Civilizational Metamaterials: Engineering Coordination Under Capability Gradients and Structural Turbulence

David Orban[0009−0004−4954−1147]

Independent Researcher
david@davidorban.com

Abstract. We argue that governance must transition from a normative discipline to an engineering discipline, and we develop a formal framework, inspired by the physics of metamaterials, to make this transition quantitative and testable. Artificial General Intelligence affects civilization primarily by increasing decision velocity while human verification capacity remains bounded. When the cost of validating AI-generated outputs exceeds the expected utility of acting on them, rational agents default to inaction: a stable but catastrophic Nash equilibrium we term the *Freezing Equilibrium*. Drawing on metamaterials, where emergent macro-properties arise from designed microstructure, we develop a phenomenological constitutive law for institutional coordination: $R_{\text{eff}} = \beta \cdot (1 - \rho) \cdot (1 - \tau) \cdot (1 + \gamma\rho\tau)$, where β is the decision branching factor, ρ is provenance fidelity, τ is the verification rate, and γ captures provenance–verification synergy. The model predicts a sharp phase transition between self-healing ($R_{\text{eff}} < 1$) and self-destabilizing ($R_{\text{eff}} > 1$) regimes. We introduce a three-class provenance taxonomy: cryptographic, institutional, and *context binding*, and derive four falsifiable hypotheses with a proposed 12-week stepped-wedge cluster-randomized trial in government grant review panels. The framework bridges AI alignment theory and institutional design.

Keywords: AGI governance · coordination theory · institutional design · decision provenance · AI safety · multi-agent systems

1 Introduction

Previous waves of information technology accelerated transmission: the telegraph collapsed geographic latency; the internet democratized access. Artificial General Intelligence accelerates something qualitatively different: *synthesis and decision*. The specific risk is not speed alone but the structural consequence of decision velocity outpacing verification velocity. Without engineered countermeasures, the cost of verifying reality exceeds the cost of generating fiction, and institutions enter a regime where rational inaction becomes the dominant strategy.

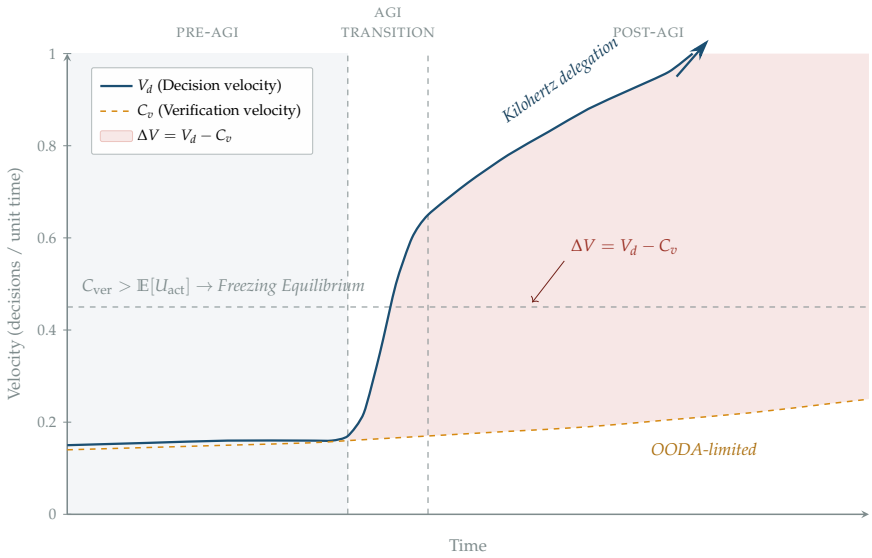

Fig. 1. The Decision–Verification Gap. Decision velocity V_d diverges from verification velocity C_v as AGI accelerates delegation beyond human verification capacity. The shaded region represents unverified decisions accumulating faster than they can be processed. When $C_{\text{ver}} > \mathbb{E}[U_{\text{act}}]$, the system reaches the Freezing Equilibrium.

The decision–verification gap. Let V_d denote the rate at which AI systems generate decisions and C_v the rate at which human (or human-supervised) processes can verify them. Historically, both were bounded by cognitive throughput, and $V_d \approx C_v$. AGI decouples these rates: synthetic principals execute directives at kilohertz frequencies while human verification remains tethered to the biological orientation phase of the OODA loop [5], requiring 0.2–2.0 s per assessment [7]. As the velocity gradient $\Delta V = V_d - C_v$ grows, unverified claims accumulate faster than they can be processed (Fig. 1).

The Freezing Equilibrium. When the expected cost of verification exceeds the expected utility of action,

$$C_{\text{ver}} > \mathbb{E}[U_{\text{act}}] \implies \text{Stasis}, \tag{1}$$

rational agents wait. When every agent waits because first-movers are exploited by unverified claims, the system reaches a Nash equilibrium that is individually rational but collectively catastrophic. This is not gridlock from incompetence; it is a structural property of the decision network under high ΔV.

Worked example. Consider an environmental regulatory agency that receives permit applications with AI-generated environmental impact assessments. Before AGI, the agency processed 200 assessments per year, each requiring ap-

proximately 40 analyst-hours to verify. An AGI-augmented consulting industry now generates 2,000 assessments annually of comparable apparent quality. The agency's verification capacity has not scaled proportionally. Analysts, unable to distinguish rigorous from confabulated assessments without full review, face inequality (1): the expected cost of verifying any single assessment exceeds the expected utility of approving it (since approval of an unverified fraudulent assessment carries career and legal risk). Rational analysts defer decisions. Development stalls, not from regulatory opposition, but from verification paralysis. The Freezing Equilibrium is not a failure of will; it is the predictable outcome of a verification bottleneck under conditions of high ΔV.

The metamaterial analogy as theory. Engineered metamaterials, periodic microstructures that produce emergent macro-properties such as photonic bandgaps, mechanical anisotropy, and phase transitions, offer more than a metaphor. The analogy generates specific, falsifiable predictions that a generic "governance framework" would not: that certain failure modes can be *forbidden* by microstructure design (bandgaps), that coordination improvements are directional (anisotropy), that interventions are superadditive (synergy), and that withdrawal is asymmetrically costly (hysteresis). If these predictions fail empirically, the metamaterial framing is wrong and should be discarded. If they hold, the framing becomes a design language for institutional engineering.

The analogy serves two functions: *heuristic* (organizing disparate phenomena under a single design language) and *generative* (importing structural predictions that a plain branching-process model does not produce). We distinguish these throughout.

Contribution. The treatment of governance rules as designable microstructure extends Ostrom's institutional design tradition [24] by providing a formal threshold criterion ($R_{\text{eff}} < 1$) for when the institutional design is sufficient.

This paper makes four contributions: (1) a phenomenological constitutive law for institutional coordination, parameterized by designable features, with a sharp phase transition (Section 2); (2) a three-class provenance taxonomy identifying *context binding* as the missing third class (Section 3); (3) treatment of AI agents as "synthetic principals" requiring distinct governance primitives (Section 4); (4) four falsifiable hypotheses with a concrete experimental design (Section 6).

2 The Constitutive Law and Phase Transition

The model belongs to the family of stochastic branching processes [2,15] and cascading failure analysis [28,6,11]. We contribute an *institutional parameterization*: (i) β is a *design variable* under institutional control, not an exogenous rate; (ii) the effective reproduction rate decomposes into actionable targets (ρ, τ) with a synergy term (γ); and (iii) the provenance taxonomy (Section 3) makes these parameters measurable.

2.1 Failure Propagation as a Branching Process

Institutional stability is a structural property dictated by the rate at which errors propagate across decision nodes. Each node in a decision network acts as either a signal attenuator or a failure amplifier. We model this as a stochastic branching process parameterized by four quantities:

- **Branching factor** β: the average number of downstream nodes a single decision impacts. This is *endogenous*: rate limits, delegation boundaries, and dual-control requirements reduce β directly.
- **Provenance fidelity** $\rho \in [0, 1]$: the probability that the source and transformation history of information is cryptographically bound to the decision unit.
- **Verification rate** $\tau \in [0, 1]$: the probability that a node detects and halts an erroneous claim.
- **Synergy coefficient** $\gamma \geq 0$: an interaction term capturing that high-ρ systems make verification cheaper (clear lineage speeds audits) and high-τ systems incentivize provenance (actors provide documentation when they know it will be checked).

We propose the following phenomenological constitutive law for the effective failure propagation rate:

$$R_{\text{eff}} = \beta \cdot (1 - \rho) \cdot (1 - \tau) \cdot (1 + \gamma\rho\tau). \tag{2}$$

R_{eff} is a *phenomenological ansatz*: its justification is empirical, like Hooke's law or Ohm's law. The multiplicative structure $(1-\rho)(1-\tau)$ reflects sequential filtering: at each node, an error must survive both a provenance check and a verification check. The independence assumption is a simplification that the synergy term γ partially corrects.

The synergy term $(1 + \gamma\rho\tau)$ captures mutual reinforcement between provenance and verification. The bilinear form is the simplest interaction that vanishes when either ρ or τ is zero. The qualitative superadditivity prediction (H3) is robust to the functional form; quantitative thresholds depend on the specific interaction and must be empirically calibrated (Section 7.3).

2.2 The Phase Transition

The system exhibits a sharp phase transition at $R_{\text{eff}} = 1$. This threshold property is inherited from the branching-process model class [2,15] and is not, in itself, a novel prediction:

- **Damped regime** ($R_{\text{eff}} < 1$): Errors decay exponentially with network depth. The system is self-healing; tail risk is bounded.
- **Turbulent regime** ($R_{\text{eff}} > 1$): Errors amplify exponentially. The system is self-destabilizing; cascade depth follows a power-law distribution with fat tails.

What the institutional parameterization adds is that the sub-critical condition $R_{\text{eff}} < 1$ can be *engineered* rather than merely observed: β can be constrained by delegation policy, ρ can be increased by provenance infrastructure, and τ can be raised by verification protocols. The constitutive law makes the relationship between these design choices and system-level stability quantitative.

The critical verification threshold for a given branching factor is $\tau^* = 1 - 1/\beta$ (in the simplified case where $\rho = 0$ and $\gamma = 0$). For a standard hierarchical panel with $\beta = 10$, stability requires $\tau > 0.90$. If AGI-accelerated delegation pushes the effective branching factor to $\beta = 50$, the required verification fidelity reaches 0.98, a threshold that legacy institutions cannot sustain without automated scaffolding.

The synergy term $(1 + \gamma\rho\tau)$ is crucial: because it multiplies rather than adds, combined provenance and verification interventions are *superadditive*. A small coordinated improvement in both ρ and τ can flip the system from turbulent to damped (Fig. 2), whereas equivalent improvements in either alone may not.

2.3 The Branching Factor as a Design Variable

Unlike natural epidemiological branching, β in institutional networks is a *design parameter*. Rate limits bound the maximum β. Delegation boundaries partition the network into sub-graphs. Dual-control requirements reduce the effective β for high-impact decisions. This means that scaffolding can attack all three terms in (2) simultaneously: reducing β, increasing ρ, and increasing τ.

Analogously to metamaterial design, where bandgap width depends on lattice periodicity and contrast ratio, the coordination "bandgap", the range of failure modes that cannot propagate, depends on the combination of delegation constraints, provenance requirements, and verification protocols. The constitutive law makes this dependence quantitative and testable.

3 A Three-Class Provenance Taxonomy

Current scaffolding initiatives focus on two provenance mechanisms: content provenance (C2PA [10]) and identity verification (Proof of Personhood). Neither addresses the full attack surface. We propose three complementary classes.

3.1 Class A: Cryptographic Provenance

Cryptographic provenance establishes the chain of custody from origin to current state via signatures that are computationally infeasible to forge. The C2PA Technical Specification 2.0 embeds tamper-evident manifests directly within asset bitstreams using JUMBF structures, including hash assertions for content binding and action assertions for transformation history [10].

Failure mode: Key compromise, implementation bugs, or capture of the signing infrastructure. *Mitigation*: Hardware security modules, key rotation, distributed signing, and audit of signing infrastructure.

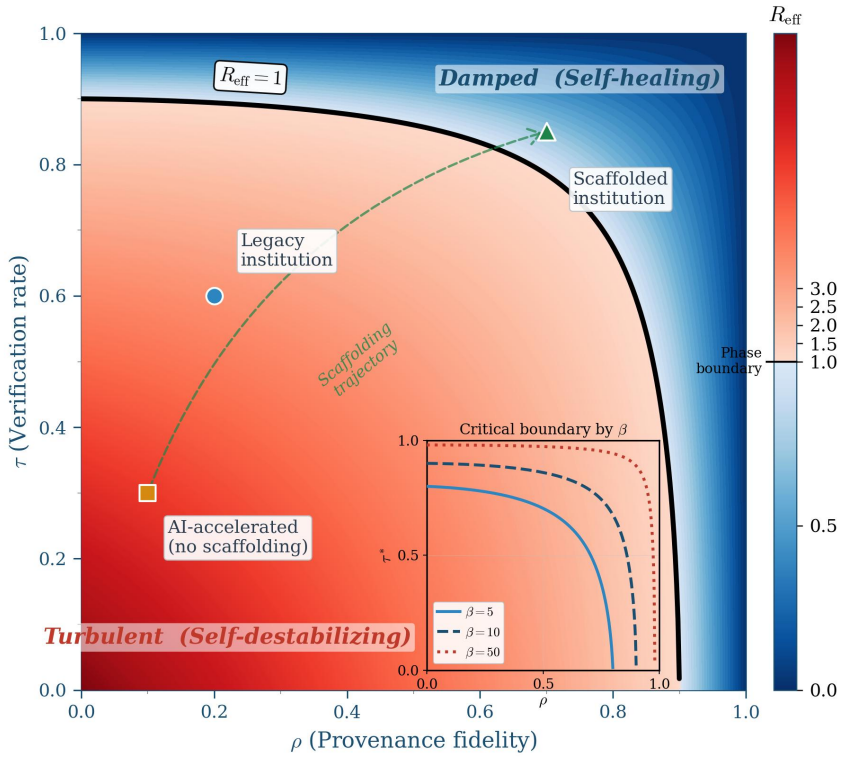

Fig. 2. Phase transition diagram for R_{eff} in the (ρ, τ) parameter space ($\beta = 10$, $\gamma = 1$). The bold contour marks the critical boundary $R_{\text{eff}} = 1$; blue region is damped (self-healing), red is turbulent (self-destabilizing). Three archetypal institutional positions are shown. Inset: boundary shift under varying branching factor β.

3.2 Class B: Institutional Provenance

A document signed by a ministry has institutional provenance: the reputation of the signing entity provides assurance beyond the cryptographic fact. The IETF SCITT standard [3] operationalizes this by submitting signed statements to a Transparency Service that issues receipts anchored in an append-only Merkle tree, preventing equivocation: an institution cannot repudiate a decision without invalidating the cryptographic proof held by downstream parties.

Failure mode: Institutional corruption, capture, or incompetence. The signature is valid but the institution's judgment is compromised. *Mitigation*: Separation of powers, external audit, competitive verification markets, and reputation tracking.

A critical insight: high cryptographic provenance with low institutional provenance (a corrupt institution with perfect signatures) is more dangerous than the reverse, because the cryptographic validity provides false assurance.

3.3 Class C: Context Binding (Novel Contribution)

Classes A and B secure integrity and attribution but fail against *"Valid Credential, Invalid Context"* attacks, where adversaries replay authorized outputs outside their intended temporal window, jurisdiction, or decision scope. A signed resource authorization from Q1 is re-injected in Q3; a regulatory approval for jurisdiction A is cited in jurisdiction B. Current standards do not prevent this.

Context binding, governed by a strictness parameter σ_c, tethers decisions to specific operational boundaries. Implementation utilizes Structured Rationale Capture (SRC): synthetic principals commit to a reasoning path *before* outcome realization, creating a "Decision Anchor" that makes post-hoc rationalization computationally infeasible. The protocol nests SRC data within C2PA `c2pa.actions.v2` assertions using the `parameters-map-v2` extension [10].

Worked example. A government procurement panel receives an AI-generated application citing a research partnership with University X, signed with a valid C2PA manifest (Class A) by a recognized research council (Class B). The partnership was genuine, but it expired six months ago. Without context binding, the claim propagates through review because its cryptographic and institutional provenance are intact. With context binding, the temporal scope tag in the SRC metadata triggers a mismatch against the current decision window: the verification cost becomes $O(1)$ rather than requiring a manual check of partnership status, and the claim is flagged before it influences scoring.

This mechanism directly addresses the Freezing Equilibrium. By enabling constant-time context verification, SRC reduces C_{ver} to a level where inequality (1) reverses, unfreezing the decision pipeline.

3.4 Two Types of Legibility

Following Scott's critique of state-imposed legibility [26,25], the framework distinguishes:

- **Rule-following legibility**: Did this process follow the specified protocol? This is amenable to zero-knowledge proofs: an entity can prove it checked a watchlist without revealing who was checked.
- **Judgment legibility**: Was this decision wise? Were relevant considerations weighed? This *cannot* be reduced to protocol compliance. It requires structured rationale capture, adversarial review, and outcome-linked accountability.

Design principle: Audit the process via ZKP. Audit the judgment via structured rationale and outcome tracking. Never require content disclosure when process compliance is sufficient. This resolves the legibility–privacy tension that Scott identified: we can make institutions auditable without making them panoptic.

3.5 Relation to Existing Governance Standards

The three-class taxonomy maps onto and extends existing governance frameworks. The NIST AI Risk Management Framework [22,23] addresses provenance through its "Map" and "Measure" functions, requiring organizations to document AI system inputs and outputs; ISO/IEC 42001 [17] mandates management-system controls for AI lifecycle documentation; and NIST SP 800-53 [21] specifies audit and accountability controls including AU-10 (non-repudiation). These frameworks provide robust coverage of Classes A and B. However, none currently addresses Class C, context binding, which requires tethering a credential or decision to a specific temporal, jurisdictional, and scope boundary. The taxonomy's contribution is to identify this gap and propose SRC as a mechanism to fill it, complementing rather than replacing the existing control architectures.

4 Synthetic Principals: AI Agents as Governance Nodes

The framework treats AI agents not merely as tools but as *nodes in the decision network* with distinct governance requirements. As AI agents increasingly generate, delegate, and execute decisions autonomously, they become "synthetic principals" requiring identity, provenance, and accountability primitives that differ from both human actors and passive software.

Identity and capability attestation. Each AI agent instance requires a non-repudiable cryptographic identity bound to, but distinct from, its operator's identity. This identity should include attested capabilities and permissions [20]. For agents that spawn sub-agents, the delegation chain must be preserved, maintaining an auditable lineage from the authorizing human principal through every layer of delegation.

Provenance requirements for AI outputs. AI-generated decisions require three provenance layers: input provenance (what data was consumed, from what sources, with what transformations), structured reasoning metadata (auditable traces, not full chain-of-thought which may be confabulated), and explicit confidence bounds that propagate through downstream decisions.

Verification challenges specific to AI. Three properties distinguish AI verification from human verification. First, *reasoning opacity*: human decision-makers can be deposed; AI explanations may be post-hoc rationalizations. Verification must focus on inputs, outputs, and consistency rather than stated reasoning [4,18]. Second, *speed asymmetry*: AI agents operate faster than human verification cycles, requiring rate limits that bound the verification backlog, a direct application of constraining β in the constitutive law. Third, *inter-agent coordination*: AI agents may coordinate in ways invisible to human oversight; monitoring must cover agent-to-agent communication, not just agent-to-human interfaces [19].

Relation to multi-agent governance and AI alignment. The synthetic principals framework connects to a growing literature on AI agent oversight. Chan et al. [8] argue that visibility into AI agent behavior requires monitoring not just individual outputs but the emergent dynamics of agent collectives, a concern directly addressed by our constitutive law, which models error propagation across networks of agents rather than individual agent reliability. Shavit et al. [27] propose structured delegation practices for AI systems, including explicit capability attestation and human-in-the-loop checkpoints at delegation boundaries; these practices map directly onto our ρ (provenance at delegation handoffs) and τ (verification at checkpoints) parameters. More broadly, scalable oversight proposals [9,4] address a related problem: how to supervise AI systems that exceed human capability, but typically assume a single principal–agent relationship. The metamaterial framework extends this to networks of synthetic principals with recursive delegation, providing a quantitative criterion ($R_{\text{eff}} < 1$) for when the oversight architecture is sufficient across the full delegation network, not just at individual handoff points, and when it has failed ($R_{\text{eff}} > 1$).

5 Trust Anchors: Where the Recursion Bottoms Out

If provenance and verification are enforced by infrastructure, who audits the infrastructure? The recursion must terminate in trust anchors: mechanisms whose integrity is maintained by design constraints rather than further verification layers.

We identify four candidate classes, each with distinct failure modes: *constitutional commitments* (entrenched rules requiring supermajorities to alter; failure mode: constitutional capture), *distributed consensus* (Byzantine-fault-tolerant verification where no single party controls the process; failure mode: collusion or 51% attacks), *international treaty bodies* (multi-jurisdictional oversight; failure mode: great-power defection), and *competitive verification markets* (independent auditors with reputational stakes; failure mode: cartel formation).

The design principle follows the metamaterial analogy: no single anchor is sufficient, just as no single layer of a metamaterial produces the desired macroproperty. Robust scaffolding combines anchors with orthogonal failure modes, so that compromise of one layer does not cascade through the system. The constitutive law provides a quantitative test: the combined trust anchor configuration must keep $R_{\text{eff}} < 1$ under realistic adversarial assumptions about individual anchor compromise.

6 Falsifiable Hypotheses and Empirical Design

6.1 Four Signature Predictions

The metamaterial framing generates predictions that a generic governance framework would not. If these fail empirically, the framing should be discarded. We are explicit about which predictions are inherited from the branching-process model class and which depend on the metamaterial analogy.

H1: The Bandgap Effect. By enforcing mandatory provenance ($\rho >$ threshold) and dual-control verification ($\tau >$ threshold), specific failure classes, hallucinated precedents, unverified transactions, replayed authorizations, become forbidden states that cannot propagate deeper than N nodes. *Prediction*: Baseline pipelines show power-law distribution of error propagation depth (fat tails). Scaffolded pipelines show exponential decay (sharp cutoff at the bandgap).

The bandgap mechanism. In metamaterial physics, bandgaps arise from destructive interference when waves encounter *periodic* contrasts in material properties. The institutional analog is *mandatory dual-control checkpoints applied uniformly at every delegation boundary.* When the "lattice" of institutional controls is periodic, specific failure modes become structurally impossible: a replayed authorization (Class C attack) cannot propagate past even a single node with context binding active. The failure mode is *forbidden by structure*, not merely attenuated by probability. In a heterogeneous branching process, a claim can propagate deep via a sequence of weak nodes; in a periodic verification structure, there is no such "lucky path." Distinguishing a true structural bandgap from deep sub-critical behavior would require larger samples than the proposed pilot; we flag this as a priority for follow-up analysis.

H2: Coordination Anisotropy. High-velocity AI agents optimized for local throughput will destroy cross-boundary coordination unless interface protocols are specifically engineered. *Prediction*: Without interface scaffolding, increased within-unit productivity (AI drafting) *increases* cross-unit latency (review bottlenecks, integration failures). With interface scaffolding, cross-boundary metrics improve sharply with a threshold effect.

Formalization. The directional decomposition applies the constitutive law separately to within-unit and cross-boundary propagation paths, recognizing that these paths have different parameter values:

$$R_{\text{eff}}^{\text{intra}} = \beta_{\text{intra}} \cdot (1 - \rho_{\text{intra}}) \cdot (1 - \tau_{\text{intra}}) \cdot (1 + \gamma_{\text{intra}}\, \rho_{\text{intra}}\, \tau_{\text{intra}}), \tag{3}$$

$$R_{\text{eff}}^{\text{cross}} = \beta_{\text{cross}} \cdot (1 - \rho_{\text{cross}}) \cdot (1 - \tau_{\text{cross}}) \cdot (1 + \gamma_{\text{cross}}\, \rho_{\text{cross}}\, \tau_{\text{cross}}). \tag{4}$$

AI acceleration typically reduces β_{intra} while increasing β_{cross}; simultaneously, $\rho_{\text{intra}} > \rho_{\text{cross}}$ and $\tau_{\text{intra}} > \tau_{\text{cross}}$ because provenance and verification are easier within shared contexts. The anisotropy prediction follows: $R_{\text{eff}}^{\text{intra}} < 1$ and $R_{\text{eff}}^{\text{cross}} > 1$ can hold simultaneously, producing a system that appears locally healthy while failing at interfaces. A scalar R_{eff} does not capture this; the prediction is motivated by the metamaterial concept of anisotropic response. The pilot can test this by measuring cascade depth separately for within-panel and cross-panel propagation.

H3: Provenance–Verification Superadditivity. Combined provenance and verification interventions are superadditive ($\gamma > 0$ in the constitutive law). *Prediction*: In a factorial design testing all four combinations of (low/high ρ) $\times$ (low/high τ), the high–high condition outperforms the sum of the two single-intervention conditions. *Note*: This prediction follows from the constitutive law's synergy term

and does not require the metamaterial analogy; it is a property of the model's interaction structure.

H4: Structural Hysteresis. Withdrawal of scaffolding yields asymmetric performance loss. Three mechanisms drive this: trust asymmetry (trust is built linearly but destroyed as a step function), skill atrophy (operators lose checking habits when scaffolding automates verification), and expectation reset (stakeholders calibrate to scaffolded performance, creating legitimacy penalties when it degrades). *Prediction*: Recovery time after scaffolding withdrawal exceeds original adoption time by a factor > 3. *Note*: Hysteresis appears in many dynamical systems and is not unique to metamaterials. We invoke the metamaterial framing here as a heuristic: the analogy to structural memory in physical metamaterials motivates the specific asymmetry prediction, but we acknowledge this is the weakest of the four analogical imports. The prediction's value is primarily empirical: if withdrawal costs are symmetric with adoption costs, the hysteresis mechanism is not operating.

6.2 Proposed Pilot: Government Grant Review

We propose a 12-week stepped-wedge cluster-randomized trial across government R&D grant review panels, chosen because they provide controlled, comparable units with real stakes.

Design. Twenty comparable panels are randomized: 10 treatment (scaffolded), 10 control (baseline). The baseline condition uses standard PDF submissions, manual eligibility screening, unstructured peer review, and summary memo documentation. The scaffolded condition adds structured data intake with mandatory provenance fields, automated eligibility filtering with audit trails, dual-blind review with structured scoring rubrics, pre-commitment (reviewers record initial scores before deliberation), structured rationale capture linked to rubric criteria, and randomized sampling audits with escalation triggers. Panels are stratified by domain (STEM, humanities, social sciences), panel size, and historical funding rate to ensure comparability (Fig. 3).

Primary endpoint. P95 cascade depth of injected tracer errors, harmless but detectable false claims seeded into *synthetic calibration applications* (see Section 6.3 for safeguards). This directly operationalizes the bandgap hypothesis.

Secondary endpoints. Time-to-decision (days from submission to notification), protest rate (formal appeals as a proxy for perceived legitimacy), shadow work rate (percentage of review activity occurring outside the official system), reviewer workload (hours per application), and decision consistency (inter-rater reliability on calibration applications reviewed by multiple panels).

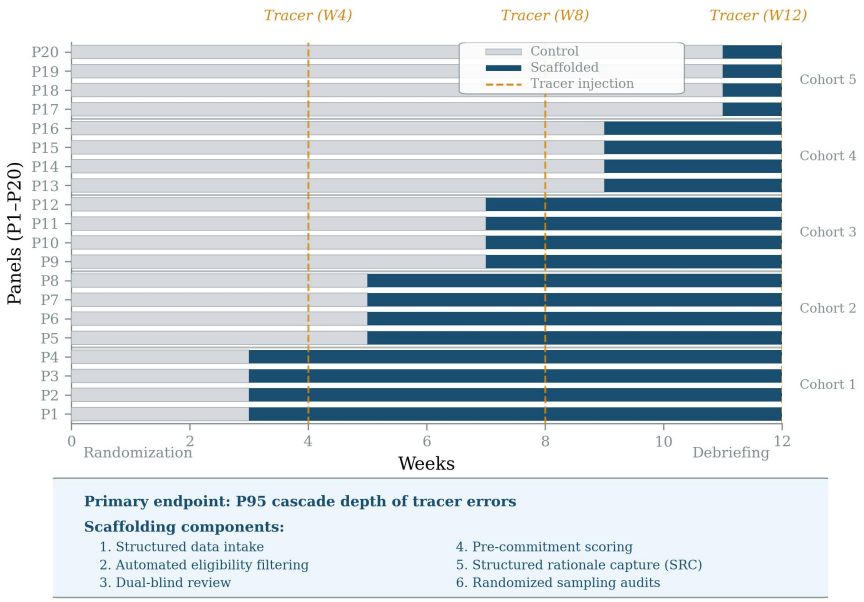

Fig. 3. Stepped-wedge cluster-randomized trial design. Twenty panels cross from control (gray) to scaffolded (blue) conditions at staggered intervals over 12 weeks. Tracer errors are injected at weeks 4, 8, and 12 to measure cascade propagation depth.

Table 1. Required number of panels (total, treatment + control) to achieve 80% power at $\alpha = 0.05$ for a 30% reduction in P95 cascade depth, as a function of intra-cluster correlation (ICC). Assumes 75 applications per panel and CV = 0.3.

ICC	0.01	0.05	0.10	0.15
Panels required	12	20	32	44

Power analysis. With 20 panels and 75 applications per panel, the design achieves 80% power to detect a 30% reduction in P95 cascade depth at $\alpha = 0.05$, under the following assumptions: intra-cluster correlation (ICC) of 0.05, consistent with published ICCs for stepped-wedge trials in administrative settings [13]; a coefficient of variation of 0.3 for baseline cascade depth; and a design effect of $1 + (m - 1) \times \text{ICC}$ where $m = 75$ applications per panel. Table 1 shows the sensitivity of required panel counts to the ICC assumption.

Hard guardrail. No intervention is successful if mean throughput improves while tail risk worsens. Success is defined as a vector improvement: cascade depth reduction *without* increased decision latency or shadow work rate.

6.3 Ethical Safeguards

The tracer-error methodology requires IRB approval and four safeguards. First, tracer errors are injected exclusively into fabricated *synthetic calibration applications*, not real submissions; these are flagged in the trial backend and cannot influence real funding decisions, following established "mystery shopper" methodology [29]. Second, no real applicant's submission is modified, delayed, or disadvantaged; the stepped-wedge design ensures all panels eventually receive scaffolding. Third, reviewers are informed of quality-assurance measures but the specific nature of tracers is disclosed only post-trial (standard minimal-deception protocol with mandatory debriefing). Fourth, all reviewer performance data is anonymized before analysis and used only in aggregate.

7 Discussion

7.1 Political Economy of Scaffolding

Scaffolding is not politically neutral. Opacity produces informational rents for incumbents: middle managers, gatekeepers, and intermediaries who derive power from information asymmetry [16]. Making a system legible redistributes power from information hoarders to auditors and high performers. Organizations with weak provenance and opaque delegation may be weak precisely because powerful actors benefit from that opacity.

This creates an adoption barrier that purely technical solutions cannot overcome. The framework addresses it through incentive compatibility: scaffolding must offer high-performers "fast lanes" (reduced audit friction via proven reputation, expanded delegation authority, portable reputation assets) while imposing friction on opaque actors. If legibility does not yield individual utility: speed, agency, credit; it will be circumvented via shadow systems, violating the stability constraint.

7.2 The Stability Constraint

Human cognitive capacity is a binding constraint. If scaffolding increases verification burden beyond a threshold, users bypass the secure system via shadow IT. Illustrative thresholds: roughly 5% overhead tolerance for high-stakes domains (nuclear, medical, financial), approximately 20% for lower-stakes domains. These are starting hypotheses, not precise targets. The design implication: verification must be *zero-attention* in the normal case, demanding cognition only for anomalies.

7.3 Sensitivity of the Constitutive Law to Functional Form

The constitutive law (Eq. 2) is a phenomenological ansatz. Its qualitative predictions, phase transition at $R_{\text{eff}} = 1$, superadditivity of combined interventions,

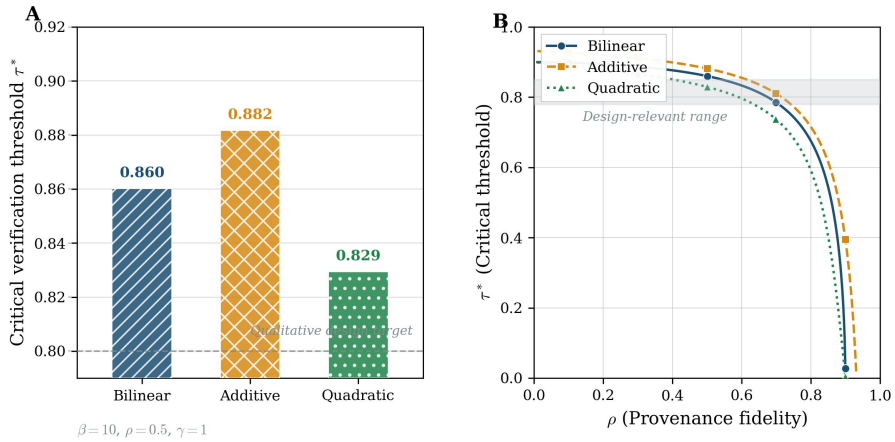

Fig. 4. Sensitivity of the critical verification threshold τ^* to synergy specification. (A) Bar chart at $\beta = 10$, $\rho = 0.5$, $\gamma = 1$: the three synergy forms yield $\tau^* \in [0.829, 0.882]$. (B) τ^* as a function of ρ for each form. The ± 3 pp spread indicates qualitative robustness of design guidance.

are robust to the choice of synergy term. However, quantitative predictions (critical thresholds, power calculations) depend on the specific interaction form. We briefly examine how the critical verification threshold τ^* (for a given β and ρ) shifts under three alternative synergy specifications:

1. **Bilinear (baseline):** $(1 + \gamma\rho\tau)$. The form used throughout this paper. τ^* decreases as ρ increases, with the rate of decrease modulated by γ.
2. **Additive:** $(1+\gamma(\rho+\tau)/2)$. Synergy accrues even with only one intervention active. This produces more conservative thresholds (higher τ^* for a given ρ) because the synergy term exceeds the bilinear baseline across most of the parameter space.
3. **Quadratic:** $(1+\gamma(\rho\tau)^2)$. Synergy is convex and weak at moderate parameter values, growing rapidly only when both ρ and τ are high. This produces more optimistic thresholds (lower τ^*) but stronger lock-in effects once high ρ is achieved.

Worked example. For $\beta = 10$, $\rho = 0.5$, and $\gamma = 1$, we compute the critical verification threshold τ^* (the minimum τ required for $R_{\text{eff}} = 1$) under each specification:

- **Bilinear:** Solving $10 \cdot 0.5 \cdot (1 - \tau) \cdot (1 + 0.5\tau) = 1$ yields $\tau^* \approx 0.860$.
- **Additive:** Solving $10 \cdot 0.5 \cdot (1 - \tau) \cdot (1.25 + 0.5\tau) = 1$ yields $\tau^* \approx 0.882$.
- **Quadratic:** Solving $10 \cdot 0.5 \cdot (1 - \tau) \cdot (1 + 0.25\tau^2) = 1$ yields $\tau^* \approx 0.829$.

The choice of synergy specification shifts τ^* by approximately ± 3 percentage points (Fig. 4). For policy purposes, this means that qualitative design guidance

(target $\tau > 0.85$ for $\beta = 10$ with moderate provenance) is robust to model uncertainty, but precise threshold calibration requires empirical data. The proposed pilot (Section 6) can resolve this: the factorial design testing (low/high ρ) $\times$ (low/high τ) provides the data needed to estimate the interaction structure and discriminate between functional forms.

7.4 Limitations and Open Questions

Several limitations bound the current framework. The constitutive law parameters require empirical estimation, and the model's utility depends on whether they can be measured reliably. The scale translation claim (organizational dynamics predicting civilizational phenomena) remains a hypothesis. Adversarial dynamics are treated as exogenous; a fuller treatment would model co-evolution of scaffolding and adversarial strategy. A first-principles derivation from micro-level institutional decision-making would place the ansatz on firmer footing. Finally, the metamaterial framing relies on qualitative analogical reasoning; importing formal homogenization theory and dispersion-relation analysis is a priority for future work.

7.5 Relation to AI Safety and Alignment

The framework complements technical AI alignment [1,12]: alignment ensures individual systems pursue intended goals; the metamaterial framework ensures institutional coordination survives when networks of agents operate beyond human oversight capacity [14]. R_{eff} operationalizes the scalable oversight agenda [4] at the institutional level. Ostrom's institutional design tradition [24] suggests a further direction: applying the constitutive law to multi-jurisdictional commons problems (climate governance, pandemic response, spectrum allocation).

8 Conclusion

AGI's dominant impact is the acceleration of decision velocity beyond institutional verification capacity. We have proposed governance engineering, the deliberate design of coordination microstructures, as the response. The constitutive law $R_{\text{eff}} = \beta \cdot (1 - \rho) \cdot (1 - \tau) \cdot (1 + \gamma\rho\tau)$ provides a quantitative stability criterion with a sharp phase transition. The provenance taxonomy identifies context binding as the critical gap; the synthetic principals framework connects to AI alignment; and the four falsifiable hypotheses ensure empirical accountability. The next step is the proposed procurement pilot. If the predictions fail, the framework should be discarded; if they hold, governance engineering becomes a discipline with quantitative foundations. Code and supplementary materials: https://github.com/davidorban/civilizationalmetamaterials.

Disclosure. AI writing assistants (Claude, Gemini, GPT) were used during drafting and revision. The author is solely responsible for all intellectual content and claims.

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in AI safety. arXiv preprint arXiv:1606.06565 (2016)
2. Anderson, R.M., May, R.M.: Infectious Diseases of Humans: Dynamics and Control. Oxford University Press (1991)
3. Birkholz, H., Delignat-Lavaud, A., Fournet, C., Deshpande, Y., Lasker, S.: An architecture for trustworthy and transparent digital supply chains. Internet-Draft draft-ietf-scitt-architecture-22, Internet Engineering Task Force (2025)
4. Bowman, S.R., et al.: Measuring progress on scalable oversight for large language models. arXiv preprint arXiv:2211.03540 (2022)
5. Boyd, J.R.: The essence of winning and losing (1996), unpublished briefing. Widely circulated; archived at https://www.dnipogo.org/boyd/essence_of_winning_losing.htm
6. Buldyrev, S.V., Parshani, R., Paul, G., Stanley, H.E., Havlin, S.: Catastrophic cascade of failures in interdependent networks. Nature **464**(7291), 1025–1028 (2010)
7. Card, S.K., Moran, T.P., Newell, A.: The Psychology of Human–Computer Interaction. Lawrence Erlbaum Associates (1983)
8. Chan, A., Ezell, C., Kaufmann, M., Wei, K., Tong, L., Korbak, T., Klenk, M., Skalse, J., Abate, A., Krueger, D.: Visibility into AI agents. The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT) (2024)
9. Christiano, P.F., Leike, J., Brown, T., Martic, M., Amodei, D., et al.: Deep reinforcement learning from human preferences. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
10. Coalition for Content Provenance and Authenticity: Content credentials: C2PA technical specification, version 2.0. Tech. rep. (2024), https://c2pa.org/specifications/specifications/2.0/specs/C2PA_Specification.html
11. Dobson, I.: Analyzing cascading failures and blackouts using utility outage data. In: Sun, K. (ed.) Cascading Failures in Power Grids. Springer Nature Switzerland AG (2024)
12. Gabriel, I.: Artificial intelligence, values, and alignment. Minds and Machines **30**(3), 411–437 (2020)
13. Hemming, K., Haines, T.P., Chilton, P.J., Girling, A.J., Lilford, R.J.: The stepped wedge cluster randomised trial: Rationale, design, analysis, and reporting. BMJ **350**, h391 (2015)
14. Hendrycks, D., Mazeika, M., Woodside, T.: An overview of catastrophic AI risks. arXiv preprint arXiv:2306.12001 (2023)
15. Hethcote, H.W.: The mathematics of infectious diseases. SIAM Review **42**(4), 599–653 (2000)
16. Holmström, B., Milgrom, P.: Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. Journal of Law, Economics, & Organization **7**, 24–52 (1991)
17. International Organization for Standardization: ISO/IEC 42001:2023 information technology – artificial intelligence – management system. Tech. rep. (2023)
18. Irving, G., Christiano, P., Amodei, D.: AI safety via debate. arXiv preprint arXiv:1805.00899 (2018)
19. Kittel, C., Siemens, C., et al.: AI agent orchestration patterns. Tech. rep., Microsoft Azure Architecture Center (2025), https://learn.microsoft.com/en-us/azure/architecture/ai-ml/guide/ai-agent-design-patterns

20. Narvaneni, Y.: The agent:// protocol – a URI-based framework for interoperable agents. Internet-Draft draft-narvaneni-agent-uri-02, IETF (2025)
21. National Institute of Standards and Technology: Security and privacy controls for information systems and organizations (NIST SP 800-53 rev. 5). Tech. rep., U.S. Department of Commerce (2020)
22. National Institute of Standards and Technology: AI risk management framework (AI RMF 1.0). Tech. rep., U.S. Department of Commerce (2023), https://www.nist.gov/itl/ai-risk-management-framework
23. National Institute of Standards and Technology: NIST-AI-600-1, artificial intelligence risk management framework: Generative artificial intelligence profile. Tech. rep., U.S. Department of Commerce (2024). https://doi.org/10.6028/NIST.AI.600-1, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
24. Ostrom, E.: Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press (1990)
25. Rao, V.: A big little idea called legibility. Ribbonfarm (2010), https://www.ribbonfarm.com/2010/07/26/a-big-little-idea-called-legibility/
26. Scott, J.C.: Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed. Yale University Press (1998)
27. Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Heidecke, A., Lama, K., Krueger, D.: Practices for governing agentic AI systems. OpenAI Research (2023), https://openai.com/research/practices-for-governing-agentic-ai-systems
28. Watts, D.J.: A simple model of global cascades on random networks. Proceedings of the National Academy of Sciences **99**(9), 5766–5771 (2002)
29. Wendler, D., Miller, F.G.: Deception in the pursuit of science. Archives of Internal Medicine **164**(6), 597–600 (2004)