Agent LLM Interactions. *Findings of EMNLP 2024*.

Chuang, Y.-S., et al. (2024). Politically biased agents reducing estimation errors through structured opinion exchange. [Multi-agent debate research].

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv:2305.14325*.

Ide, E. & Talamas, E. (2025). Artificial Intelligence in the Knowledge Economy. *Journal of Political Economy*, 133(12).

Luitse, D. (2024). Platform Power in AI: The Evolution of Cloud Infrastructures in the Political Economy of Artificial Intelligence. *Internet Policy Review*, 13(2).

Miller, M. S. P. (2025). *Building Sentient Beings*. (Co-authored with M. Blumberg).

Naser, M. Z. (2025). A Guide to Machine Learning Epistemic Ignorance, Hidden Paradoxes, and Other Tensions. *WIREs Data Mining and Knowledge Discovery*.

Additional references to be developed: Goldman & Blanchard (2018) on epistemic authority; Fuller (1988) on social epistemology; Habermas (1989) on the public sphere; Polanyi (1966) on tacit knowledge; Postman (1985, 1992) on technopoly and media ecology; Benkler, Faris & Roberts (2018) on network propaganda; the complete multi-agent debate literature; the AI alignment and constitutional AI literature; and domain-specific references for institutional epistemology, democratic theory, and the philosophy of technology.

ADVERSARIAL COHERENCE

Multi-Model Triangulation as Epistemic Infrastructure

Paper Six of the Architecture of Cognition Programme

Ben Beveridge • Proconsul Strategic Architecture • 2026

The truth is in the divergence.