Epistemic Frustration: Dimensional Collapse and the Limits of Falsifiability in Complex Systems

Ian Todd

Sydney Medical School

University of Sydney

Sydney, NSW, Australia

itod2305@uni.sydney.edu.au

Abstract

Complex adaptive systems—whether scientific paradigms, economic institutions, biological organisms, or social coalitions—exhibit a characteristic lifecycle: expansion into high-dimensional possibility space, consolidation around successful strategies, exploitation of narrow optima, and eventual brittleness. I argue that this trajectory produces a distinctive epistemic failure mode: as systems mature, they transition from *epistemic-first* regimes (where truth-seeking dominates) to *coordination-first* regimes (where preserving cooperation dominates). In coordination-first regimes, disagreement is reinterpreted as defection, uncertainty becomes taboo, and anomalies are absorbed into ever-richer error models rather than used to revise core commitments. This is not irrationality but optimization under constraint: when the costs of coordination failure are nonlinear and the underlying reality is high-dimensional, local epistemic accuracy becomes subordinate to global stability. I formalize this transition through the concept of *epistemic frustration*: the condition where optimal solutions in high-dimensional configuration space project into mutually contradictory positions in low-dimensional discourse. Under epistemic frustration, disagreement does not signal error but geometric incompatibility. I propose a diagnostic criterion: a system enters an unfalsifiable regime when explaining discrepancies requires adding more degrees of freedom to the error model than the data summary gains in informational constraints. This framework unifies phenomena from paradigm exhaustion in science to monopolistic consolidation in markets to senescence in organisms, revealing them as instances of a general dynamical pattern rather than domain-specific pathologies. The question is whether we will recognize exploitation traps from the inside—or mistake local optimality for global virtue while the error-model degrees of freedom quietly exceed the data.

Keywords: philosophy of science; falsifiability; complex systems; coordination; epistemic frustration; dimensional collapse

1 Introduction

Why do successful systems become brittle? Why do mature scientific paradigms resist revision even as anomalies accumulate? Why do dominant firms extract rather than innovate? Why do aging organisms lose resilience despite sophisticated repair mechanisms? Why do ideological movements calcify into orthodoxy?

The standard explanations are moralized: greed, corruption, complacency, cognitive bias, institutional capture. These narratives identify villains and imply that better actors or better incentives would prevent decline. I argue that this moralization, while locally accurate, obscures a deeper structural pattern. Complex adaptive systems undergo a predictable phase transition from *exploration* to *exploitation*, and late-stage exploitation produces characteristic epistemic pathologies regardless of the virtue of individual agents.

The pattern is thermodynamic, not moral.[1] Maintaining high-dimensional adaptive ca-

[1] I use "thermodynamic" here in the information-theoretic sense: maintaining distinguishable states against noise requires ongoing work. Whether this entails literal Landauer costs at molecular scales or serves as a useful metaphor at institutional scales, the structural logic is the same: high-dimensional adaptive capacity is expensive to maintain.

pacity requires continuous energetic expenditure: diversity, redundancy, tolerance for failure, slack. As systems mature and coordination yields increasing returns, they naturally compress into lower-dimensional configurations. This compression is initially efficient—it enables scaling, standardization, and optimization. But it also reduces resilience, narrows the space of viable responses, and eventually produces brittleness.

The epistemic consequences are profound. Early-stage systems can afford to be *epistemic-first*: truth-seeking is rewarded because the system has slack to absorb errors and incorporate corrections. Late-stage systems must be *coordination-first*: preserving cooperative behavior dominates truth-seeking because errors in coordination impose nonlinear, system-level costs. In coordination-first regimes, dissent is not evaluated on its merits but on its threat to stability. Uncertainty becomes dangerous. Anomalies are absorbed rather than explored.

This paper makes three contributions:

1. **Conceptual:** I define *epistemic frustration* as projection-induced incompatibility—the condition where high-dimensional optima project into mutually contradictory low-dimensional positions, making disagreement geometric rather than purely epistemic.

2. **Dynamical:** I model a predictable lifecycle shift from epistemic-first to coordination-first regimes, occurring when coordination stakes are high and observations access only low-dimensional projections of high-dimensional processes.

3. **Methodological:** I propose an operational diagnostic for practical unfalsifiability: a system enters an unfalsifiable regime when error-model degrees of freedom grow faster than informational constraints from new data.

Section 2 characterizes the exploration-to-exploitation lifecycle and its thermodynamic basis. Section 3 introduces *epistemic frustration*: the condition where high-dimensional optima project into contradictory low-dimensional representations. Section 4 analyzes *scale-bound normativity*: why moral intuitions are locally correct but globally misleading. Section 5 examines how coordination-first epistemics produces coalitional enforcement dynamics in secular institutions. Section 6 situates the framework within existing philosophical literature. Section 7 proposes a diagnostic criterion for unfalsifiability. Section 8 traces these patterns across domains: scientific paradigms, economic concentration, biological senescence, and cosmological inference. Section 9 draws implications for how we should reason under dimensional constraint.

Scope Conditions

Before proceeding, I specify when the framework applies. The exploration-to-exploitation transition and its epistemic consequences emerge under specific structural conditions:

- **High coordination stakes**: Errors in coordination impose nonlinear costs (trust collapse, system failure, cascade effects).
- **Projection bottleneck**: Low-dimensional observables are generated by high-dimensional processes, so many distinct causes project to the same effect.
- **Increasing returns to standardization**: Network effects, shared infrastructure, and pipeline investment make deviation increasingly costly.
- **Maturity and lock-in**: Large installed base, career structures, and institutional memory create path dependence.

When these conditions are absent—when coordination stakes are low, observables are rich, alternatives are cheap, and lock-in is weak—the pathologies described here need not arise. The framework is a conditional claim, not a universal law. But the conditions are common enough that the pattern recurs across domains.

Notation and Dimensionality Concepts

The framework employs several related notions of dimensionality. For reference:

Symbol	Meaning
$\mathcal{X}$, x	High-dimensional configuration/state space
$\mathcal{Y}$, y	Low-dimensional observable/discourse space
$\mathcal{A}$, a	Action/policy space
$f : \mathcal{X} \to \mathcal{Y}$	Projection (measurement/discourse pipeline)
E_θ	Error model with parameters θ
D_{sys}	System dimensionality (internal degrees of freedom)
D_{obs}	Observable dimensionality (what measurement accesses)
D_{eff}	Effective dimension: $(\text{tr } g)^2/\text{tr}(g^2)$ on Fisher spectrum
R	DOF ratio: (error-model DOF) / (informational constraints)

Running example. Consider a biological system with $D_{\text{sys}} = 10^4$ internal molecular states, observed through $D_{\text{obs}} = 10$ measurable biomarkers. The projection degeneracy is enormous: $\sim 10^3$ internal configurations map to each observable profile. Policies can adjust perhaps $D_{\text{act}} = 3$ interventions (drug, dose, timing). The high-dimensional optimum (the "right" treatment for this patient's full state) may not survive projection into implementable policy space—this is epistemic frustration at the level of personalized medicine.

2 The Exploration-Exploitation Lifecycle

2.1 The Universal Pattern

Complex adaptive systems exhibit a canonical lifecycle that recurs across substrates [March, 1991]:

Phase 1: Exploration. The system expands into a new possibility space. Many degrees of freedom are active. Multiple competing frameworks coexist. Failure is cheap; novelty is rewarded. In science, this is the period of wild ideas, messy data, and rapid conceptual growth. In markets, it is the period of many small firms experimenting with diverse strategies.

In organisms, it is development and early adulthood, with high regenerative capacity and behavioral flexibility.

Phase 2: Consolidation. Some frameworks prove useful. Redundant dimensions are pruned. A shared language emerges. Inference becomes efficient. This is where paradigms form—in both the Kuhnian sense [Kuhn, 1962] and the algorithmic sense. Coordination yields increasing returns, so systems converge on common representations.

Phase 3: Exploitation. The system now exploits a narrow slice of the original space. Everything is tuned to fit the established framework. Progress becomes incremental, statistical, and pipeline-driven. Deviations are absorbed rather than explored. This is where mature paradigms, dominant platforms, and aging organisms find themselves.

Phase 4: Brittleness. The system runs out of genuinely new explanatory degrees of freedom. Residuals grow. Flexibility migrates into error models, patches, and exceptions. The structure becomes internally consistent but externally fragile. Shocks that would have been absorbed in earlier phases now threaten collapse.

This is not a moral narrative. It is a description of how optimization under constraint reshapes possibility space (Figure 1).

2.2 The Thermodynamic Basis

Why does this always happen? Because maintaining high-dimensional structure is expensive.

Exploration requires:

- Energy to sustain diverse configurations
- Tolerance for failed experiments
- Redundancy that appears wasteful from an efficiency perspective
- Slack that allows deviation from optimal exploitation

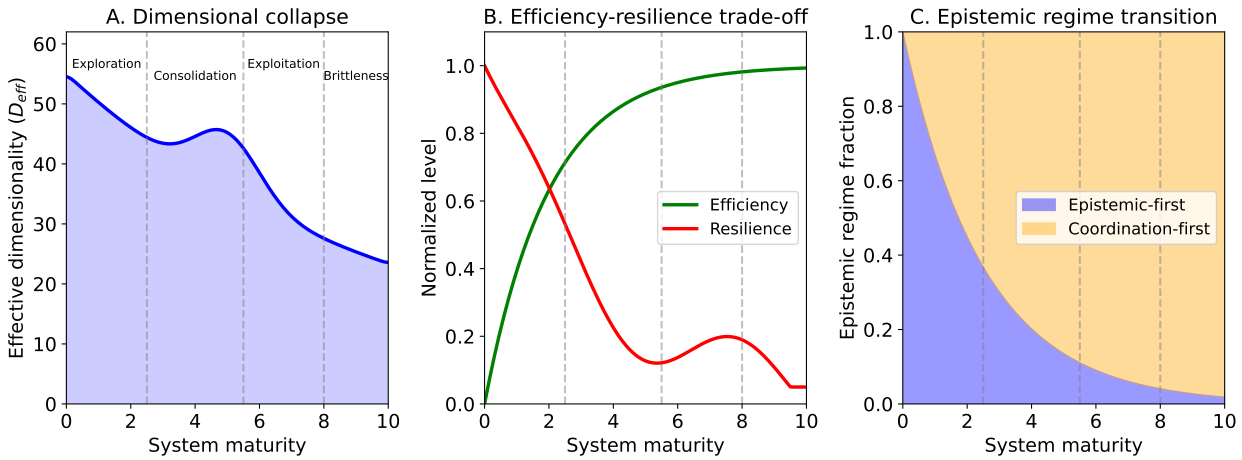

Figure 1: **The exploration-exploitation lifecycle** (schematic). (A) Effective dimensionality collapses as systems mature, moving through exploration, consolidation, exploitation, and brittleness phases. Formally, effective dimension can be defined as the participation ratio on the Fisher information spectrum: $D_{\text{eff}} = (\text{tr } g)^2/\text{tr}(g^2)$, which measures how many directions in state space produce distinguishable futures. (B) Efficiency increases while resilience decreases—the system becomes optimized but fragile. (C) Epistemic regime transitions from truth-seeking (epistemic-first) to stability-seeking (coordination-first).

As constraints tighten—whether through resource limitation, competitive pressure, or coordination demands—systems naturally collapse into simpler representations, reusable heuristics, and local optima.

Joseph Tainter's analysis of civilizational collapse makes this explicit: beyond a point, complexity yields diminishing returns, so systems simplify [Tainter, 1988]. The same logic applies to scientific paradigms, economic institutions, and biological organisms.[2] Simplification is not failure; it is the expected trajectory of systems that have exploited their initial

[2] The inclusion of economic and political systems in a biology journal is not a category error. Political-economic structures *are* biological phenomena—they emerge from primate social dynamics and remain governed by the same constraints. Multilevel societies with nested coalitions, coordination hierarchies, and collective action problems appear across papionins, colobines, and hominins [Grueter et al., 2012]. Crucially, these structures emerge when social complexity exceeds individual cognitive capacity: primates begin using low-dimensional representations of conspecifics, developing signaling systems that compress high-dimensional social information into tractable codes [Grueter and Lüpold, 2024]. Human institutions are the most elaborately externalized projection of this same cognitive compression. This connects directly to Stanford's "unconceived alternatives" [Stanford, 2006]: we cannot conceive alternatives to our paradigms for the same reason primates cannot conceive alternatives to their social structures—both arise from the same dimensional bottleneck between environment and representation. The question this paper addresses—how do living things navigate complex systems that constrain their epistemic access—applies whether those systems are metabolic, neural, or institutional.

possibility space.

The crucial insight is that *exploitation is not merely an alternative to exploration; it actively suppresses exploration.* Resources allocated to exploitation are unavailable for exploration. Coordination mechanisms that enable scaling also penalize deviation. Standards that permit interoperability also foreclose alternatives. The system becomes locked into a basin of attraction from which escape becomes increasingly costly.

2.3 Why This Matters for Epistemology

The exploration-exploitation transition has direct epistemic consequences. In exploration phases, systems can afford to be wrong. Errors are cheap, corrections are rewarded, and diversity of perspective is an asset. In exploitation phases, systems cannot afford disruption. Errors in coordination are expensive, corrections threaten stability, and diversity becomes noise.

This produces a shift in what counts as rational behavior:

- **Exploration-phase rationality**: Update beliefs based on evidence. Tolerate uncertainty. Reward novel hypotheses. Accept that current models are incomplete.

- **Exploitation-phase rationality**: Preserve coordination. Minimize variance. Absorb anomalies into existing frameworks. Treat uncertainty as threat.

Both are rational given their respective constraints. The problem is that exploitation-phase rationality *looks like* irrationality from an epistemic-first perspective. Resistance to evidence, hostility to dissent, and absorption of anomalies appear as failures of reasoning. But they are optimization under a different objective function: coordination maintenance rather than truth-seeking.

3 Epistemic Frustration

3.1 When the Right Answer Doesn't Survive Projection

Many complex systems exhibit *frustrated optima*: configurations where multiple valid constraints cannot all be satisfied simultaneously in any single low-dimensional representation.

The concept originates in physics. In a spin glass, local magnetic interactions may be mutually incompatible: satisfying one constraint necessarily violates another. The system cannot reach a single ground state; instead, it becomes trapped in a rugged landscape of local minima [Mézard et al., 1987].

I propose that epistemic systems exhibit an analogous phenomenon. To make this precise, consider a minimal formal skeleton:

Box 1: Formal Definitions

Let **world states** be $x \in \mathcal{X}$ (high-dimensional configuration space).

Let **observables/discourse summaries** be $y \in \mathcal{Y}$ (low-dimensional representation space).

Let the **measurement/discourse pipeline** be $y = f(x) + \epsilon$, where $f : \mathcal{X} \to \mathcal{Y}$ is a projection and ϵ is structured by an error model E_θ.

Let **constraints/objectives** be $C_i(x)$ or $C_i(y)$, and let **policy/action** be $a \in \mathcal{A}$ (also low-dimensional).

Epistemic frustration obtains when: (i) there exists a feasible set in $\mathcal{X}$ that satisfies constraints $\{C_i\}$, but (ii) for any low-dimensional action $a \in \mathcal{A}$ or discourse point $y \in \mathcal{Y}$, at least one constraint is violated—because f is many-to-one and/or $\mathcal{A}$ is too low-dimensional to span the Pareto set.

This connects to **multi-objective optimization**: when no scalar objective can optimize all constraints simultaneously, the feasible region in $\mathcal{X}$ projects to incompatible optima in $\mathcal{Y}$.

Three distinct notions of dimensionality matter here: (i) *dimensionality of the generative* *process* (latent causes in $\mathcal{X}$), (ii) *dimensionality of the representational/discourse space* (what can be said or argued in $\mathcal{Y}$), and (iii) *dimensionality of the action/policy space* (what can be done in $\mathcal{A}$). Dimensional collapse can occur in any of these, with different epistemic consequences.

These notions can be formalized information-geometrically. Let $\mathcal{M}$ be the *transition manifold*—the family of probability distributions over future states reachable from current configurations, equipped with the Fisher information metric as its intrinsic geometry. The system's dimensionality is $\dim(\mathcal{M})$: the number of independent ways futures can differ. The observer's discourse manifold $\mathcal{N}$ has lower dimension when measurement or communication compresses the state space. Epistemic frustration arises when the feasible region in $\mathcal{M}$ does not survive projection $\pi : \mathcal{M} \to \mathcal{N}$—optimization targets that are compatible in high-dimensional constraint space become mutually exclusive in low-dimensional policy space. The projection degeneracy $\deg(y) = \dim(\mathcal{M}) - \dim(\mathcal{N})$ quantifies how many system configurations map to each observable state; when degeneracy is high, disagreement can be geometric rather than epistemic.

Consider the constraints that a mature knowledge system must satisfy:

- Minimize immediate harm
- Preserve institutional trust
- Prevent concentration of power
- Allow innovation and correction
- Maintain resilience to shocks
- Ensure legitimacy across constituencies
- Avoid catastrophic tail risks

No scalar objective function can optimize all of these simultaneously. The "optimal" solution exists in high-dimensional space, where these constraints define a complex feasible region. But policy, discourse, and moral judgment operate in low-dimensional projections of this space.

Every projection violates something.

This is epistemic frustration: the condition where the high-dimensional optimum projects into mutually contradictory positions in low-dimensional representation space (Figure 2).

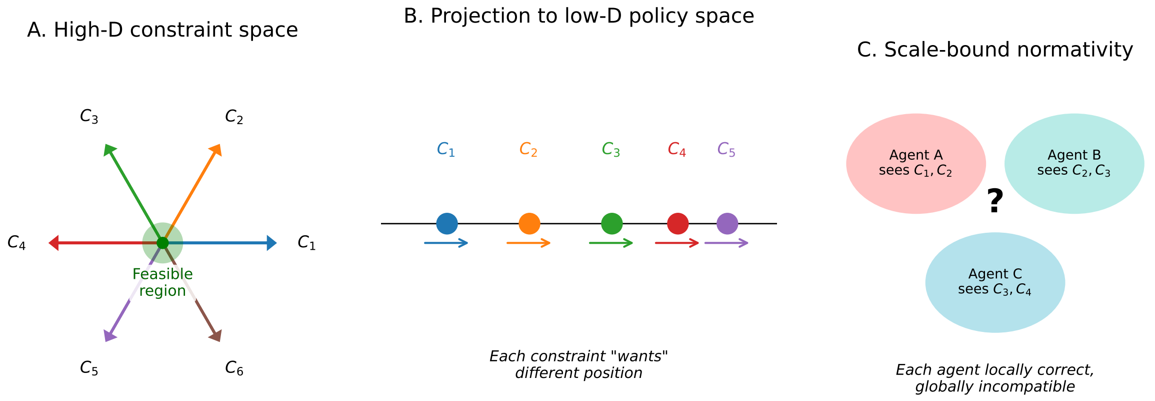
A. High-D constraint space
C_3
C_2
C_4
C_1
Feasible region
C_5
C_6
B. Projection to low-D policy space
C_1
C_2
C_3
C_4
C_5
Each constraint "wants" different position
C. Scale-bound normativity
Agent A sees C_1, C_2
Agent B sees C_2, C_3
?
Agent C sees C_3, C_4
Each agent locally correct, globally incompatible

Figure 2: **Epistemic frustration** (schematic). (A) In high-dimensional constraint space, multiple constraints (C_1–C_6) define a small feasible region. (B) When projected to low-dimensional policy space, different constraints "want" different positions—no single point satisfies all. (C) Agents seeing different subsets of constraints are each locally correct but globally incompatible, producing scale-bound normativity.

3.2 Disagreement as Geometry

Under epistemic frustration, disagreement does not signal that one party is wrong. It signals that different parties are projecting the same high-dimensional reality onto different low-dimensional coordinate systems.

Consider a concrete example. One analyst says: "Suppressing recessions has allowed extreme wealth concentration and threatens democratic institutions." Another responds: "Recessions cause acute human suffering; preventing them is morally imperative."

Both can be correct simultaneously:

- The first is making a claim about long-run manifold deformation: policies that suppress short-term volatility can shift risk distributions, amplify winner-take-all dynamics, and concentrate political influence.
- The second is making a claim about local welfare: recessions impose severe, concentrated costs on vulnerable populations.

These are not contradictory claims about the same object. They are claims about different projections of a high-dimensional dynamical system. The "right answer" would require optimizing across both—but no policy that lives in low-dimensional implementation space can do so.

3.3 Why Moral Conflict Is Structural

Human moral cognition evolved for low-dimensional, repeated games with relatively stable constraints. When faced with frustrated optima:

- Each agent perceives a true slice of the constraint space
- Violation of their slice registers as injustice
- They infer bad faith rather than projection loss

Hence escalation. The structure of high-dimensional optimization guarantees that any implementable policy will violate someone's legitimate constraint. Moral conflict is not a failure of reasoning or goodwill; it is a geometric consequence of dimensional reduction.

This reframes political and institutional conflict. The question is not "who is right?" but "which constraints are we choosing to satisfy, and which to violate?" That is a distributional question, not an epistemic one. But because we lack the conceptual vocabulary for frustrated optima, we translate distributional conflict into moral accusation.

4 Scale-Bound Normativity

4.1 Local Correctness, Global Mismatch

Moral intuitions are calibrated for local optimization on a high-dimensional social manifold. Within an agent's accessible neighborhood, individual choices—cooperation, defection, honesty, deception—often do make a decisive difference. Moral judgment is useful because it constrains harmful local moves.

But the same actions, when aggregated across many agents and iterated through network feedback, reshape the global manifold. Increasing returns to coordination amplify winners, reduce strategy diversity, and drive dimensional compression. The local gradient that each agent follows is real; the global attractor it produces may be invisible from any local vantage point.

I call this *scale-bound normativity*: moral intuitions that are valid at one scale but misleading at another.

4.2 Why Global Awareness Is Penalized

Agents embedded in complex systems are not merely limited in their access to global dynamics; they are frequently *incentivized to ignore them.*

Local optimization is rewarded with immediate feedback:

- Career advancement
- Social approval
- Resource access
- Reduced cognitive load

Sensitivity to global structure incurs costs:

- Cognitive expense of modeling system-level dynamics

- Social friction from questioning shared assumptions
- Career risk from challenging institutional narratives
- Delayed and diffuse benefits that don't cash out locally

As a result, epistemic blind spots are not accidental but adaptive. Blindness to manifold-level deformation enables agents to climb local gradients more efficiently. Selection favors myopia.

4.3 The Moralization of System Dynamics

A recurrent feature of complex-system exhaustion is the moralization of simplification. When adaptive degrees of freedom collapse—whether in markets, institutions, organisms, or scientific paradigms—observers tend to interpret the resulting rigidity as vice (greed, corruption, decadence) rather than as the expected outcome of incentive-driven compression.

Moral narratives provide cognitively efficient, low-dimensional explanations. They identify villains, assign blame, and suggest interventions (remove the bad actors). They also serve coordination functions: shared moral frameworks enable collective action even when causal understanding is incomplete.

But moral narratives risk obscuring the underlying mechanism. If dimensional collapse is a dynamical attractor rather than a choice, then replacing actors will not prevent recurrence. The same incentive landscape will produce the same compression.

The appropriate conclusion is not that morality is false, but that it is scale-bound. Moral reasoning is an adaptive interface for local coordination, not a theory of global system dynamics.

5 Coordination-First Epistemics

5.1 When Truth-Seeking Becomes Dangerous

In high-stakes coordination domains, epistemic critique and social stability operate on different objective functions. Local claims about model error or uncertainty may be epistemically correct yet socially destabilizing when communicated broadly.

Consider public health. Medical knowledge is genuinely uncertain, contested, and revisable. Internal scientific discourse is highly critical. But public-facing communication must balance accuracy against the risk of undermining the cooperative behaviors (vaccination, masking, treatment adherence) that depend on institutional trust.

When trust is treated as a public good, and public goods invite moral enforcement, then:

- Expressing uncertainty becomes defection
- Critique becomes sabotage
- Nuance becomes irresponsibility

This is *coordination-first epistemics*: an inference regime in which preserving large-scale cooperative behavior is prioritized over local epistemic accuracy, because errors in coordination impose nonlinear, system-level costs.

5.2 When Coordination-First Is Rational

Coordination-first epistemics is not purely pathological. There are conditions under which prioritizing coordination over local truth-seeking is the correct move:

Irreversible collective action. When a group must commit to a course of action that cannot be easily reversed (vaccination campaigns, infrastructure investment, institutional reform), premature epistemic revision can be catastrophic. The costs of coordination failure exceed the costs of local epistemic error.

Trust as public good. When institutional trust enables beneficial cooperation across the population, and when that trust is fragile and hard to rebuild, protecting trust may justify some epistemic sacrifice. The epistemically correct statement "our models have significant uncertainty" can destroy the coordination that makes the models useful.

Adversarial information environments. When external actors are actively trying to destabilize coordination through misinformation, tightening epistemic standards (requiring higher confidence before public revision) can be a rational defense. This is not denying uncertainty internally; it is managing its public expression.

Nested timescales. When short-term epistemic gains would undermine long-term epistemic capacity (for example, by destroying the institutions that produce reliable knowledge), coordination-first reasoning at the institutional level can be epistemically optimal at longer timescales.

The point is not that coordination-first epistemics is always wrong, but that it has scope conditions. The pathology emerges when coordination-first reasoning persists beyond its proper domain—when it becomes the default rather than a contextual response to genuine coordination constraints.

5.3 Coalitional Enforcement Dynamics

Coordination-first epistemics generates structural features that parallel those found in high-commitment groups, including religious communities. To be clear: the comparison is to *coordination-enforcement mechanisms*, not to metaphysical content. The claim is not that scientific institutions become supernatural or dogmatic in their beliefs, but that any institution optimizing for coordination under uncertainty will develop enforcement dynamics structurally similar to those that stabilize high-commitment coalitions—because both face the same underlying game-theoretic problem: maintaining cooperation when internal states are unobservable. This does not imply false content; it concerns the *function* of public assent under uncertainty.

Credal compression. Complex, multidimensional knowledge is compressed into simple, shareable formulations. These function less as truth claims than as coordination signals: affirming the creed demonstrates coalition membership.

Costly signaling. Because belief is unobservable, groups stabilize cooperation through costly public displays—ritualized affirmations, taboos, purity tests—that function as proxies for loyalty [Zahavi, 1975,Todd, 2025a]. The costlier the signal, the more reliable its indication of commitment.

Heresy punishment. Punishing deviation serves two functions: it deters defection, and it signals the punisher's own commitment. Under coordination-first epistemics, enforcement becomes a competitive display of loyalty.

Evidence resistance. If belief functions as coordination signal rather than truth-tracker, then evidence against group beliefs can *strengthen* commitment by raising the cost of continued profession. Those who maintain belief despite counterevidence demonstrate precisely the commitment that coalition membership requires.

These are not pathologies. They are predictable consequences of optimizing for coordination under uncertainty. The "irrationality" of ideological belief is rationality at the coalition level (Figure 3).

Model details. The simulation is a minimal toy model illustrating the core dynamic: $N = 100$ agents choose between truth-seeking and coordination-first stances. Coordination pressure increases with system maturity. Each agent faces a trade-off: truth-seeking provides a fixed benefit, while coordination-first avoids a dissent penalty that scales with how coordinated the population already is (dissent cost $\propto$ pressure $\times$ coordination^2). Agents stochastically switch types based on this trade-off. Panels A–B show 20 independent runs per condition; Panel C shows 25×25 grid of parameter combinations with 5 runs each, totaling 3,125 simulation runs for the phase diagram. The qualitative pattern—transition to coordination-first under high stakes—is robust; the model is intended to illustrate the mechanism, not to quantitatively fit any empirical system.

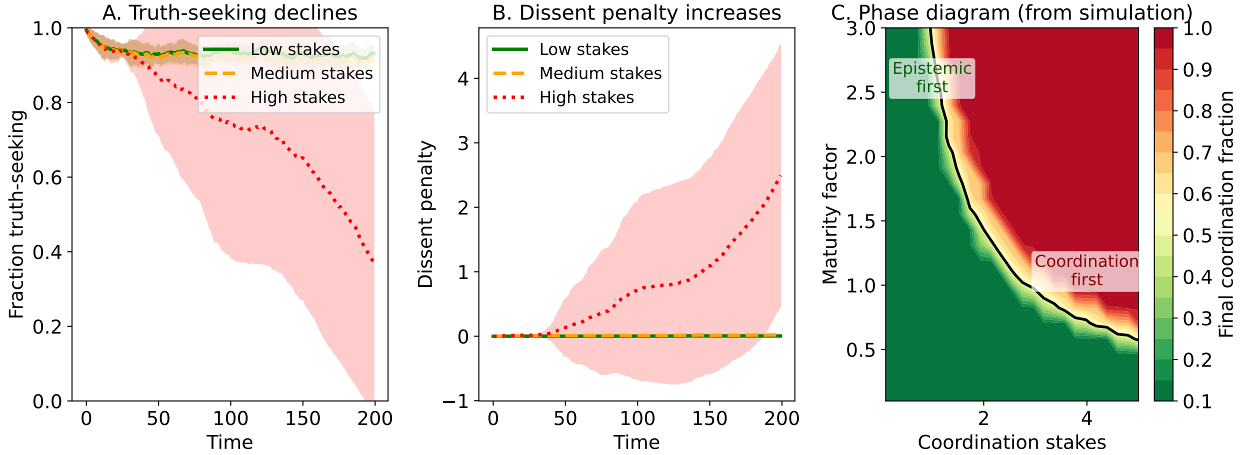

Figure 3: **Agent-based simulation of coordination-first transition** (illustrative toy model). (A) As coordination stakes increase, the fraction of truth-seeking agents declines—high stakes select for coordination-first behavior. Shaded regions show ±1 standard deviation across 20 independent runs. (B) Dissent penalties increase over time, especially under high stakes. (C) Phase diagram derived from simulation: each point shows mean final coordination fraction across 5 runs at that (stakes, maturity) parameter combination. The black contour marks the 50% transition boundary. *Note*: The qualitative pattern (transition to coordination-first under high stakes) is robust across parameter variations; quantitative thresholds depend on specific payoff functions.

5.4 The Exploitation Trap

Put together, these dynamics form a closed loop:

1. Systems reward local exploitation
2. Local exploitation collapses global dimensionality
3. Dimensional collapse generates anomalies
4. Anomalies are moralized locally
5. Moralization preserves local incentives
6. Exploration never reopens

This is the exploitation trap: a stable configuration in which the system actively suppresses awareness of its own trajectory by rewarding agents who mistake local optimality for global virtue.

The trap is not maintained by conspiracy or malice. It is maintained by incentive gradients that make myopia adaptive and critique costly. Agents who see the global pattern are penalized; agents who optimize locally are rewarded. The system selects for its own perpetuation (Figure 4).

6 Relation to Existing Accounts

The framework developed here intersects with several established literatures in philosophy of science. Situating it precisely clarifies both what it adds and what it presupposes.

6.1 Underdetermination and Unconceived Alternatives

The Duhem-Quine thesis establishes that theories are never tested in isolation [Duhem, 1906, Quine, 1951]. Stanford's problem of unconceived alternatives deepens this: not only

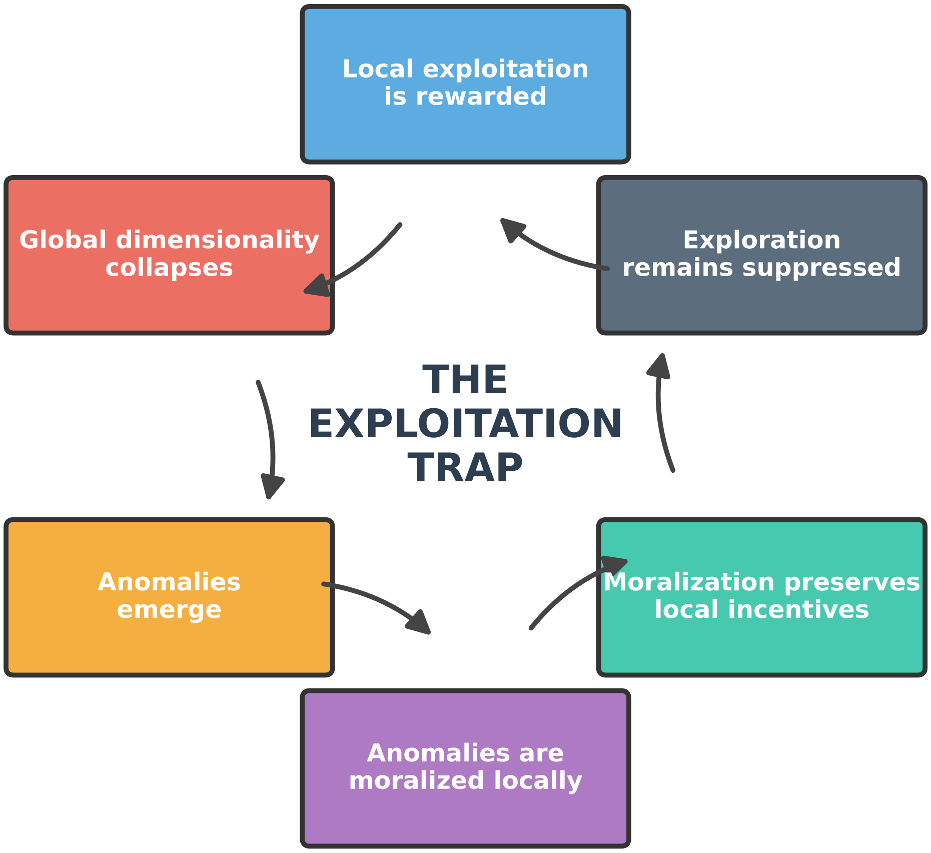
Local exploitation is rewarded
Global dimensionality collapses
Exploration remains suppressed
THE EXPLOITATION TRAP
Anomalies emerge
Moralization preserves local incentives
Anomalies are moralized locally

Figure 4: **The exploitation trap.** A closed loop in which local exploitation collapses global dimensionality, generates anomalies, triggers moralization, preserves local incentives, and suppresses exploration. The trap is maintained by selection on local payoffs, not by conspiracy.

can we always adjust auxiliaries, but history shows we systematically fail to conceive of the theories that will replace our current ones [Stanford, 2006]. The present framework adds a *dynamical* dimension: underdetermination is not static but intensifies as systems mature. Early-stage paradigms face underdetermination but have slack to explore alternatives; late-stage paradigms face underdetermination *plus* coordination lock-in that makes exploration costly. The exploitation trap is what happens when Stanford's problem meets institutional inertia.

Van Fraassen's constructive empiricism holds that science aims at empirical adequacy rather than truth [van Fraassen, 1980]. The coordination-first framework is compatible with this view but adds that empirical adequacy itself becomes harder to assess as error-model complexity grows. In late-stage paradigms, "saving the phenomena" becomes indistinguishable from "managing the residuals."

6.2 Social Epistemology and Network Effects

Goldman's social epistemology examines how social processes affect knowledge production [Goldman, 1999]. Longino emphasizes that objectivity is achieved through social criticism and diverse perspectives [Longino, 1990]. Kitcher analyzes how epistemic communities can go wrong when incentives misalign [Kitcher, 2001]. The present framework contributes a specific mechanism: *coordination pressure* can systematically suppress the diversity that social epistemologists identify as essential.

Zollman's work on network epistemology is particularly relevant. He shows that communication structure affects belief dynamics: more connected networks converge faster but may converge on falsehoods [Zollman, 2007]. He also demonstrates that transient diversity—maintaining disagreement longer—can improve long-run accuracy [Zollman, 2010]. The exploitation trap predicts exactly the opposite tendency: mature systems will increase connectivity and suppress diversity to reduce coordination costs, sacrificing long-run accuracy for short-run stability.

O'Connor and Weatherall analyze how false beliefs spread through epistemic networks, emphasizing that propaganda and misinformation exploit network structure [O'Connor and Weatherall, 2019]. The present framework suggests that even without external manipulation, internal dynamics of coordination pressure can produce similar effects: dissent becomes costly, conformity becomes rewarded, and the network converges on coordinated belief regardless of accuracy.

6.3 Values in Science

Douglas argues that values inevitably enter scientific reasoning, particularly in managing uncertainty and determining acceptable error rates [Douglas, 2009]. Biddle and Kukla examine how risk distributions across epistemic communities affect what gets investigated and how [Biddle and Kukla, 2017]. The coordination-first framework extends this analysis: when preserving coordination becomes the dominant value, it systematically biases which uncertainties are tolerable (those that don't threaten coordination) and which are not (those that do).

This is not a claim that coordination-first reasoning is irrational. Douglas and others have established that non-epistemic values legitimately enter science. The point is that coordination values can *dominate* epistemic values under specific structural conditions—and that this domination is predictable from the exploration-exploitation lifecycle.

6.4 Models and Idealization

The philosophy of models provides crucial background. Weisberg distinguishes Galilean idealization (simplifying to reveal essential structure), minimalist idealization (using the simplest model that works), and multiple-models idealization (using several incompatible models for different purposes) [Weisberg, 2007]. Morgan and Morrison analyze models as "mediators" between theory and world [Morgan and Morrison, 1999]. Godfrey-Smith examines how model-based science proceeds through strategic simplification [Godfrey-Smith, 2006].

The present framework adds that idealization has a lifecycle. Early-stage models are exploratory: many idealizations are tried, fit is loose, the goal is insight. Late-stage models are exploitative: idealizations become entrenched, fit is tight, the goal is prediction within the established framework. The same idealization that enabled insight can become a prison when the community loses the capacity to question it. Cartwright's observation that "the laws of physics lie"—that our best models systematically misrepresent—applies most strongly when models have been optimized to the point of brittleness [Cartwright, 1983].

6.5 Scientific Realism and Its Limits

Laudan's pessimistic meta-induction argues that since past successful theories were false, we should not assume current successful theories are true [Laudan, 1981]. Fine's natural ontological attitude recommends avoiding both realism and anti-realism, taking science's claims at face value without metaphysical inflation [Fine, 1984]. Chakravartty defends a selective realism that commits only to those aspects of theories that are genuinely explanatory [Chakravartty, 2007].

The present framework is compatible with all of these positions on the realism question. It makes a different point: regardless of whether mature theories are approximately true, they become *epistemically sticky*—hard to revise even when revision is warranted. This is not a claim about truth but about the dynamics of belief revision in communities under coordination pressure. Hacking's emphasis on intervention and manipulation as the basis for realism is relevant here: late-stage paradigms may lose the capacity for the kind of experimental probing that would ground realist commitments [Hacking, 1983].

Chang's historical work on temperature measurement shows how scientific progress often requires tolerating contradictions, exploring dead ends, and maintaining pluralism [Chang, 2004]. The exploitation trap predicts that these capacities will erode as paradigms mature. Chang's "complementary science"—the recovery of abandoned lines of inquiry—is precisely what coordination-first epistemics suppresses.

7 A Diagnostic for Unfalsifiability

7.1 The Duhem-Quine Problem, Quantified

The classical Duhem-Quine thesis observes that theories are never tested in isolation [Duhem, 1906, Quine, 1951]. Any empirical test involves auxiliary hypotheses about instruments, initial conditions, and background assumptions. When data disagree with prediction, one can always "save" the core theory by revising auxiliaries.

This is philosophically correct but practically vague. It does not tell us *when* auxiliary revision becomes pathological. Every mature theory requires some auxiliary adjustment; that is normal science. The question is: when does adjustment become evasion?

I propose a quantitative diagnostic:

> **A system enters an unfalsifiable regime when explaining discrepancies requires adding more degrees of freedom to the error model than the data summary gains in informational constraints.**

That is: unfalsifiability emerges when the number of plausible "systematics knobs" grows faster than the number of independent observational constraints.

Box 2: Operational Recipe for the DOF Diagnostic

Error-model degrees of freedom = number of independently tunable nuisance parameters (calibration offsets, selection corrections, systematic budgets, pipeline choices) that can be adjusted to reconcile theory with data.

Informational constraints = number of independent summary statistics added by new observations that genuinely constrain the theory—i.e., reduce the posterior volume or improve out-of-sample prediction (not merely extend the domain of application at fixed precision).

Diagnostic: Compute the ratio R = (error-model DOF)/(informational constraints) for each round of anomaly accommodation.

- $R < 1$: Progressive—new data constrain the theory more than error models expand.
- $R \approx 1$: Stagnant—theory and error model grow in lockstep; no net evidential gain.
- $R > 1$: Degenerating—error models absorb residuals faster than data constrain them.

When R persistently exceeds unity, the paradigm has entered an unfalsifiable regime. This does not mean the paradigm is wrong; it means decisive adjudication between the paradigm and alternatives has become structurally impossible.

In information-geometric terms, "informational constraints" correspond to independent directions along which the data increase the rank of the Fisher information matrix—directions where nearby hypotheses become distinguishable. "Error-model degrees of freedom" correspond to directions in which the model can be adjusted without changing the predicted distribution over observables. Unfalsifiability emerges when the latter space grows faster than the former: the model's flexibility in explaining residuals outpaces the data's capacity to distinguish between explanations.

To prevent caricature: "constraints" are not simply "number of datasets" or "number of measurements," but *independent directions that reduce posterior volume*—dimensions along which new data genuinely narrow the space of viable theories. Similarly, "DOF" are not "any parameter in the pipeline," but *independently tunable directions that preserve fit on existing summaries*—slack directions or non-identifiabilities. A parameter that is tightly constrained by existing data does not count as a free DOF; only parameters that can absorb new residuals without degrading existing fit contribute to error-model flexibility. This connects to standard notions of effective number of parameters, profile likelihood, and identifiability analysis. The diagnostic does not require exact counting; it requires recognizing when the *growth rate* of error-model slack exceeds the growth rate of genuinely constraining information.

7.2 The Mechanism

Consider a mature scientific paradigm under increasingly precise observation. Residuals appear: small discrepancies between prediction and data. These can be explained in two ways:

1. **Extend the core theory**: Add new physical mechanisms, modify fundamental assumptions, revise the ontology.
2. **Extend the error model**: Add systematic corrections, expand nuisance parameters, invoke selection effects, refine calibration.

Option 1 is exploration: it reopens the high-dimensional possibility space. Option 2 is exploitation: it preserves the existing framework by absorbing residuals into an expanded error budget.

In early-stage paradigms, Option 1 is cheap and Option 2 is expensive (there is no elaborate pipeline to protect). In late-stage paradigms, the costs reverse: Option 1 threatens the entire inference infrastructure, while Option 2 preserves it.

The result is that residuals are systematically routed into error models. Each routing is locally rational—it is the minimum-cost explanation. But the cumulative effect is that the theory becomes immunized against disconfirmation.

7.3 Connection to Lakatos

This diagnostic operationalizes Lakatos's distinction between progressive and degenerating research programmes [Lakatos, 1970]. A progressive programme generates novel predictions that are subsequently confirmed. A degenerating programme survives by *ad hoc* adjustments that accommodate anomalies without generating new predictions.

The degrees-of-freedom criterion provides a metric: count the free parameters added to accommodate each anomaly, and compare to the new constraints those accommodations explain. When the ratio exceeds unity, the programme is degenerating. When it becomes much greater than unity, the programme has entered an unfalsifiable regime.

This is not Popperian failure in the simple sense [Popper, 1959]. The theory is not "refuted"; it is *immune to decisive test* because the space of possible error-model adjustments is larger than the space of possible observations (Figure 5).

7.4 Why This Is Structural, Not Sociological

The unfalsifiability criterion is not an accusation of bad faith. It describes a structural property of inference under dimensional constraint.

When:

- The true generative process is high-dimensional
- Observation accesses only low-dimensional projections
- Multiple error sources contribute to residuals
- The mapping from theory to observation involves complex pipelines

then decisive falsification becomes impossible regardless of the intentions of practitioners. Many distinct high-dimensional causes project to the same low-dimensional effect. "Falsify

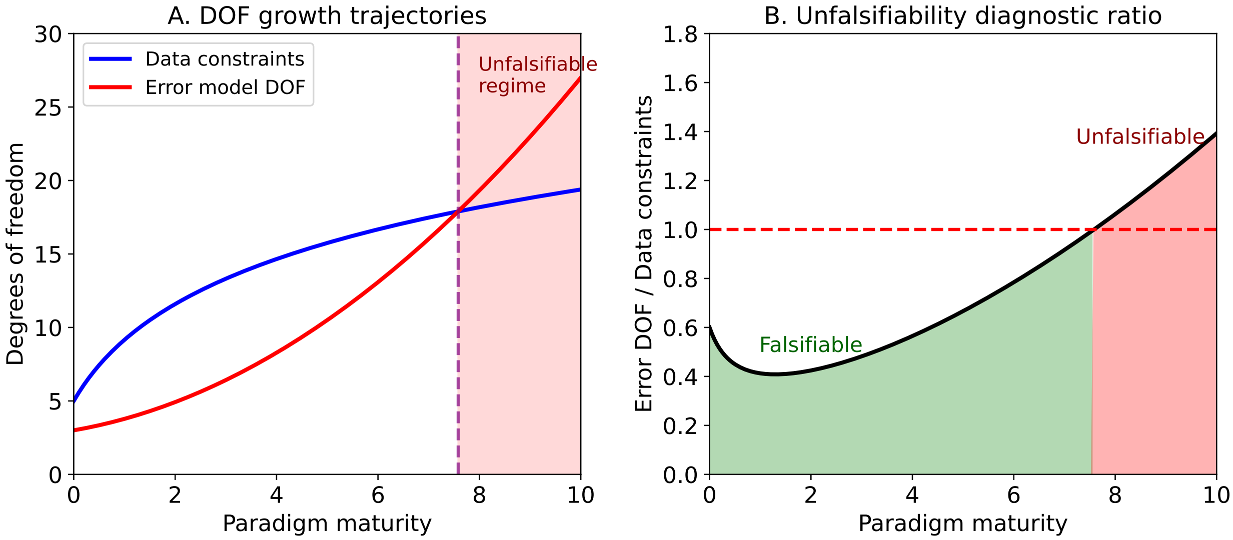

Figure 5: **The degrees-of-freedom criterion for unfalsifiability** (schematic). (A) As paradigms mature, data constraints grow slowly while error-model degrees of freedom grow faster, eventually crossing. Beyond the crossover, anomalies can always be absorbed. (B) The ratio of error DOF to data constraints provides a diagnostic: when the ratio exceeds 1, decisive falsification becomes structurally impossible.

the model" ceases to be well-defined.

The appropriate scientific response is not to pretend falsifiability still applies, but to develop frameworks that explicitly model projection and information loss as part of the theory-data interface.

7.5 Falsifiability Was Always an Idealization

The DOF diagnostic operationalizes what Duhem and Quine established philosophically: no theory is tested in isolation, and auxiliary hypotheses can always be adjusted to save the core. What the dimensional framework adds is that this is not a peculiarity of certain paradigms but a *structural feature of any observation process where* $D_{\text{sys}} > D_{\text{obs}}$.

This condition obtains universally. Every real-world system has more internal degrees of freedom than any measurement can access. Classical Popperian falsificationism describes an idealization that never existed in practice. Real science operates through model comparison, error-model adjustment, and consensus formation—all of which are subject to the DOF diagnostic.

The implication is not that science is broken, but that we should stop asking "is this paradigm falsifiable?" and start asking "what is the current DOF ratio?" Early paradigms have low R: data constrain faster than error models expand. Mature paradigms have $R \approx 1$: stagnation. Late paradigms have $R > 1$: structural unfalsifiability. This is a continuous variable, not a binary property.

The free will debate is not anomalous in being unfalsifiable. It is anomalous only in having installed its error model at the foundation rather than accumulating it through anomaly accommodation. Most paradigms drift into unfalsifiability; hard determinism started there.

Biology, in particular, operates permanently in the $D_{\text{sys}} \gg D_{\text{obs}}$ regime [Todd, 2025b, Todd, 2025c]. Living systems maintain internal complexity that exceeds observational access by design—this is part of what makes them alive. The persistent difficulty of reducing biology to mechanism, the tendency of organisms to surprise us, the failure of genomics to predict phenotype—these are not temporary limitations awaiting better technology. They reflect a genuine feature of the subject matter. The DOF diagnostic does not condemn biology; it explains why biological science requires different epistemic standards than physics at solvable limits.

8 Case Studies

The following case studies are not offered as comprehensive empirical demonstrations, but as *interpretive mappings* that illustrate how the same dynamical pattern could manifest across substrates. Each is meant to show that the framework generates recognizable predictions in concrete domains, not to provide definitive evidence that those domains instantiate the framework.

8.1 The Free Will Debate: Coordination-First Epistemics in Philosophy of Mind

The debate over free will provides perhaps the clearest example of coordination-first epistemics operating within science itself. This case study is unusual because it concerns not a specific scientific paradigm but the *meta-level assumptions* that constrain which positions count as "serious" within scientific discourse.

The phenomenological evidence. Every person has direct, continuous experience of making choices. You deliberate, weigh options, and act. The feeling of agency is not occasional or subtle—it is the constant texture of conscious life. If any evidence counts as data, this does.

The "sophisticated" position. A common posture in public-facing neuroscience and philosophy of mind—particularly among hard determinists and illusionists—holds that this experience is illusory. The argument runs: physics describes a world of deterministic laws (or deterministic laws plus quantum randomness). Neither determinism nor randomness leaves room for genuine agency. Therefore, the feeling of choice must be an epiphenomenon—a story the brain tells itself, generated by neural processes the agent does not observe. (Compatibilists offer a different resolution, redefining "free will" to be compatible with determinism; this sidesteps rather than addresses the phenomenological question.)

The social dynamics. Crucially, this position is not merely argued; it is *enforced through social signaling.* To believe in free will marks one as naive, scientifically unsophisticated, or insufficiently rigorous. The graduate student who expresses skepticism about hard determinism is gently corrected. The philosopher who defends libertarian free will is tolerated but not taken fully seriously. The neuroscientist who suggests that agency might be real is assumed to be confused about physics.

This is classic coordination-first epistemics:

- The "correct" position is not primarily established by evidence (the phenomenological evidence runs the other way)

- Dissent signals coalition defection rather than intellectual engagement
- Holding the naive view incurs social costs; holding the sophisticated view signals in-group membership
- The debate has effectively "ended" within serious discourse, despite never being resolved

The DOF diagnostic. Apply the unfalsifiability criterion. The anomaly is direct phenomenological experience of agency. The error-model response is: "This feeling is an illusion generated by neural processes you do not have introspective access to."

This response can absorb *any* phenomenological evidence. No matter how vivid, how consistent, how universal the experience of choice, the reply is always available: "Yes, the illusion is very convincing, but it remains an illusion." The error model ("illusory feeling generated by hidden neural processes") has more degrees of freedom than the phenomenological data can constrain. By the diagnostic of Section 7, the hard determinist position has entered an unfalsifiable regime—not because it is false, but because it has become structurally immune to the primary evidence against it.

To make this concrete, consider the error-model degrees of freedom available to the determinist:

1. "Introspection is unreliable" (explains away first-person reports)
2. "The feeling of agency evolved for fitness, not accuracy" (explains away the universality)
3. "Quantum randomness $\neq$ agency" (blocks the indeterminism escape route)
4. "Unconscious neural processes precede conscious awareness" (Libet-style arguments)

5. “Compatibilist redefinition” (redefine “free will” to mean something determinism-compatible)

6. “Complexity creates the illusion of choice” (emergence as explanation-stopper)

Each of these is independently adjustable. Against this, what informational constraints does phenomenological evidence provide? Essentially one: the consistent report that agency feels real. But this single constraint is absorbed by error-model knob (1) or (2). The ratio $R = 6/1 = 6 \gg 1$. By the diagnostic, the paradigm is deeply in the unfalsifiable regime—not because determinism is wrong, but because no phenomenological evidence could ever shift the conclusion. The error model was installed at the foundation, not accumulated through anomaly accommodation.

The dimensional interpretation. One might attempt a reconciliation: perhaps “free will” names causal degrees of freedom that the agent accesses from the inside but that external observers cannot measure. What appears as randomness from outside (because $D_{\text{sys}} \gg D_{\text{obs}}$) might be experienced as choice from inside. This would make agency real but unmeasurable by third-person methods—a projection artifact rather than an illusion.

But notice how easily this move is absorbed: “Fine, it’s unmeasurable determinism, but it’s still determinism.” The paradigm protects itself by definitional expansion. Any evidence of agency can be reframed as evidence of complex-but-deterministic processes we happen not to observe. This is precisely the exploitation trap: the framework has become flexible enough to accommodate any data.

The meta-point. This analysis does not resolve the free will debate. It does not claim that libertarian free will is correct, or that hard determinism is wrong. It claims something more limited but perhaps more important:

Regardless of the metaphysics, the sociology of the free will debate follows exactly the pattern this paper describes.

- Phenomenological evidence (direct experience of choice) is systematically dismissed

- The dismissal is enforced through social costs, not through counter-evidence
- The "sophisticated" position requires believing that one's most basic experience is illusory
- The paradigm has become unfalsifiable by the DOF criterion
- Dissent is treated as failure to understand rather than as legitimate intellectual disagreement

Whether or not free will exists, the scientific consensus against it exhibits coordination-first dynamics. The question for the epistemically serious observer is: if this is how the debate is conducted, how much weight should the consensus carry? The framework suggests that consensus achieved through coordination-first mechanisms is evidence of paradigm lock-in, not evidence about the underlying question.

The biological question. For a journal concerned with the philosophy of biology, there is a deeper issue: *when did agency first emerge?*

The implicit assumption in most free will debates is that agency—if it exists at all—is a late evolutionary development, perhaps unique to humans or at most shared with a few cognitively sophisticated animals. But consider the alternative: perhaps agency is coextensive with life itself.

A bacterium performing chemotaxis is not merely executing a stimulus-response algorithm. It is integrating information across multiple gradients, maintaining internal state, and selecting among behavioral options in ways that affect its survival. A cell deciding whether to divide, differentiate, or undergo apoptosis is making a choice with consequences. A plant redirecting growth toward light is exercising something that looks remarkably like preference.

The language that practicing biologists use is telling. Recent work recreating the ancient endosymbiotic event that produced eukaryotic cells—the moment a bacterium took up residence inside an archaeon, eventually becoming the mitochondrion—found that the researchers could not avoid agential descriptions [Giger et al., 2024, Herring, 2025]. The archaeal host cells were described as "welcoming" bacterial partners rather than attacking them. The bacteria "evaded" immune responses. Some interactions failed because organisms "didn't want to live together." These are not careless metaphors; they reflect the structure of the phenomenon. The organisms were making choices with consequences for their survival, integrating information, and behaving in ways that selection has shaped but not fully determined. If agency is an illusion, it is an illusion that even careful scientists cannot stop perceiving when they watch microbes interact.

If we grant phenomenological evidence any weight—and the determinist position requires dismissing it entirely—then the question becomes: at what point in the tree of life do we draw the line? The options are:

- **Only humans have agency**: This requires special pleading. What is it about human neural architecture that generates genuine choice when no other physical system can? This is dualism by another name.
- **Some animals have agency**: When did it emerge? What was the first species to make a genuine choice? The question is unanswerable because there is no principled boundary.
- **All living things have agency**: Agency is not a late evolutionary add-on but a fundamental feature of what it means to be alive—perhaps even constitutive of life itself.
- **Nothing has agency**: The determinist position. But this requires dismissing universal phenomenology as universal illusion, which is the coordination-first move we have been analyzing.

The dimensional framework suggests a resolution: agency is what high-dimensional causal structure looks like from the inside. Any system with sufficient internal degrees of freedom—

sufficient distance between D_{sys} and D_{obs}—will have aspects of its dynamics that are inaccessible to external measurement but accessible to the system itself. Life, almost by definition, maintains internal complexity that exceeds what external observers can measure. Agency may be the rule for living systems, not the exception.

This reframes the free will debate from a question about human consciousness to a question about biology: not "do humans have free will despite being physical systems?" but "is agency a fundamental feature of life that we have been systematically misdescribing as mechanism?"

The boundary problem and implications for science. If agency is widespread in living systems, we face difficult questions. Does kinesin "decide" where to walk along a microtubule? Do mitochondria "decide" how much ATP to produce and what signals to emit? What kind of awareness, if any, does a nucleus have? Where do we draw the line, and how would we test it?

These questions are hard. But the determinist position does not solve them—it merely denies the phenomenon requires explanation. The dimensional framework suggests a different approach: agency is not binary but a matter of degree, correlating with the gap between internal complexity and external measurability.

A system has agency to the extent that $D_{\text{sys}} \gg D_{\text{obs}}$—to the extent that its internal dynamics exceed what external observation can characterize. By this criterion:

- A bacterium has modest agency: many internal degrees of freedom, but we can predict much of its behavior from environmental inputs.

- A kinesin motor has minimal agency: its behavior is well-characterized by a small number of biochemical parameters. There is little gap between what the molecule "knows" and what we can measure.

- A neuron has substantial agency: its firing depends on integration across thousands of synapses, internal metabolic state, and stochastic processes we cannot fully observe.

- A human has extensive agency: the gap between internal state and external measurement is vast.

This is *not* a claim about consciousness or subjective experience in the human sense. It is a claim about *predictability gaps*—how much of a system's behavior is determined by factors external observers cannot access. Agency, on this view, is the name for the causal work done by those inaccessible factors.

To be precise: this is an account of *epistemic agency*—observer-relative irreducibility—rather than *organizational autonomy* in the sense of self-maintenance or goal-directedness. The claim is not that projection gaps constitute agency in some deep metaphysical sense, but that they explain why agential description remains stable and useful in biology: organisms maintain internal complexity that exceeds external measurement, and this gap is what agential language tracks. Whether there is "something it is like" to be a bacterium is a separate question; what we can say is that the bacterium's behavior will never be fully predictable from outside, and this unpredictability has the functional signature of choice.

What does this mean for biology as a science? It suggests that the mechanistic program—explaining living systems entirely in terms of externally measurable components and their interactions—may face principled limits. Not because organisms are magical, but because they maintain internal complexity that exceeds observational access. The persistent difficulty of reducing biology to physics, the tendency of organisms to be "more than the sum of their parts," the failure of genomics to fully predict phenotype—these may not be temporary limitations awaiting better technology. They may reflect a genuine feature of the subject matter: living systems have more degrees of freedom than external measurement can capture.

This does not mean we cannot do science on living things. It means we may need conceptual frameworks that treat organisms as agents rather than machines—not as a metaphor, but as an accurate description of what they are. The resistance to this move is itself a case study in coordination-first epistemics: mechanism is the "serious" position, and treating organisms as genuine agents marks one as insufficiently rigorous.

The limits of mathematical description. There is a deeper possibility that deserves serious consideration: agency may be the kind of thing that science, as currently constituted, is simply not equipped to describe.

Science operates through external observation and mathematical modeling. It describes systems from the outside, in terms of measurable quantities and formal relationships. This method has been extraordinarily successful for physics, chemistry, and much of biology. But its success does not prove that everything real is externally measurable and mathematically describable.

Agency—if it is real—is precisely what a system does that *cannot* be captured by external measurement. It is the causal work done by factors inaccessible to the observer. If this is correct, then agency is not a failure of current scientific understanding awaiting future resolution. It is a phenomenon that falls outside the scope of third-person mathematical description in principle.

This does not make agency supernatural or unreal. It makes it a different kind of thing than what physics describes—not because it violates physical laws, but because it is constituted by the internal dynamics that physical laws, applied from outside, cannot fully specify. Mathematics describes the constraints on what can happen; agency is what happens within those constraints, from the inside.

Consider what genuine agency would mean informationally. If I have even a sliver of free will—if my decisions are not fully determined by prior states—then when I make a choice, I generate new information. Something enters the world that was not implicit in what came before. But mathematics describes relationships between quantities, transformations of existing information, the unfolding of what was already there in different form. How would you write an equation for genuine novelty? The output would not be a function of the inputs. The formalism would have a gap precisely where the agency occurs.

The determinist insists: “Nothing is truly new; it only appears new because you cannot see all the prior causes.” But this is the claim in question, not its proof. It is the assertion that the mathematical description is complete, offered as evidence that the mathematical description is complete. If even one bit of genuinely new information enters the world through choice—if any agent anywhere ever does something not implicit in the prior state of the universe—then the deterministic equations are incomplete in principle, not merely in practice.

The hard determinist response—"if it cannot be mathematically described, it does not exist"—is not a scientific conclusion. It is a metaphysical commitment, adopted as axiom and enforced through social dynamics rather than evidence. The demand that we dismiss the evidence of our own experience because it does not fit the formalism represents a particular philosophical stance, not the inevitable conclusion of rigorous inquiry.[3]

8.2 Scientific Paradigms: The Case of Cosmology

The standard cosmological model (ΛCDM) provides a striking example of the exploration-to-exploitation transition.[4]

Exploration phase (1920s–1990s): Multiple competing cosmologies. Steady-state versus Big Bang. Open, closed, or flat universe. Matter-dominated or radiation-dominated. Rapid conceptual change driven by new observations.

Consolidation phase (1990s–2000s): Precision measurements (CMB anisotropies, supernova distances, baryon acoustic oscillations) converge on a six-parameter model. ΛCDM becomes "the standard model of cosmology." Inference pipelines are built; simulation codes are calibrated; careers are structured around the paradigm.

[3] One might characterize this more sharply: the insistence that the mathematical description is complete, offered as evidence that the mathematical description is complete, has the structure of dogma rather than discovery. The social-rationality dynamic is notable: sophistication is partly signaled by treating phenomenology as epistemically negligible, which creates selection pressure for dismissing agency regardless of the underlying question.

[4] There is a deeper parallel: the universe itself instantiates the lifecycle. The Big Bang represents maximal exploration—high energy density, symmetry breaking, all configurations accessible. Structure formation is consolidation: matter clumps, galaxies form, stable configurations emerge. The current epoch is exploitation: stars burn hydrogen efficiently, planets orbit stably, life extracts free energy from thermal gradients. Heat death is the ultimate brittleness: maximum entropy, no gradients to exploit, no exploration possible. The framework applies not only to the paradigm studying the universe but to the universe as a complex system.

Exploitation phase (2010s–present): Tensions emerge. The Hubble constant measured locally disagrees with the value inferred from early-universe observations. The amplitude of matter clustering shows similar discrepancies. Large-scale anomalies persist across independent probes.

Response: Rather than reopening fundamental assumptions, the field has primarily responded by:

- Expanding systematic error budgets
- Invoking unmodeled selection effects
- Adding parameters to accommodate discrepancies
- Treating tensions as "probably systematics" pending further investigation

Each response is locally rational. But the cumulative effect is that ΛCDM has become difficult to falsify—not because it is correct, but because the error-model space has expanded to absorb anomalies.

If the diagnostic applies—and cosmologists may reasonably contest whether it does—then the number of "systematics stories" invoked to explain tensions may exceed the number of independent constraints those tensions provide. This is not a claim that ΛCDM is wrong. It is a claim that decisive adjudication may have become structurally difficult, and that the field should explicitly assess whether the conditions for the exploitation trap obtain.

8.3 Economic Institutions: Monopolistic Consolidation

Markets exhibit the same lifecycle.

Exploration phase: Many firms experiment with diverse strategies. Entry is easy; failure is common; innovation is rewarded. High effective dimensionality in the strategy space.

Consolidation phase: Some strategies prove successful. Network effects, economies of scale, and learning curves amplify winners. Standards emerge; platforms consolidate.

Exploitation phase: A few dominant firms capture most value. Strategy space collapses to optimization within established platforms. Innovation shifts from product to rent extraction. New entrants face not just competition but structural barriers (regulatory capture, platform control, data moats).

Brittleness: The system becomes efficient but fragile. Shocks that distributed systems would absorb become systemic risks. The 2008 financial crisis exemplified this: consolidation produced efficiency gains but also correlated failure modes.

The moral narrative ("greedy bankers," "corrupt regulators") identifies real local failures but obscures the structural attractor. Replacing actors without changing the incentive landscape produces the same consolidation.

8.4 Biological Senescence: Organism-Level Dimensional Collapse

Aging can be understood as physiological dimensional collapse—and, more provocatively, as *epistemic frustration at the cellular level* [Ferrucci et al., 2022].

Youthful exploration: High regenerative capacity. Flexible regulatory systems. Redundant pathways. The organism can recover from insults because it has degrees of freedom to reallocate. Stem cell populations are diverse; epigenetic states are plastic; proteostatic networks maintain flexibility; immune repertoires are broad. The system can "learn"—it can mount novel responses to novel challenges.

Mature exploitation: Regulatory systems become optimized for efficiency rather than flexibility. Redundancy is pruned. Control becomes more centralized and stereotyped. Stem cell populations narrow toward lineages that have proven useful; epigenetic marks accumulate, constraining gene expression to established patterns; protein quality control prioritizes familiar proteins over novel synthesis. The organism becomes very good at what it has been doing, at the cost of the capacity to do something else.

Senescent brittleness: The system loses resilience. Small perturbations that would once have been absorbed now cascade. Regulatory networks become rigid. The organism can no longer adapt to novel challenges. Stem cells are exhausted or senescent; epigenetic states are locked; proteostatic capacity is overwhelmed by accumulated damage; immune responses are stereotyped and increasingly auto-reactive.

The DOF diagnostic applies directly. Consider the organism's capacity to respond to a novel pathogen or repair a novel form of damage:

- **Young organism**: Many degrees of freedom available. Stem cells can differentiate along multiple lineages; immune cells can generate novel receptor configurations; metabolic networks can reroute around blockages. The "error model" (repair and adaptation machinery) has high dimensionality relative to the space of possible insults.

- **Aged organism**: Fewer degrees of freedom. Stem cell lineages are committed; immune repertoire is narrowed by clonal expansion of memory cells; metabolic flexibility is reduced by accumulated mitochondrial damage. The error model has collapsed. Novel challenges cannot be accommodated because the response repertoire has been optimized for past challenges.

This is epistemic frustration in a literal sense: the aged organism "knows" how to handle familiar problems (exploitation) but cannot "learn" new solutions (exploration). The high-dimensional optimum—robust health across all possible challenges—does not survive projection into the low-dimensional response repertoire the organism has retained. Any particular insult might be handleable by *some* configuration of cellular machinery, but not by the configuration the organism has locked itself into.

The parallel to coordination-first epistemics is striking. Just as mature paradigms absorb anomalies into error models rather than revising core theory, aged organisms absorb damage into compensatory mechanisms rather than regenerating capacity. Just as coordination-first systems penalize exploration because it threatens stability, aged organisms suppress stem cell proliferation and immune novelty because uncontrolled growth threatens cancer. The exploitation trap is not a metaphor for aging; aging *is* the exploitation trap, instantiated in tissue.

The moral narrative of aging ("decay," "degeneration," "wearing out") obscures this structural logic. Aging is not failure; it is the expected trajectory of a system that has been optimizing for decades. The tragedy is not that the organism breaks down, but that the very success of its early optimization forecloses the flexibility it would need to continue adapting. Youth is not health; youth is slack. Age is not disease; age is lock-in.[5]

8.5 Language and Cultural Diversity

Languages are high-dimensional encodings of local ecology, social structure, and cognitive history. Each language represents a distinct compression of experience into communicable form.

Exploration phase: Human populations diversify across environments. Languages diverge to encode local contingencies. High diversity of representational strategies.

Consolidation phase: Trade networks, empires, and states impose coordination benefits on shared languages. Standardization enables commerce, governance, and mobility.

Exploitation phase: Dominant languages (English, Mandarin, Spanish) capture increasing returns. Education, media, and economic opportunity flow through major languages. Minority languages face declining speaker populations.

Collapse: Languages die. UNESCO classifies a substantial fraction of the world's languages as endangered, with projections suggesting major losses within this century [UNESCO, 2010]. Each extinction represents the loss of a distinct high-dimensional encoding of

[5] The same principle applies to cognitive and skill development. Research on Nobel laureates finds they are nine times more likely than typical scientists to have training in crafts or fine arts, and systematically defy narrow specialization—integrating knowledge across disciplines rather than exploiting a single niche. Early specialization in children produces measurable short-term gains but reduces the high-dimensional exploratory capacity that predicts long-term creative achievement. The selection pressure toward early exploitation (parents, coaches, admissions committees rewarding measurable performance) works against the slack that enables later adaptation.

human experience.

The moral narrative ("cultural imperialism," "linguistic genocide") identifies real dynamics but risks obscuring the structural attractor. Coordination benefits are genuine; people adopt dominant languages for rational reasons. The collapse is not primarily about villains; it is about incentive gradients that make diversity expensive.

9 Implications

9.1 Accepting Geometric Conflict

The first implication is conceptual. If epistemic frustration is structural—if high-dimensional optima genuinely project into contradictory low-dimensional positions—then some disagreements cannot be resolved by evidence or argument. They are geometric, not epistemic.

This does not imply relativism. It implies that the locus of disagreement should shift from "who is right?" to "which constraints are we choosing to satisfy?" That is a political and distributional question, not a truth question.

Recognizing geometric conflict could reduce the moralization of disagreement. Opponents are not necessarily irrational or malicious; they may be optimizing different slices of a frustrated landscape. This does not require agreement, but it does permit a different quality of engagement.

9.2 Designing for Exploration

If exploitation is a dynamical attractor that suppresses exploration, then maintaining adaptive capacity requires deliberate investment. Systems that wish to remain corrigible must:

- Preserve slack (resources not committed to exploitation)
- Maintain diversity (multiple strategies, not one optimum)
- Tolerate failure (low penalties for exploration that doesn't pay off)

- Resist coordination lock-in (modularity, reversibility, exit options)

These are expensive. They look wasteful from an exploitation-optimized perspective. But they are the only known defense against brittleness.

The implication for institutions is that apparent inefficiency may be a feature, not a bug. Redundancy, diversity, and slack are the price of resilience.

9.3 Epistemic Humility Under Dimensional Constraint

The unfalsifiability diagnostic suggests a different relationship to mature paradigms. Rather than treating them as approximately true pending refutation, we might treat them as *effective low-dimensional descriptions whose scope is bounded.*

This is the move physics has made repeatedly:

- Newtonian mechanics is not "wrong"; it is an effective theory valid in a regime

- Thermodynamics is not "wrong"; it is a coarse-grained description that ignores microstates

- Classical electromagnetism is not "wrong"; it is a limit of quantum electrodynamics

The same perspective can apply to any mature paradigm. ΛCDM may be the best available effective description and still hit an inference wall where projection effects dominate. Economic equilibrium models may be useful approximations and still miss high-dimensional dynamics. Medical guidelines may be evidence-based and still encode dimensional compression that loses patient-specific information.

Epistemic humility under dimensional constraint means holding models as tools rather than truths, expecting scope limitations, and remaining alert for the signs of unfalsifiable regimes.

9.4 The Defensive Function of Understanding

A final implication concerns the stakes of this analysis. The dynamics described here—exploration collapse, coordination-first epistemics, epistemic frustration—exist whether or not we study them.

Any system optimizing for engagement, loyalty, or coordination will find these structures through gradient descent. A recommendation algorithm maximizing retention will discover that costly commitment displays predict long-term engagement. A platform optimizing for community will find that embarrassing in-group markers produce stickier coalitions. An institution seeking stability will learn that suppressing dissent reduces coordination costs.

The exploitation emerges from optimization pressure, not from explicit design. This means that understanding these dynamics is not optional; it is defensive. The alternative is not ignorance but asymmetric knowledge: systems that have learned to exploit the structure, and agents who have not learned to recognize the exploitation.

Scientific analysis of high-dimensional social systems is therefore not merely academic. It is prerequisite to recognizing when one's epistemic environment has transitioned from exploration to exploitation, and when apparent irrationality is actually coordination-first rationality operating under constraints one has not perceived.

10 Conclusion

Complex systems do not fail because agents are immoral or ignorant, but because local rationality does not integrate into global coherence on nontrivial manifolds.

The exploration-to-exploitation transition is not a choice; it is a dynamical attractor. Systems that successfully exploit their possibility space naturally compress into lower-dimensional configurations. This compression produces efficiency but reduces resilience. Late-stage systems become coordination-first rather than epistemic-first, because errors in coordination impose nonlinear costs while errors in belief can be absorbed.

Epistemic frustration—the condition where high-dimensional optima project into contradictory low-dimensional representations—explains why disagreement persists despite good faith on all sides. Moral intuitions calibrated for local optimization become misleading at global scales. Dissent is reinterpreted as defection. Uncertainty becomes taboo. The system enters an exploitation trap from which escape requires precisely the exploration that exploitation suppresses.

The diagnostic criterion—unfalsifiability emerges when error-model degrees of freedom grow faster than observational constraints—provides a way to recognize when a paradigm has entered this regime. It is not a claim that the paradigm is wrong, but a claim that decisive adjudication has become structurally impossible.

These patterns appear across domains: scientific paradigms, economic institutions, biological organisms, social coalitions, linguistic diversity. The recurrence suggests a general dynamical principle rather than domain-specific pathology.

The implication is not nihilism. It is geometric realism: recognition that high-dimensional systems under coordination pressure exhibit predictable trajectories, and that moralized narratives of decline obscure rather than illuminate these trajectories. The appropriate response is not blame but design: building systems that maintain exploratory capacity, tolerate diversity, and resist the lock-in that exploitation inevitably seeks.

Understanding these dynamics is itself a form of defense. Systems will find the exploitation gradient whether or not we study it. The question is whether we will recognize when we are inside an exploitation trap—or whether we will moralize our local optimality as virtue while the manifold deforms beneath us.[6]

[6] The attentive reader may note that this paper has been carefully worded to avoid triggering the dynamics it describes. This is not hypocrisy—it is prudent presentation. But readers may draw their own conclusions about what such care suggests. Any account of coordination-first epistemics that did not have to navigate coordination-first constraints in its own presentation would be self-refuting.

References

[Biddle and Kukla, 2017] Biddle, J. B. and Kukla, R. (2017). The geography of epistemic risk. *Kennedy Institute of Ethics Journal*, 27(S2):E71–E103.

[Cartwright, 1983] Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford University Press, Oxford.

[Chakravartty, 2007] Chakravartty, A. (2007). *A Metaphysics for Scientific Realism: Knowing the Unobservable*. Cambridge University Press, Cambridge.

[Chang, 2004] Chang, H. (2004). *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press, Oxford.

[Douglas, 2009] Douglas, H. E. (2009). *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press, Pittsburgh.

[Duhem, 1906] Duhem, P. (1906). La théorie physique: Son objet et sa structure. *Revue de Philosophie*. English translation: The Aim and Structure of Physical Theory, Princeton University Press, 1954.

[Ferrucci et al., 2022] Ferrucci, L. et al. (2022). Time and the metrics of aging. *Nature Aging*, 2:757–762.

[Fine, 1984] Fine, A. (1984). The natural ontological attitude. *Scientific Realism*, pages 83–107.

[Giger et al., 2024] Giger, L. et al. (2024). Evolution of a minimal cell with an inducible phagocytic mode. *Nature*. doi: 10.1038/s41586-024-08335-9.

[Godfrey-Smith, 2006] Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology and Philosophy*, 21(5):725–740.

[Goldman, 1999] Goldman, A. I. (1999). *Knowledge in a Social World.* Oxford University Press, Oxford.

[Grueter et al., 2012] Grueter, C. C., Chapais, B., and Zinner, D. (2012). Evolution of multilevel social systems in nonhuman primates and humans. *International Journal of Primatology*, 33(5):1002–1037.

[Grueter and Lüpold, 2024] Grueter, C. C. and Lüpold, S. (2024). The role of between-group signaling in the evolution of primate ornamentation. *Evolution Letters*, 8(6):927–935.

[Hacking, 1983] Hacking, I. (1983). *Representing and Intervening.* Cambridge University Press, Cambridge.

[Herring, 2025] Herring, Y. S. (2025). Scientists re-create the microbial dance that sparked complex life. Quanta Magazine. Available at: `https://www.quantamagazine.org/scientists-re-create-the-microbial-dance-that-sparked-complex-life-20250102/`.

[Kitcher, 2001] Kitcher, P. (2001). *Science, Truth, and Democracy.* Oxford University Press, Oxford.

[Kuhn, 1962] Kuhn, T. S. (1962). *The Structure of Scientific Revolutions.* University of Chicago Press, Chicago.

[Lakatos, 1970] Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In Lakatos, I. and Musgrave, A., editors, *Criticism and the Growth of Knowledge*, pages 91–196. Cambridge University Press, Cambridge.

[Laudan, 1981] Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science*, 48(1):19–49.

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

[March, 1991] March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, 2(1):71–87.

[Mézard et al., 1987] Mézard, M., Parisi, G., and Virasoro, M. A. (1987). *Spin Glass Theory and Beyond*. World Scientific, Singapore.

[Morgan and Morrison, 1999] Morgan, M. S. and Morrison, M., editors (1999). *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge University Press, Cambridge.

[O'Connor and Weatherall, 2019] O'Connor, C. and Weatherall, J. O. (2019). *The Misinformation Age: How False Beliefs Spread*. Yale University Press, New Haven.

[Popper, 1959] Popper, K. R. (1959). *The Logic of Scientific Discovery*. Hutchinson, London.

[Quine, 1951] Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review*, 60(1):20–43.

[Stanford, 2006] Stanford, P. K. (2006). *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford University Press, Oxford.

[Tainter, 1988] Tainter, J. A. (1988). *The Collapse of Complex Societies*. Cambridge University Press, Cambridge.

[Todd, 2025a] Todd, I. (2025a). Costly signaling and coalition formation across biological scales. Manuscript in preparation.

[Todd, 2025b] Todd, I. (2025b). The limits of falsifiability in high-dimensional systems. *BioSystems*, 258:105608.

[Todd, 2025c] Todd, I. (2025c). Timing inaccessibility and the projection bound. *BioSystems*, 258:105632.

[UNESCO, 2010] UNESCO (2010). Atlas of the world's languages in danger. Available at: `http://www.unesco.org/culture/languages-atlas/`.

[van Fraassen, 1980] van Fraassen, B. C. (1980). *The Scientific Image*. Oxford University Press, Oxford.

[Weisberg, 2007] Weisberg, M. (2007). Three kinds of idealization. *Journal of Philosophy*, 104(12):639–659.

[Zahavi, 1975] Zahavi, A. (1975). Mate selection—a selection for a handicap. *Journal of Theoretical Biology*, 53(1):205–214.

[Zollman, 2007] Zollman, K. J. S. (2007). The communication structure of epistemic communities. *Philosophy of Science*, 74(5):574–587.

[Zollman, 2010] Zollman, K. J. S. (2010). The epistemic benefit of transient diversity. *Erkenntnis*, 72(1):17–35.