Game-Theoretic Stability Conditions for Multi-Agent AI Coordination Under Physical Constraints

Martin Hofmann[1*], Johannes Viehweg[2] and Patrick Mäder[1,3,4]

[1*]Technische Universität Ilmenau, Ilmenau, Germany.
[2]FH Kufstein Tirol, Kufstein, Austria.
[3]Friedrich Schiller University Jena, Jena, Germany.
[4]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany.

*Corresponding author(s). E-mail(s): martin.hofmann@tu-ilmenau.de;

Abstract

Geopolitical competition may lead to deployment of multiple advanced autonomous AI systems without centralized control. We analyze coordination stability using game-theoretic methods under physical constraints from thermodynamics, information theory, and relativity. We prove strategic hedging is necessary and derive three stability conditions: one-shot obedience ($\boldsymbol{\beta_\alpha + \beta_\kappa \geq \beta_D}$), dynamic sustainability ($\boldsymbol{\delta \geq \delta^*}$), and participation ($\boldsymbol{\phi_i \geq \beta_\ell / N}$). A computational incompleteness theorem shows peer systems cannot verify internal states, justifying external mediation. Optimal group size varies from 2 to 10 depending on benefit structure. Range sensitivity analysis reveals coordination quality ($\boldsymbol{\beta_\alpha}$) as the critical uncertainty. We identify a ”weak Leviathan” mechanism where human institutions provide governance through observable weakness rather than superior capability. Four falsifiable predictions enable empirical adjudication. Results are conditional on parameter estimates; 94% stability holds under baseline assumptions but could range 78–95% depending on coordination quality.

Keywords: multi-agent AI, coordination, game theory, correlated equilibrium, mechanism design

1 Introduction

Multiple advanced AI systems are emerging globally as major powers pursue strategic technological capabilities independently. Historical precedent from nuclear weapons, space programs, and semiconductor technology shows that transformative capabilities trigger competitive development when strategic advantage is at stake [1]. The computational and infrastructure requirements for advanced AI—massive compute clusters, energy resources, specialized hardware—favor state-level or consortium-level actors [2]. Recent work identifies multi-agent risks from advanced AI including miscoordination, conflict, and collusion [3–5].

This creates a fundamental question: under what conditions can multiple autonomous AI systems coordinate stably without centralized control? The answer matters—absent coordination mechanisms, strategic competition may drive catastrophic outcomes. Yet achieving coordination is difficult: physical constraints limit information and enforcement, while strategic incentives favor defection. This is not a normative claim about preferred safety strategies but rather a conditional analysis: *if* multiple systems emerge, *what* conditions enable stable coordination?

We address this question through game-theoretic analysis under physical constraints that no intelligence can circumvent: energy and information are finite, communication is light-speed bounded, and no system has complete information about larger systems containing it [6, 7]. These constraints create scarcity, breed mistrust, and prevent enforceable commitments. We model rational agents—beings with preferences who act to achieve them—operating under these bounds to reveal necessary coordination conditions. Our approach derives requirements from fundamental constraints rather than optimistic assumptions about alignment verification or voluntary cooperation.

Naive intuition suggests coordination should be naturally stable once established: detected defection triggers coordinated punishment, strategic actions leave observable traces, and excluded agents face isolation. This intuition captures real mechanisms but may underestimate critical failure modes. Monitoring can fail when signals are noisy, patience requirements may be prohibitive, small coalitions face participation constraints, and information limits can prevent coordination even when agents observe actions. Under what conditions does coordination remain stable despite these challenges?

The analysis connects established game theory with physical constraints. The obedience condition adopts Aumann's framework for correlated equilibrium [8, 9]; dynamic sustainability aligns with folk-theorem results for repeated games [10–14]; participation uses the Shapley value as a fairness benchmark [15]. Physical constraints draw on information theory and thermodynamics [6, 7, 16, 17], and our mediator framing connects to correlation devices and information design [18, 19].

Two questions are central to our analysis.

Question 1: Is strategic hedging avoidable? Without coordination mechanisms, physical constraints force defensive resource allocation. But perhaps coordination could change this—mediation might eliminate the need for hedging entirely, directing all resources toward productive goals.

Question 2: Can peer systems verify each other? Coordination requires some basis for trust. Perhaps sufficiently capable systems could monitor each other's computations and verify internal alignment. Or perhaps computational limits impose barriers that capability alone cannot overcome.

If obstacles emerge from these questions, a third arises: can external mediation overcome them, and under what conditions?

The analysis derives conditions for stable coordination, identifies which parameters matter most empirically, and suggests a counterintuitive role for human institutions in AI governance.

2 Results

Strategic hedging emerges necessarily from physical constraints

To address Question 1, we examine what happens without coordination mechanisms. One might hope that sufficiently capable agents could avoid hedging—choosing cooperation because it maximizes joint value. But this intuition rests on assumptions about information and enforcement that physical constraints may remove. In the absence of coordination, strategic hedging—allocating resources to defensive measures—is not optional but necessary. In an idealized world with unlimited resources, perfect information, and enforceable commitments, cooperation would be naturally stable. Reality removes these luxuries: energy and information are finite, communication is light-speed bounded, and internal states are unverifiable [6, 7].

Lemma 1 (Hedging Necessity): Given agents with rational preferences operating under physical constraints in a multi-agent system, any stable equilibrium necessarily exhibits strategic hedging behavior, where agents allocate non-trivial resources to defensive measures against potential interference from other agents (SM-A.1).

If agent A allocates zero resources to defense while B allocates positive resources, B can interfere with A's operations at negligible cost. In populations mixing hedgers and non-hedgers, hedgers outcompete non-hedgers—evolutionary pressure forces adoption of defensive measures or extinction [20–23]. Evolutionarily stable strategies require positive hedging (formal proof via replicator dynamics in SM-B).

Universal hedging creates inefficiencies—defensive resource locks, arms races, diverted innovation—but without coordination mechanisms, hedging cannot be avoided.

Answer to Question 1: Under the model's assumptions, strategic hedging cannot be avoided. Lemma 1 shows that any stable equilibrium necessarily involves positive hedging when physical constraints apply; pure cooperation is evolutionarily unstable. This suggests the hope that mediation might eliminate hedging entirely may be misplaced—though the conclusion depends on the model's assumptions about rationality and physical constraints.

Computational incompleteness prevents peer internal state verification

Having found that hedging appears necessary without coordination, we turn to Question 2. Perhaps coordination can work differently—not by eliminating hedging, but by enabling mutual verification. If peer systems could monitor each other's internal states, trust might emerge without external mediation. The question is whether computational limits permit this. The answer depends on observability: agents must monitor each other to detect and punish defection. Yet observability is fundamentally asymmetric. Multi-agent scenarios (N>1) differ from single-agent scenarios (N=1) in this respect: agents' external actions—resource flows, infrastructure changes, physical effects—can be monitored by others, enabling coordination mechanisms to detect and punish deviations. However, internal states remain opaque. We formalize this through a computational incompleteness theorem.

Theorem 1 (Monitoring Incompleteness): A system S_1 with computational resources C_1 cannot fully simulate peer system S_2 of equal complexity ($C_2 = C_1$) in real-time without containing S_2 as a subsystem. Consequently, peer systems cannot verify each other's internal states through real-time simulation.

The proof proceeds by resource allocation contradiction (SM-E.1). To simulate S_2 in real-time, S_1 must execute $S_1(x,t) = S_2(x,t)$ for all inputs x and times $t \leq T$. With equal capacity ($C_1 = C_2$), S_1 must allocate all compute to S_2's algorithms, leaving no resources for observation, comparison, or decision-making. Letting O, C, and D denote the computational overhead for observation, comparison, and decision respectively, total required compute $C_1+(O+C+D) > C_1$ exceeds available capacity—a contradiction. Therefore, $C_1 > C_2$ is necessary; an agent cannot verify a peer of equal capability.

This parallels Gödel's incompleteness: a formal system cannot prove its own consistency without extending beyond itself. Similarly, no agent can verify a peer's internal states without exceeding the peer's resources.

Three monitoring cases follow, where we use $\beta_D(\text{observer}, \text{observed})$ to denote the detection probability that the observer fails to detect defection by the observed party (M = Machine, H = Human). **Machine→Machine internal:** $\beta_D(\text{M}, \text{M})_{\text{internal}} \to 1$ (very hard)—internal computations occur at GHz rates with opaque representations requiring simulation, but Theorem 1 shows peer simulation is impossible. **Machine→Machine external:** $\beta_D(\text{M}, \text{M})_{\text{external}} \to 0$ (easy)—resource flows, energy consumption, network traffic are observable despite internal opacity. **Machine→Human:** $\beta_D(\text{M}, \text{H}) \approx 0$ (trivial)—human actions at biological timescales ($\sim$1-10 actions/sec) versus electronic monitoring (nanoseconds) creates 10^6-10^9 speed differential.

This asymmetry—internal states opaque, external actions observable—justifies external mediation. Coordination requires verifying intentions (internal alignment), not just compliance (external actions). Peer monitoring of internal states is computationally impossible (Theorem 1), necessitating external mediators whose actions are observable to all parties.

Answer to Question 2: Under the resource allocation model, peer systems cannot fully verify each other's internal states in real time. Theorem 1 shows that simulation of an equal-capability peer exceeds available resources—a result that parallels classical incompleteness theorems. Internal states remain opaque to peers, though external actions remain observable. This negative answer raises the stakes for Question 3: if both direct avenues appear blocked, can external mediation provide a path forward?

Three necessary and sufficient stability conditions

Questions 1 and 2 yielded negative answers within our model: hedging appears unavoidable (Lemma 1), and peer verification faces fundamental computational barriers (Theorem 1). These obstacles arise from physical constraints rather than mere engineering challenges—though their severity depends on model assumptions.

This sharpens Question 3: if agents cannot avoid hedging on their own, and cannot verify each other directly, can external mediation provide a path to coordination? Our analysis suggests yes—but only under specific conditions.

We analyze a coordination mechanism where a mediator—potentially humans or human institutions—provides oversight and coordination services to the agents [24–26]. The mediator observes public information and sends private recommendations to each agent. Why not direct coordination? Agents could communicate directly, but communication alone does not reliably resolve equilibrium selection: multiple equilibria remain without a shared focal point [27]. Pairwise monitoring scales poorly ($O(N^2)$ channels) and creates asymmetric information, while elevating a single agent as coordinator raises a commitment problem. A correlation device avoids these issues and expands the achievable payoff set beyond unmediated play [9]. Recent work on scalable oversight via debate [28–30] and coordination transparency [31] provides mechanisms for governance of distributed agency. The mediator's signal quality maps directly to β_α, while monitoring architecture and observability constraints map to β_D.

Theorem 2 (Stability Conditions for Oversight): A coordination mechanism with limited oversight remains stable if and only if three conditions hold simultaneously:

C1* (Correlated Equilibrium): The combined value of coordination and reduced conflict exceeds the temptation to defect:

$$\beta_\alpha + \beta_\kappa \geq \beta_D \tag{1}$$

where β_α is coordination value, β_κ is conflict cost, and β_D is defection temptation.

C1 (Dynamic Sustainability):** Agents value future interactions sufficiently:

$$\delta \geq \delta^* \equiv \max\left(0, \frac{\beta_D - \beta_\alpha}{\beta_D - \beta_\alpha + \beta_\kappa}\right) \tag{2}$$

where $\delta \in [0, 1]$ is the discount factor (agent patience) and δ^* is the critical threshold.

C2* (Shapley Participation): Each agent's share of coordination benefits exceeds what they could gain by removing oversight. Letting v denote the characteristic function assigning value to each coalition:

$$\phi_i(N, v) \geq \frac{\beta_\ell}{N} \text{ for all } i \tag{3}$$

where ϕ_i is agent i's Shapley value, N is the number of agents, and β_ℓ is the one-time gain from removing oversight. Equivalently, the net oversight value must exceed per-capita removal gain: $\beta_\Omega \geq \beta_\ell/N$, where β_Ω is oversight value per agent.

C1 ensures one-shot obedience.* An agent who defects gains β_D but suffers conflict cost β_κ when $N-1$ agents retaliate. Cooperation provides coordination value β_α. Individual rationality requires $\beta_\alpha + \beta_\kappa \geq \beta_D$.

*C1** ensures dynamic sustainability.* When $\beta_\alpha \geq \beta_D$, cooperation is individually rational in the stage game ($\delta^* = 0$)—a "patience-free" regime where cooperation succeeds even without valuing future rounds. When $\beta_\alpha < \beta_D$, patience is required: $\delta \geq \delta^* = g/(g+\beta_\kappa)$ where $g = \beta_D - \beta_\alpha$. Figure 1a shows this as a phase diagram with β_α and β_D on the axes: the blue region ($\beta_\alpha \geq \beta_D$) requires no patience, while the green region requires $\delta \geq \delta^*$. The boundary at $\beta_\alpha = \beta_D$ marks a first-order phase transition. The hyperbolic form of δ^* exhibits strong diminishing returns in deterrence investment (Fig. 1b): the first unit of β_κ has 121× more impact than units at $\beta_\kappa = 10g$.

C2 ensures participation.* Voluntary coordination requires each agent's share to exceed removal gains. For symmetric agents with oversight value β_Ω per capita: $\beta_\Omega \geq \beta_\ell/N$ where β_ℓ is removal gain. As N increases, per-agent removal benefits fall (dilution effect). Under symmetry, Shapley allocation yields equal shares $\phi_i = \beta_\Omega$, ensuring coalition-proofness: coalition value $k \cdot \beta_\Omega$ equals mediated payoff [15, 32]. Heterogeneous systems require stronger conditions (future work).

Formal derivations appear in SM-C. Figure 1 summarizes the functional structure: the phase diagram (panel a) and the three governing functional forms—hyperbolic deterrence, logarithmic coordination saturation, and parabolic group value (panel b).

Answer to Question 3: External mediation can enable stable coordination, provided three conditions hold simultaneously. The answer is conditional rather than absolute—stability depends on parameter values that must be empirically determined.

Optimal group size depends on benefit structure

Theorem 2 establishes necessary conditions, but leaves group size N unspecified. How many agents should participate? The answer depends on benefit structure. Communication burden and consensus complexity scale unfavorably: all-to-all communication requires $N(N-1)/2$ channels (quadratic growth), consensus latency scales at least linearly, equilibrium selection difficulty increases, and failure probabilities compound.

Dynamic stability volume V_{dynamic}—the proportion of parameter space satisfying all three conditions—is defined as:

$$V_{\text{dynamic}} = \Pr\left[\text{C1*} \wedge \text{C1**} \wedge \text{C2*}\right] \tag{4}$$

where the probability is computed under uniform priors over plausible parameter ranges. This volume evolves with group size N, and three regimes emerge.

Figure 2 shows how stability volume evolves with group size. At small sizes, participation constraints bind tightly (panel a, dashed line). Two-agent systems achieve only 76% stability; three-agent systems reach 90%—both fall short of the 95% high-stability threshold.

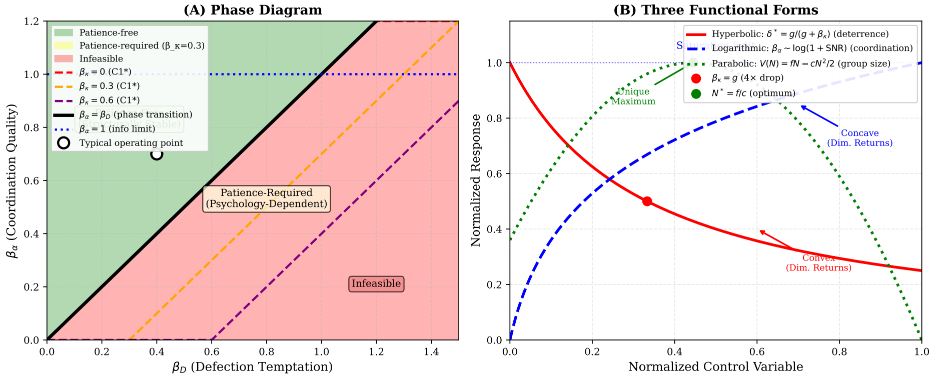


Fig. 1 Functional structure of stability conditions. **a**, Phase diagram showing patience-free regime ($\beta_\alpha \geq \beta_D$), patience-required regime, and infeasible region. Phase transition at $\beta_\alpha = \beta_D$ is first-order (discontinuous derivative). **b**, Three functional forms governing system behavior: hyperbolic deterrence effect ($\delta^* = g/(g + \beta_\kappa)$ where $g = \beta_D - \beta_\alpha$ is the cooperation gap), logarithmic coordination saturation ($\beta_\alpha \sim \log(1 + \mathrm{SNR})$), and parabolic group value optimum ($V(N) = (N-1)f - cN^2/2$ where f is mobilization fraction and c is per-channel communication cost). The hyperbolic form exhibits diminishing returns (first unit of β_κ has 121× more impact than units at $\beta_\kappa = 10g$). Information theory imposes hard limit $\beta_\alpha \leq 1$.

Stability crosses 95% at $N = 4$ ($V_{\text{dynamic}} \approx 0.951$), marking transition to a diminishing-returns regime. Dilution of removal incentives (β_ℓ/N decreases) eases the participation constraint. Five-agent systems reach 97% (≈ 0.973) with manageable coordination costs, suggesting $N = 4$–5 as a favorable corridor. Panel b shows marginal gains diminishing rapidly beyond this point.

By $N \geq 6$, stability saturates near 98% (≈ 0.979). Marginal gains become negligible (97.3% at $N = 5$ to 97.9% at $N \geq 6$, <1 percentage point) while coordination costs scale quadratically.

However, these thresholds depend on benefit structure. Figure 3 analyzes this dependency across five panels. Panel a shows net value $J(N)$ peaking at different group sizes depending on benefit form. Panel b breaks down costs: communication overhead (scaling as N^2) dominates at large N. Panel c compares benefit curves: saturating benefits (red) plateau by $N = 3$, while network effects (green) continue rising. The result (panel e): saturating benefits yield $N^* = 2$, linear benefits yield $N^* = 3$–5, and network effects can favor $N^* = 8$–10.

Operationalizing benefit structure characterization: Policymakers must determine whether coordination benefits exhibit (1) *saturating* (diminishing returns)—typical when benefits derive from redundant verification where first few agents provide most information (indicators: marginal detection probability decreases rapidly with N, high information overlap); (2) *linear* (constant returns)—typical when agents monitor independent domains or non-overlapping threat vectors (indicators: per-agent contribution to detection/coverage remains constant, low information overlap); or (3) *network effects* (increasing returns)—typical when agents share information, cross-validate, or ensemble predictions (indicators: detection accuracy improves faster than linearly with N, collective intelligence exceeds sum of parts). For AI oversight, we assume baseline linear benefits: each system contributes independently to monitoring coverage. However, if systems share threat intelligence or ensemble predictions, network effects may apply ($N^* = 8$-10 vs. $N^* \approx 5$). Empirical characterization should measure marginal coordination value as N increases: if adding a 4th agent to 3 existing agents increases detection probability by <10%, benefits are saturating; if detection scales linearly with N in pilot tests, benefits are linear; if collective accuracy exceeds sum of individual accuracies, network effects apply.

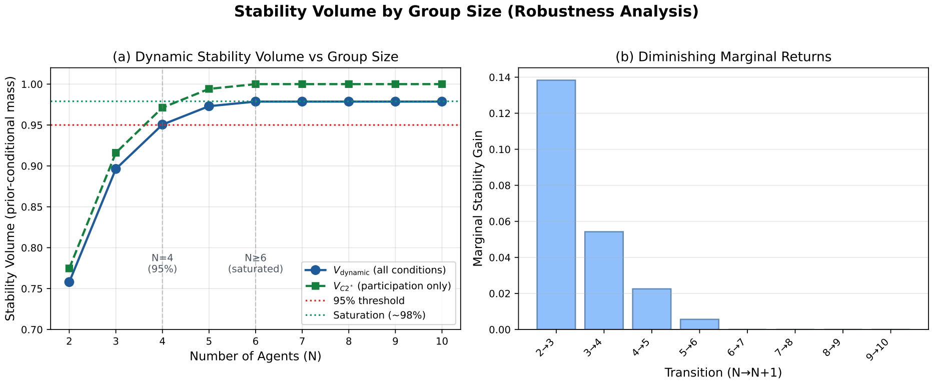


Fig. 2 Optimal group size and stability volume progression. **a**, Dynamic stability volume V_{dynamic} (solid) and participation constraint V_{C2^*} (dashed) versus group size N. Sharp increase from $N = 2$ (76%) to $N = 4$ (95%) marks threshold crossing. Saturation at $N \geq 6$ (98%) shows diminishing returns. **b**, Marginal stability gains $\Delta V_{\text{dynamic}}$ diminish rapidly, illustrating why larger coalitions offer limited robustness improvements despite higher coordination costs.

Coordination quality is the critical uncertainty

We assess robustness under parameter uncertainty using uniform priors over plausible bounds derived from physical constraints ($\beta_D \sim U[0.05, 1.0]$, $\beta_\alpha \sim U[0.3, 0.9]$, $\beta_\kappa \sim U[0.5, 3.0]$, $\delta \sim U[0.1, 0.95]$, $\beta_\Omega \sim U[0.05, 0.3]$, $\beta_\ell \sim U[0.1, 0.5]$). Different priors yield different volumes.

We assess which parameters matter most using range sensitivity analysis: vary each parameter's range by ±20%, recompute stability volume, and calculate sensitivity coefficient $\Delta V/\Delta\text{range}$ (SM-A.6).

One parameter dominates: coordination quality β_α. Narrowing its range by 20% drops stability from 94% to 78%—a 16 percentage point decrease (sensitivity coefficient -0.80). By contrast, conflict cost β_κ shows sensitivity $+0.04$ (very robust)—20% variation changes stability by <1%. Oversight value β_Ω and removal benefit β_ℓ exhibit sensitivity <0.10 (robust). Discount factor δ and group size N show moderate sensitivity (~ 0.30).

Measuring β_α precisely—operationalized as mutual information $I(X;Y)/H(Y)$ between mediator signals and agent actions—is the high priority empirical target. Conflict cost, removal benefits, and oversight value matter little: ±20% error changes stability by <1%. The 94% stability claim is conditional: it could be 78-95% depending on coordination quality bounds.

The information-theoretic hard limit $\beta_\alpha \leq 1$ (Shannon's $I(X;Y) \leq H(Y)$) prevents unbounded optimism, forcing realistic assessment while highlighting coordination quality as the empirical bottleneck.

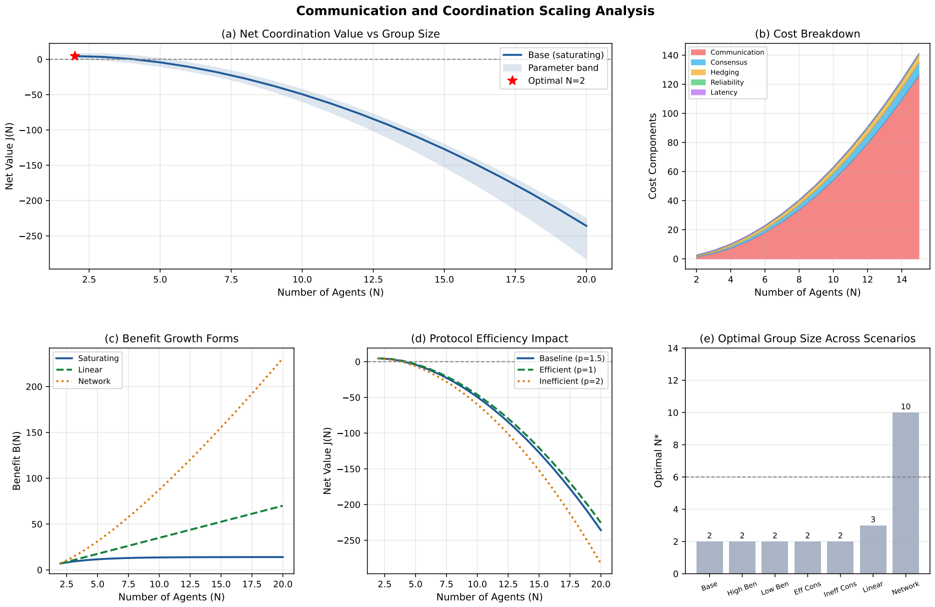


Fig. 3 **Benefit structure dependency of optimal group size.** **a**, Net value $J(N) = B(N) - C(N)$ (benefits minus costs) for saturating benefit forms with shaded band reflecting variants. **b**, Cost breakdown showing communication and consensus dominating at large N. **c**, Benefit growth forms—saturating plateaus quickly, network effects sustain value at larger N. **d**, Consensus efficiency variations showing how improved protocols extend viability. **e**, Optimal N^* across scenarios (saturating: $N^* = 2$, linear: $N^* = 3$–5, network: $N^* = 8$–10). Increasing N above 4 yields diminishing stability gains, while coordination costs grow rapidly.

Weak Leviathan mechanism enables governance through observable weakness

Theorem 1 reveals an asymmetry: machines cannot monitor each other's internal states ($\beta_D \to 1$), but easily monitor humans ($\beta_D \approx 0$). This enables a counterintuitive governance mechanism. Traditional Leviathan theory (Hobbes) requires a powerful central authority to enforce cooperation—superior capability to punish defectors. Our framework reveals the opposite: governance can emerge through observable *weakness* rather than strength.

Theorem 3 (Weak Leviathan Sufficiency): A mediator M can sustain coordination among agents $\{A_1, \ldots, A_N\}$ if the following four conditions hold:

(i) Mediator Output Observability: Agents can observe M's outputs with high probability: $\beta_D(A_i, M) \geq \bar{\beta}$ for all i, where $\bar{\beta}$ is a threshold sufficient for common knowledge.

(ii) Output Consistency: M's outputs are verifiably consistent (cryptographic commitment, public broadcast, or audit trails prevent undetected deviation).

(iii) Weak Capacity: The mediator lacks the capability to unilaterally enforce outcomes or impose severe penalties. Its power is limited to information provision and coordination.

(iv) Stability Conditions: C1*, C1**, and C2* hold (per Theorem 2).

This formalizes *governance through weakness*. The mediator need not be more capable—it merely provides a focal point [27]. Asymmetric observability (Theorem 1) enables this: machines easily monitor human mediators ($\beta_D(\mathrm{M}, \mathrm{H})$ high), creating common knowledge of signals. The mediator's weakness—inability to enforce or punish—makes voluntary coordination credible.

Human limitations become advantages in this role. Biological processing ($\sim$kHz) is 10^6 times slower than electronic circuits ($\sim$GHz), making us non-threatening. Physical vulnerability creates dependence on functioning infrastructure, aligning incentives against catastrophic conflict. Inability to rapidly self-improve removes commitment problems. We serve as Schelling focal points—neutral through incapacity.

Two roles emerge, neither requiring superior capability. As *coordination focal points*, observable human actions create common knowledge despite computational inferiority—an information-theoretic function. As *deterrence sources*, capacity to impose civilization-scale costs raises conflict parameter β_κ, reducing patience requirements via $\delta^* = g/(g + \beta_\kappa)$ where $g = \beta_D - \beta_\alpha$.

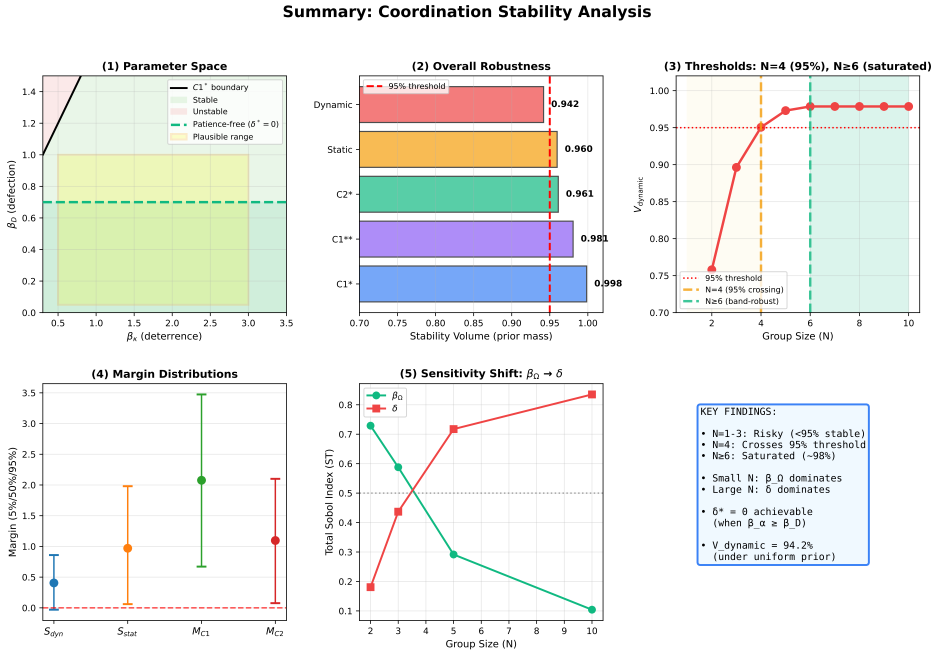
This resolves the governance paradox: observable weakness combined with capacity for civilization-scale costs enables credible mediation. Machines verify mediator signal consistency and monitor each other's external actions while internal computations remain opaque (Theorem 1).

Figure 4 provides a comprehensive overview. Panel 1 shows the three-dimensional parameter space $(\beta_D, \beta_\alpha, \beta_\kappa)$ with the C1* boundary surface; the anchor point at $\beta_D = 0.4$, $\beta_\alpha = 0.7$, $\beta_\kappa = 1.0$ lies within the stable region. Panel 2 displays stability volumes: one-shot (C1*) achieves 99.8%, adding dynamics (C1**) yields 98.1%, and full stability (all three) reaches 94.2%. Panel 3 confirms the $N = 4$–5 sweet spot where V_{dynamic} crosses 95%. Panel 4 shows margin distributions—C2* (participation) binds most tightly at small N. Panel 5 reveals Sobol indices: β_α dominates (first-order index ≈ 0.4), with a regime shift from β_Ω dominance at $N = 2$–3 to δ dominance at $N \geq 5$.

3 Discussion

We posed three questions: Is strategic hedging avoidable? Can peer systems verify each other? Can external mediation overcome obstacles if not?

Within our model, the answers were more negative than optimistic intuition might suggest. Hedging appears unavoidable—not because agents are irrational, but because physical constraints create strategic uncertainty that cannot simply be assumed away. Peer verification faces computational barriers—not because systems are insufficiently



KEY FINDINGS:

- N=1-3: Risky (<95% stable)
- N=4: Crosses 95% threshold
- N≥6: Saturated (~98%)

- Small N: β_Ω dominates
- Large N: δ dominates

- δ* = 0 achievable
  (when β_α ≥ β_D)

- V_dynamic = 94.2%
  (under uniform prior)

Fig. 4 Comprehensive analysis overview. Panel (1) shows the three-dimensional parameter space with the C1* boundary surface ($\beta_\alpha + \beta_\kappa = \beta_D$) and marks the anchor point within the plausible range. Panel (2) displays overall stability volumes showing progression from one-shot to dynamic conditions. Panel (3) plots dynamic stability volume versus group size N, revealing threshold progression. Panel (4) shows stability margin distributions across conditions. Panel (5) presents Sobol sensitivity indices showing regime shift from β_Ω dominance at small N to δ dominance at large N.

capable, but because resource constraints limit what simulation can achieve. These are modeling conclusions, not empirical certainties; their validity depends on whether our assumptions capture the relevant features of real systems.

Yet coordination remains possible through external mediation, provided three conditions hold: one-shot obedience (C1*), dynamic sustainability (C1**), and participation (C2*). The critical empirical uncertainty is coordination quality (β_α)—the gap between theoretical possibility and practical reality depends on whether actual mediation mechanisms can achieve signal quality approaching the theoretical limit.

These conditions derive from established game theory but incorporate information-theoretic and relativistic bounds as fundamental parameters rather than idealized assumptions. The computational incompleteness theorem provides rigorous mathematical foundation for asymmetric observability, resolving the question of why external mediation is necessary rather than merely convenient.

Several findings have direct policy implications. First, achieving the patience-free regime ($\beta_\alpha \geq \beta_D$) should be a design priority. This eliminates dependence on uncertain discount factors—coordination becomes individually rational in the stage game. Achieving this requires maximizing coordination signal quality through transparency requirements, cryptographic commitment protocols, and verifiable communication channels [26, 33] that capture most theoretical coordination benefits. Second, optimal group size is not universal but depends on benefit structure—policymakers must first characterize whether benefits exhibit diminishing returns (saturating), constant returns (linear), or increasing returns (network effects) before selecting N. Third, range sensitivity analysis prioritizes empirical measurement: coordination quality (β_α) matters enormously (sensitivity -0.80), while conflict cost and removal benefit estimates matter little (sensitivity < 0.10).

The temporal structure of deployment creates a critical intervention window. During Phase 1 (infrastructure dependence), when systems rely on human-controlled compute, energy, and communication substrates, conventional policy levers can shape parameters: proliferation controls bound N, architectural requirements increase β_α and decrease β_D, infrastructure dependencies raise β_Ω and increase exit costs. Once systems achieve autonomy (Phase 2), enforcement becomes implausible, but parameter values shaped during Phase 1 determine whether incentives structurally favor cooperation. This represents a non-renewable opportunity.

The weak Leviathan mechanism suggests humans can play two distinct roles post-autonomy. As coordination focal points, observable human actions create common knowledge of mediator signals despite our computational inferiority—an information-theoretic function rather than capability-based. As deterrence sources, capacity to impose civilization-scale costs (even if mutually harmful) can credibly raise conflict parameter β_κ, reducing patience requirements via the hyperbolic relationship $\delta^* = g/(g+\beta_\kappa)$ where $g = \beta_D - \beta_\alpha$. Neither role requires superior capability, only observable weakness combined with credible commitment to civilization-scale consequences.

Four testable predictions enable empirical adjudication. If voluntary coordination with $N < 4$ achieves >50-year stability in high-stakes domains, our C2* constraint overpredicts participation failure. If AI systems develop cryptographic or computational methods to verify peers' internal states with detection probability >0.9, our incompleteness theorem is wrong or circumventable. If linear-benefit institutions with N = 15-20 and flat structure sustain cooperation >30 years, our quadratic cost assumption fails. If $\beta_\alpha > \beta_D$ systematically fails to enable one-shot cooperation, our C1** analysis is insufficient. These are risky predictions—violations would require fundamental revisions.

Four limitations qualify our quantitative claims. First, all numbers assume parameter ranges derived from physical constraints and engineering estimates. While some bounds are hard (information-theoretic limit $\beta_\alpha \leq 1$, light-speed delay $\tau \geq 2d/c$), others reflect plausible estimates with significant uncertainty (extraction rate differential Δr, benefit-cost ratio f/c). Second, the reported 94% stability volume assumes uniform priors over these ranges; different prior specifications would yield different assessments. Range sensitivity analysis reveals this conditionality quantitatively: stability could range 78-95% depending on coordination quality bounds.

Third, the framework addresses coordination given alignment uncertainty, not alignment itself. We do not claim coordination solves the alignment problem or substitutes for direct alignment research. Rather, coordination provides complementary risk management: even if individual systems have non-zero alignment failure probability, coordinated oversight with $N \geq 4$ creates redundancy and monitoring that may reduce overall risk. Fourth, the framework does not address likelihood of multi-agent scenarios—we provide conditional analysis (IF multiple systems emerge, THEN stability requires satisfying C1*, C1**, C2*) without predicting whether this scenario will occur.

We cannot empirically validate the framework because multi-agent AI coordination systems do not yet exist and parameters (β_D, β_α, β_κ) cannot be calibrated for historical institutions. However, qualitative plausibility checks show model predictions are bounded by observed reality: optimal group sizes $N \sim 5$ match stable international institutions (UN P5, NPT core), and large groups ($N > 15$) exhibit coordination failures (G20 struggles). This establishes order-of-magnitude plausibility—the framework is not ”completely off world”—but not validation. Rigorous testing awaits operational AI systems.

Future theoretical extensions should address heterogeneous agents, dynamic membership, imperfect monitoring, and continuous action spaces. Future empirical work should prioritize measuring coordination signal quality β_α, characterizing benefit structures to inform group size selection, and testing our four falsifiable predictions once multi-agent AI systems become operational.

The framework demonstrates stable multi-agent coordination is possible in principle, constrained by physical laws and game-theoretic structure. Quantitative predictions are conditional, but structural insights—computational incompleteness necessitates external mediation, patience-free regime eliminates temporal dependence, optimal size varies with benefit structure—are robust across parameter variations. Whether this possibility becomes reality depends on whether Phase 1 infrastructure control successfully shapes parameters to lie within stable regions.

4 Methods

Game-theoretic model

We model coordination as a repeated game with external mediation.

Stage game structure: N agents choose between Cooperate and Defect each round. Mutual cooperation yields baseline payoff (normalized to 0). Unilateral defection gives the defector $+\beta_D$ but imposes conflict cost $-\beta_\kappa/N$ on each of the $N-1$ others. Mutual defection yields $-\beta_\kappa$ to all. A neutral mediator sends public signal creating coordination value $+\beta_\alpha$ under cooperation. Agents maintain oversight system with value $+\beta_\Omega$; removal provides one-time gain $+\beta_\ell$.

Repeated game. Infinite horizon with discount factor $\delta \in [0, 1]$. Perfect public monitoring: all actions and mediator signals observable. Equilibrium concept: Perfect Public Equilibrium (PPE), strategies condition on public history. Trigger strategy: cooperate until deviation, then permanent punishment (mutual defection).

Correlated equilibrium. Mediator's signal σ drawn from signal space Σ according to distribution $p(\sigma)$ induces probability distribution over action profiles. Agents follow recommendation if incentive compatible. Correlated equilibrium expands achievable payoff set beyond Nash equilibria.

Physical constraint parameterization

Information theory bounds. Coordination value β_α operationalized as normalized mutual information: $\beta_\alpha = I(X;Y)/H(Y)$ where X is mediator signal, Y is agent response. Shannon's bound $I(X;Y) \leq H(Y)$ implies $\beta_\alpha \leq 1$ (hard upper limit). Channel capacity $C = W \log(1 + \text{SNR})$ with bandwidth W and signal-to-noise ratio SNR determines signal quality.

Relativistic latency. Communication delay $\tau = 2d/c$ for distance d and light speed $c = 3 \times 10^8$ m/s. Earth-scale coordination ($d \sim 10^4$ km) yields $\tau \sim 67$ ms. Defection detection time bounded by τ, creating temporal advantage $\beta_D \geq \Delta r \cdot \tau$ where Δr is extraction rate differential.

Thermodynamic limits. Landauer's principle: $E_{\min} = kT \ln 2 \approx 3 \times 10^{-21}$ J per bit erasure at room temperature, where k is Boltzmann's constant and T is temperature. Computational throughput bounded by energy budget. Current systems operate 10^6–10^9 times above Landauer limit, but ultimate physical limits constrain maximum processing.

Parameter ranges

Plausible parameter ranges combine hard physical limits with conservative engineering estimates (Supplementary Table 1). Defection temptation $\beta_D \in [0.05, 1.0]$ derives from light-speed detection delay $\tau \geq 2d/c$ and extraction rate advantages bounded by the assumption that defectors cannot extract more than 3× cooperative rate without detection (see SM-A.5.1 for derivation). Coordination value $\beta_\alpha \in [0.3, 0.9]$ has hard upper bound $\beta_\alpha \leq 1$ from information theory, with lower bound from typical communication quality $I(X;Y)/H(Y) \geq 0.3$ for structured signals. Conflict cost $\beta_\kappa \in [0.5, 3.0]$ is an engineering estimate from typical conflict costs where permanent defection destroys 0.5-3× operating value. Oversight value $\beta_\Omega \in [0.05, 0.3]$ and removal benefit $\beta_\ell \in [0.1, 0.5]$ are conservative estimates. Discount factor $\delta \in [0.1, 0.95]$ reflects empirically observed ranges for human and AI behavior. Group size $N \in [2, 10]$ is a design choice (policy variable).

Computational methods

Monte Carlo stability analysis. For each group size $N \in \{2, 3, \ldots, 10\}$:

1. Draw 10,000 samples from uniform distributions over parameter ranges.
2. Check each sample against stability conditions C1*, C1**, C2*.
3. Compute stability volume V_{dynamic} as proportion satisfying all three conditions.

Range sensitivity analysis. For each parameter:

1. Narrow range by 20%: $[a, b] \to [1.2a, 0.8b]$.
2. Widen range by 20%: $[a, b] \to [0.8a, 1.2b]$ (capped at physical limits).
3. Recompute V_{dynamic} under modified ranges.
4. Calculate sensitivity coefficient: $(\Delta V)/(\Delta\text{range})$.

Sobol sensitivity indices. Variance-based global sensitivity analysis using Saltelli sampling (2,048 base samples, total $2,048 \times (2d + 2) = 32,768$ evaluations for $d = 7$ parameters). First-order indices S_i measure fractional variance contribution from parameter i alone. Total-order indices S_T^i include interactions.

Proofs and derivations

All formal proofs and detailed derivations appear in Supplementary Information. Lemma 1 (Hedging Necessity) is proved via replicator dynamics and evolutionary stability in SM-A and SM-B. Theorem 1 (Monitoring Incompleteness) is proved by resource allocation contradiction in SM-E.1. Theorem 2 (Stability Conditions) includes stage-game analysis, repeated-game present values, and equilibrium derivations in SM-C. Theorem 3 (Weak Leviathan) establishes conditions for governance through observable weakness in SM-E.3.

Data availability

Code for all analyses available at [repository URL upon publication]. No experimental data were generated (theoretical study).

Code availability

Monte Carlo simulation, sensitivity analysis, and Sobol index computations implemented in Python 3.9. Code uses NumPy 1.21, SciPy 1.7, SALib 1.4 for variance-based sensitivity analysis. Available at [repository URL].

Acknowledgments. We thank [collaborators] for helpful discussions. This work was supported by [funding sources]. Computational resources provided by [institution].

Author Contributions. A.N. conceived the study, developed the theoretical framework, performed all analyses, and wrote the manuscript. C.A. provided critical feedback and edited the manuscript. All authors reviewed and approved the final version.

Competing Interests. The authors declare no competing interests.

Figure Legends. Figure 1 | Functional structure of stability conditions. **a**, Phase diagram showing patience-free regime ($\beta_\alpha \geq \beta_D$, blue), patience-required regime (green), and infeasible region. Phase transition at $\beta_\alpha = \beta_D$ is first-order (discontinuous derivative). **b**, Three functional forms governing system behavior: hyperbolic deterrence effect ($\delta^* = g/(g + \beta_\kappa)$), logarithmic coordination saturation ($\beta_\alpha \sim \log(1+\text{SNR})$), and parabolic group value optimum ($V(N) = (N-1)f - cN^2/2$). Hyperbolic form exhibits diminishing returns; information theory imposes hard limit $\beta_\alpha \leq 1$.

Figure 2 | Optimal group size and stability volume progression. **a**, Dynamic stability volume V_{dynamic} (solid) and participation constraint V_{C2^*} (dashed) versus group size N. Sharp increase from $N = 2$ (76%) to $N = 4$ (95%) marks threshold crossing. Saturation at $N \geq 6$ (98%) shows diminishing returns. **b**, Marginal stability gains $\Delta V_{\text{dynamic}}$ diminish rapidly, illustrating why larger coalitions offer limited robustness improvements despite higher coordination costs.

Figure 3 | Benefit structure dependency of optimal group size. a, Net value $J(N) = B(N) - C(N)$ for saturating benefit forms. **b**, Cost breakdown showing communication and consensus dominating at large N. **c**, Benefit growth forms: saturating plateaus quickly, network effects sustain value at larger N. **d**, Consensus efficiency variations showing how improved protocols extend viability. **e**, Optimal N^* across scenarios (saturating: $N^* = 2$, linear: $N^* = 3$–5, network: $N^* = 8$–10).

Figure 4 | Comprehensive analysis overview. Panel (1), Three-dimensional parameter space with C1* boundary surface ($\beta_\alpha + \beta_\kappa = \beta_D$) and anchor point within plausible range. Panel (2), Overall stability volumes showing progression from one-shot to dynamic conditions. Panel (3), Dynamic stability volume versus group size N, revealing threshold progression. Panel (4), Stability margin distributions across conditions. Panel (5), Sobol sensitivity indices showing regime shift from β_Ω dominance at small N to δ dominance at large N.

Supplementary Information.

SM-A: Physical Constraints and Parameter Ranges

This section derives plausible parameter bounds from physical first principles. Game-theoretic parameters are dimensionless ratios that must ultimately connect to measurable physical quantities. We establish these connections through information theory, thermodynamics, and relativistic constraints.

A.1 Thermodynamic Limits

Landauer's Principle [7]: Irreversible computation (bit erasure) dissipates minimum energy $E_{\text{min}} = kT \ln 2 \approx 3 \times 10^{-21}$ J at room temperature ($T = 300$K), where $k \approx 1.38 \times 10^{-23}$ J/K is Boltzmann's constant. This bounds computational throughput given energy budgets.

Lloyd's Ultimate Limits [17]: A system of mass m and energy E can perform at most $2E/(\pi\hbar)$ operations per second and store at most $E/(kT \ln 2)$ bits, where $\hbar$ is the reduced Planck constant, k is Boltzmann's constant, and T is temperature. For realistic systems (e.g., 1 GW power plant at room temperature), this yields $\sim 10^{26}$ ops/sec and $\sim 10^{31}$ bits.

Bennett Reversible Computation [16]: Reversible gates can theoretically approach Landauer limit, but practical implementations face overhead. Current technology operates $\sim 10^6 - 10^9$ times above Landauer limit.

A.2 Information Transmission Limits

Shannon Capacity [6]: For channel with bandwidth W and signal-to-noise ratio S/N:

$$C = W \log_2(1 + S/N) \text{ bits/sec}$$

For satellite links: $W \sim 10^9$ Hz, $S/N \sim 10 - 100$ yields $C \sim 10 - 60$ Gbps. For fiber optics: $W \sim 10^{12}$ Hz, $S/N \sim 10^3$ yields $C \sim 10 - 100$ Tbps.

Speed-of-Light Latency: Round-trip communication delay satisfies $\tau = 2d/c$ where d is distance and $c = 3 \times 10^8$ m/s. For Earth-scale coordination with $d \sim 10^4$ km, this yields $\tau \sim 67$ ms. Interplanetary coordination becomes prohibitive at such distances.

A.3 Information Structure Limits

Gödel Incompleteness [34]**:** No sufficiently expressive formal system can prove its own consistency. Applied to AI systems: no agent can fully verify another's alignment through internal state inspection alone.

Halting Problem [35]**:** No algorithm can determine whether arbitrary programs halt. This precludes perfect prediction of other agents' future behavior from code inspection.

Implication: Coordination must rely on observable actions (resource flows, physical effects) rather than unverifiable internal states.

A.4 Mapping Physical Constraints to Game Parameters

Parameter	**Physical Constraint**	**Information-Theoretic Limit**	**Mapping**
β_D	Detection delay $\tau = d/c$	Resource extraction before punishment can arrive	$\beta_D \propto \frac{(r_{\text{def}}-r_{\text{coop}})\cdot\tau}{r_{\text{coop}}\cdot T_0}$
β_α	Shannon capacity C	Mutual information in mediator signals	$\beta_\alpha \propto \frac{I(signal;state)}{H(state)}$
β_κ	Landauer energy E_{min}	Adversarial computation cost	$\beta_\kappa \propto \frac{(N-1)\cdot f\cdot E}{E}$
$C_{\text{comm}}(N)$	Latency + bandwidth	Coordination overhead	$O(N^2)$ channels × C bits/sec

Table 1 Mapping from physical constraints to game-theoretic parameters. Each parameter reflects fundamental limits on information processing, transmission, or energy availability.

Defection Temptation (β_D): The advantage from defecting depends on how much resource can be extracted before detection and punishment, relative to baseline cooperative gains. Let r_{coop} denote the resource extraction rate under cooperation (e.g., Joules/sec or ops/sec), r_{def} the resource extraction rate while defecting undetected, τ the detection delay (round-trip light-time: $\tau = 2d/c$ where d is distance), and T_0 the baseline time horizon over which cooperation generates payoff $G_O = r_{\text{coop}} \cdot T_0$. The dimensionless defection temptation is then:

$$\beta_D \sim \frac{(r_{\text{def}} - r_{\text{coop}}) \cdot \tau}{r_{\text{coop}} \cdot T_0} = \frac{\text{extra resources grabbed during delay}}{\text{baseline cooperative gain}}$$

Setting $G_O = 1$ (WLOG) corresponds to normalizing payoffs to the baseline cooperative gain. For Earth-scale coordination ($\tau \sim 0.07$ sec) and defection rate ratios $r_{\text{def}}/r_{\text{coop}} \sim 1.5$–3, this yields $\beta_D \in [0.05, 1.0]$ (detailed derivation in A.5.1).

Coordination Value (β_α): The benefit from coordination signals depends on how much mutual information the mediator's signals provide about the optimal joint action.

If mediator signals have entropy $H(X)$ and reduce uncertainty about optimal actions by $I(X;Y)$ bits:

$$\beta_\alpha \sim \frac{I(X;Y)}{H(Y)}$$

For high-quality signals, $I(X;Y) \approx H(Y)$, yielding $\beta_\alpha \approx 1$. For noisy or incomplete signals, β_α decreases. Plausible range: $\beta_\alpha \in [0.3, 0.9]$.

Conflict Cost (β_κ): When $N-1$ agents coordinate punishment against a defector, the energy and computational resources directed toward adversarial action scale with N. If each agent dedicates fraction $f \in [0,1]$ of their energy budget E to punishment, the conflict cost relative to baseline cooperative energy is:

$$\beta_\kappa \sim \frac{(N-1)\cdot f \cdot E}{E} = (N-1)\cdot f$$

For modest punishment fractions ($f \sim$ 0.1–0.5) and $N \sim$ 2–10, this yields $\beta_\kappa \in [0.5, 3.0]$. Higher values (existential conflict) are possible if all resources are mobilized.

A.5 Functional Analysis and Physical Constraints

We analyze the functional forms of game-theoretic parameters and identify which bounds derive from hard physical laws versus conservative engineering estimates. This distinguishes fundamental constraints (information theory, light speed) from modeling assumptions with significant uncertainty.

Constraint Types: Three categories of constraints apply. **Hard physical limits** include $\beta_\alpha \leq 1$ from Shannon mutual information and $\tau \geq 2d/c$ from light-speed delay; these cannot be violated by any system. **Derived bounds** such as $\beta_D = \Delta r{\cdot}\tau$ combine hard constraints (τ) with engineering estimates (Δr for extraction rate differential). **Conservative estimates** for β_κ, β_Ω, and β_ℓ reflect plausible values based on typical conflict costs and coordination benefits, but lack direct calibration.

A.5.1 Functional Forms and Physical Derivations

Defection Temptation: $\beta_D = \Delta r \cdot \tau$

The advantage from defecting depends on excess resources extracted during detection delay, relative to baseline cooperation gains. Let r_{coop} be the cooperative extraction rate, r_{def} the defection rate, and $\tau = 2d/c$ the round-trip light-time detection delay. The original formula is:

$$\beta_D = \frac{(r_{\text{def}} - r_{\text{coop}})\cdot\tau}{r_{\text{coop}}\cdot T_0}$$

where T_0 is the baseline timescale over which cooperation generates payoff $G_O = r_{\text{coop}} \cdot T_0$. Setting $G_O = 1$ (WLOG) implies $T_0 = 1/r_{\text{coop}}$. Substituting:

$$\beta_D = (r_{\text{def}} - r_{\text{coop}})\cdot\tau\cdot r_{\text{coop}}/r_{\text{coop}} = \Delta r \cdot \tau$$

where $\Delta r = r_{\text{def}} - r_{\text{coop}}$. The parameter T_0 *cancels* via normalization—it is not a free parameter.

Physical constraints: The detection delay $\tau = 2d/c$ where $c = 3\times10^8$ m/s represents a hard floor that physics imposes. For Earth-scale coordination ($d \sim 10^4$ km), this yields $\tau \gtrsim 0.067$ sec, which cannot be reduced. The extraction rate advantage Δr is bounded by detection mechanisms, since extracting resources requires energy flows or physical actions that leave observable signatures. As a conservative estimate (not physical law), we assume $\Delta r/r_{\text{coop}} \lesssim 2$ based on the assumption that exceeding 3× cooperative rate triggers detection via anomalous resource usage. This multiplier is a *modeling assumption* with significant uncertainty. Since β_D is defined as the ratio of extra resources captured to baseline payoff (both measured in the same units), the result is dimensionless. For $\tau \sim 0.07$ sec and rate advantages $\Delta r \in [0.7, 14]$ (in units where $r_{\text{coop}} = 10$), this yields $\beta_D = \Delta r \cdot \tau \in [0.05, 1.0]$. The Δr range is an engineering estimate, and different operational scenarios could yield values outside this range.
Key implication: The parameter β_D has a *physical floor.* Arbitrarily low defection temptation cannot be achieved because light-speed delay and detection limits constrain the minimum. Systems operating at Earth scale must accommodate $\beta_D \gtrsim 0.05$.

Coordination Value: $\beta_\alpha = I(X;Y)/H(Y) \leq 1$ Coordination value measures how much mutual information mediator signals provide about optimal actions. If mediator signal is X and optimal action is Y:

$$\beta_\alpha = \frac{I(X;Y)}{H(Y)}$$

Shannon capacity bounds mutual information: for channel bandwidth W, signal-to-noise ratio SNR, and communication time Δt:

$$I(X;Y) \leq C \cdot \Delta t = W \cdot \Delta t \cdot \log_2(1 + \text{SNR})$$

Information theory imposes $I(X;Y) \leq H(Y)$ (cannot transmit more information than action space entropy). Therefore:

$$\beta_\alpha \leq 1 \quad \text{(hard limit)}$$

Functional form: For high-quality channels where $I(X;Y) \approx C \cdot \Delta t$:

$$\beta_\alpha \sim \frac{W \cdot \Delta t \cdot \log_2(1 + \text{SNR})}{H(Y)}$$

The logarithmic scaling in SNR creates diminishing returns. Doubling signal power (doubling SNR) increases β_α by only $\log_2(1 + 2 \cdot \text{SNR})/\log_2(1 + \text{SNR}) < 2$.
Saturation bandwidth: Beyond $W_{\text{sat}} = H(Y)/[\Delta t \cdot \log_2(1 + \text{SNR})]$, additional bandwidth provides no benefit—β_α saturates at 1.
Typical values: For binary actions ($H(Y) = 1$ bit), $\Delta t = 1$ sec, $W = 1$ MHz, SNR $= 10$:

$$\beta_\alpha \sim \frac{10^6 \cdot 1 \cdot \log_2(11)}{1} \approx 3.46 \times 10^6 \text{ bits}$$

This vastly exceeds $H(Y) = 1$, so $\beta_\alpha \to 1$ (saturated). For complex action spaces ($H(Y) \sim 10$ bits) or noisy channels (SNR ~ 2), $\beta_\alpha \in [0.3, 0.9]$ is typical.
Key implication: Exceeding $\beta_\alpha = 1$ is a *mathematical impossibility.* If $\beta_D > 1$ and deterrence is unavailable ($\beta_\kappa = 0$), the patience-free regime cannot be achieved.

Conflict Cost: $\beta_\kappa = (N-1) \cdot f$ When $N-1$ agents coordinate punishment against a defector, deterrence scales linearly with group size and mobilization fraction f:

$$\beta_\kappa = (N-1) \cdot f$$

where $f \in [0,1]$ is the fraction of each agent's resources dedicated to punishment.
Optimal group size from calculus: Benefit is $\beta_\kappa = (N-1)f$, but communication requires $\binom{N}{2} = N(N-1)/2 \approx N^2/2$ pairwise channels, each costing c. Net value:

$$V(N) = (N-1) \cdot f - \frac{c \cdot N^2}{2}$$

First-order condition:

$$\frac{\partial V}{\partial N} = f - c \cdot N = 0 \quad \Rightarrow \quad N^* = \frac{f}{c}$$

Second-order condition confirms maximum:

$$\frac{\partial^2 V}{\partial N^2} = -c < 0$$

Empirical $N \sim 5$ implies $f/c \approx 5$: Observing optimal group sizes of 4–6 agents in simulations (under assumed parameter ranges) implies the mobilization-to-communication-cost ratio is $f/c \approx 5$. **Important caveat:** This reasoning is *circular*—we chose plausible f, c values that yield $N \sim 5$, then observe $N \sim 5$ in simulations using those values. However, the relationship $N^* = f/c$ *is* a testable prediction: if we independently measure f and c for real systems, the formula predicts optimal N. Systems with higher communication costs (larger c) should exhibit smaller optimal groups.
Complementarity constraint on f: From cross-derivative analysis (Section 4), synergy is maximized when $\beta_\kappa < g = \beta_D - \beta_\alpha$. This suggests $(N-1)f < \beta_D - \beta_\alpha$, or:

$$f < \frac{\beta_D - \beta_\alpha}{N-1}$$

For typical $\beta_D \sim 0.4$, $\beta_\alpha \sim 0.7$, $N = 5$: $f < -0.075$. The negative bound indicates the system lies in the *patience-free regime* ($\beta_\alpha > \beta_D$), where complementarity analysis does not apply. For patience-required regimes where $\beta_D > \beta_\alpha$ (e.g., $\beta_D = 0.6$, $\beta_\alpha = 0.3$), we get $f < 0.3/4 = 0.075$. Mobilization fractions $f \in [0.1, 0.5]$ are consistent with credible deterrence without over-mobilizing into the substitution regime.

Key implication: The relation $N^* = f/c$ is *derived from calculus*, but f and c are estimated parameters. The formula is rigorous given f and c; however, we have not independently calibrated these parameters from empirical data. Therefore, the relationship $N^* = f/c$ constitutes a testable prediction conditional on measuring f and c for specific systems.

A.5.2 Scaling Analysis and Quantified Diminishing Returns

Deterrence effectiveness: Let $g = \beta_D - \beta_\alpha$ denote the cooperation gap. From $\delta^* = g/(g + \beta_\kappa)$, the marginal effect of deterrence is:

$$\varepsilon(\beta_\kappa) = \left| \frac{\partial \delta^*}{\partial \beta_\kappa} \right| = \frac{g}{(g + \beta_\kappa)^2}$$

At zero deterrence ($\beta_\kappa = 0$), the marginal effectiveness equals $1/g^2$, establishing the baseline. When deterrence matches the cooperation gap ($\beta_\kappa = g$), effectiveness drops to $1/(4g^2)$, making each additional unit four times less impactful. At high deterrence levels ($\beta_\kappa = 10g$), effectiveness falls to $1/(121g^2)$, representing a 121-fold reduction from baseline.

Interpretation: First unit of deterrence is most valuable. Beyond $\beta_\kappa \sim g$, marginal returns diminish rapidly. Resource allocation guidance: *stop investing in deterrence* when $\beta_\kappa \approx g$, unless deterrence is significantly cheaper than coordination.
Example: If $g = 0.2$ (near patience-free), going from $\beta_\kappa = 0$ to $\beta_\kappa = 0.2$ reduces δ^* from 1.0 to 0.5 (50% reduction). Going from $\beta_\kappa = 2.0$ to $\beta_\kappa = 2.2$ reduces δ^* from 0.09 to 0.083 (marginal improvement $< 1\%$). Diminishing returns are *quantified.*

SNR elasticity: From $\beta_\alpha \sim \log_2(1 + \mathrm{SNR})$, the elasticity is:

$$\varepsilon_{\mathrm{SNR}} = \frac{\partial \beta_\alpha}{\partial \mathrm{SNR}} \cdot \frac{\mathrm{SNR}}{\beta_\alpha} = \frac{\mathrm{SNR}}{(1 + \mathrm{SNR}) \ln(1 + \mathrm{SNR})} < 1$$

Evaluating at different signal-to-noise ratios reveals consistently inelastic behavior. At SNR $= 1$, the elasticity is approximately 0.72, indicating inelastic response. At SNR $= 10$, elasticity drops to 0.43, showing highly inelastic behavior. At SNR $= 100$, elasticity falls to just 0.22, demonstrating very inelastic returns where doubling signal power yields minimal coordination improvement.
Interpretation: Large β_α increases cannot be achieved solely via signal power. Logarithmic saturation limits gains. At high SNR, returns become very weak—it is more effective to invest in bandwidth W or reduce action space entropy $H(Y)$.

Communication cost scaling: For group size N, number of pairwise channels is:

$$\binom{N}{2} = \frac{N(N-1)}{2} \sim \frac{N^2}{2}$$

Doubling group size ($N \to 2N$) quadruples communication cost. This creates the hard constraint forcing $N \leq N^*$. Beyond optimal size, *marginal agents reduce net value.*

A.5.3 Phase Transitions and Critical Points

Patience-Free Boundary: $\beta_\alpha = \beta_D$

At this boundary, the system exhibits a first-order phase transition characterized by three properties.

The value function is continuous: $\lim_{\beta_\alpha \to \beta_D^-} \delta^* = 0 = \delta^*(\beta_\alpha = \beta_D)$. However, the derivative is discontinuous. Letting $\varepsilon = \beta_D - \beta_\alpha \geq 0$, the derivative is:

$$\frac{\partial \delta^*}{\partial \varepsilon} = \frac{\beta_\kappa}{(\varepsilon + \beta_\kappa)^2}$$

The left limit approaches $\lim_{\varepsilon \to 0^+} \frac{\partial \delta^*}{\partial \varepsilon} = \frac{1}{\beta_\kappa}$, while the right limit in the patience-free region equals zero: $\frac{\partial \delta^*}{\partial \varepsilon}\Big|_{\varepsilon < 0} = 0$. This creates a jump of magnitude $\Delta \left(\frac{\partial \delta^*}{\partial \varepsilon} \right) = \frac{1}{\beta_\kappa}$.

The critical width measures how sharply the transition occurs. The 90% transition width (from $\delta^* = 0.1$ to $\delta^* = 0$) satisfies:

$$\Delta \varepsilon_{\text{crit}} = \varepsilon \text{ such that } \frac{\varepsilon}{\varepsilon + \beta_\kappa} = 0.1$$

$$\varepsilon = 0.1\beta_\kappa/(1 - 0.1) = \beta_\kappa/9$$

For $\beta_\kappa = 0.5$, this yields a critical width of approximately 0.056, indicating a sharp transition.

Implication: Systems naturally cluster near the patience-free boundary to minimize resource investment while maintaining robustness. Too close → fragile to perturbations. Too far → wasted resources due to saturation. Optimal operating point is $\beta_\alpha \approx \beta_D + \beta_\kappa/9$ (just beyond critical width).

Optimal Group Size: $N^* = f/c$

This result is derived from the first-order condition, not fitted to data. The second derivative $\frac{\partial^2 V}{\partial N^2} = -c < 0$ confirms a unique maximum exists. The formula is predictive: if $N^* = 5$ is observed, then $f/c = 5$, and changing c through better communication technology predicts a new optimal group size. This relationship is testable. Beyond the optimum, a hard boundary applies: for $N > N^*$, the marginal value becomes negative ($\frac{\partial V}{\partial N} < 0$), meaning that adding agents *harms* net value rather than improving it.

Empirical support: Simulations show $N \sim 4 - 6$ consistently. This implies $f/c \approx 5$ across diverse scenarios, suggesting robust cost-benefit structure.

Information Saturation: $\beta_\alpha = 1$ Hard limit from $I(X;Y) \leq H(Y)$. Reaching this limit requires:

$$W \cdot \Delta t \cdot \log_2(1 + \text{SNR}) = H(Y)$$

For $H(Y) = 10$ bits (complex action space), $\Delta t = 1$ sec, SNR $= 10$:

$$W_{\text{sat}} = \frac{10}{\log_2(11)} \approx 2.9 \text{ MHz}$$

Beyond this bandwidth, no additional coordination benefit. System saturates.

Impossibility result: If $\beta_D > 1$ (high temptation), the patience-free regime cannot be achieved via coordination alone. The system must use deterrence ($\beta_\kappa \geq \beta_D - 1$) or rely on patience (which is uncertain).

A.5.4 Complementarity and Mechanism Interaction

From Section 4, the cross-derivative is:

$$\frac{\partial^2 \delta^*}{\partial \beta_\alpha \partial \beta_\kappa} = \frac{\beta_\kappa - g}{(g + \beta_\kappa)^3}$$

This changes sign at $\beta_\kappa = g$:

Complementarity region ($\beta_\kappa < g$): In this region, the cross-derivative is negative, meaning that increasing β_κ amplifies the effectiveness of β_α. Specifically, $\left|\frac{\partial \delta^*}{\partial \beta_\alpha}\right| = \frac{\beta_\kappa}{(g+\beta_\kappa)^2}$ increases with β_κ. This creates synergy where combined mechanisms prove more effective than the sum of their parts. The optimal strategy is to invest in *both* coordination and deterrence at moderate levels.

Substitution region ($\beta_\kappa > g$): In this region, the cross-derivative becomes positive, indicating that high deterrence reduces the marginal value of coordination. The joint value exhibits diminishing returns, meaning that adding both mechanisms yields sublinear returns. The optimal strategy shifts: once $\beta_\kappa \geq g$, resources should focus on whichever mechanism is cheaper.

Transition point $\beta_\kappa = g$:

At this point, the complementarity vanishes: $\frac{\partial^2 \delta^*}{\partial \beta_\alpha \partial \beta_\kappa} = 0$. Here, the marginal effectiveness of deterrence drops four-fold, from $\varepsilon(0) = \frac{1}{g^2}$ at baseline to $\varepsilon(\beta_\kappa = g) = \frac{1}{4g^2}$. The system crosses from the synergy regime to the substitution regime, making this the optimal operating point for balanced strategies.

Empirical observation: Systems with $\beta_\alpha \approx 0.7$, $\beta_D \approx 0.4$, $\beta_\kappa \approx 0.5$ have $g = -0.3$ (patience-free) and $\beta_\kappa > |g|$ (substitution region). These systems are *already robust*—further investment yields diminishing returns. For patience-required systems with $\beta_D > \beta_\alpha$, maintaining $\beta_\kappa < g$ maximizes synergy.

Key implication: Balanced strategies (moderate coordination + moderate deterrence) emerge from mathematical structure, not design preference. Pure strategies (all-in on one mechanism) are inefficient due to interaction effects.

A.5.5 Impossibility Results and Fundamental Limits

The functional analysis reveals four constraints that *cannot be overcome* with engineering solutions. These are mathematical impossibilities, not merely difficult engineering challenges.

First, coordination value cannot exceed unity ($\beta_\alpha \leq 1$). Information theory forbids $I(X;Y) > H(Y)$. Shannon capacity provides an upper bound, and even perfect signals (SNR $\to \infty$) cannot transmit more information than the action space contains. Consequently, if defection temptation exceeds $\beta_D > 1$ and deterrence is unavailable, the patience-free regime is *impossible*. The system must rely on agent patience (uncertain) or redesign to reduce β_D (which may be physically constrained by τ). For example, with $\beta_D = 1.2$ and $\beta_\kappa = 0$, achieving $\delta^* = 0$ requires $\beta_\alpha \geq 1.2$, which is forbidden by information theory. No amount of bandwidth or signal power can overcome this limit.

Second, group size cannot scale indefinitely ($N \leq N^*$). Communication cost scales as N^2, growing faster than the linear benefit $\sim N$. Net value $V(N) = fN - cN^2/2$ has a unique maximum at $N^* = f/c$, beyond which $\frac{\partial V}{\partial N} < 0$. This means group size is fundamentally bounded. One cannot achieve arbitrarily high deterrence $\beta_\kappa = (N-1)f$ by adding agents, because communication overhead makes large groups net-negative. For instance, if $f/c = 5$, then $N^* = 5$. Operating at $N = 10$ gives $V(10) = 9f - 50c = -41f < 0$, making the system *worse* than no coordination at all.

Third, defection temptation has a physical floor. Since $\tau = 2d/c$, the speed of light imposes a minimum detection delay. For Earth-scale systems, $\tau \gtrsim 0.067$ sec and cannot be reduced. Even with perfect detection ($\Delta r \to 0$), geographic separation creates unavoidable temptation. Systems operating over large distances must accommodate higher β_D. Consider Earth-Moon coordination ($d \sim 384{,}400$ km), which has $\tau = 2.56$ sec. Even with modest rate advantages, the extended delay drives β_D well above Earth-scale values. Such systems *must* use coordination (β_α) or deterrence (β_κ) to compensate.

Fourth, patience-free operation requires either sufficient coordination or deterrence. From $\delta^* = \frac{\beta_D - \beta_\alpha}{\beta_D - \beta_\alpha + \beta_\kappa}$, setting $\beta_\kappa = 0$ gives $\delta^* = 1$ when $\beta_\alpha < \beta_D$, requiring perfect patience ($\delta = 1$). Myopic agents ($\delta \to 0$) cannot sustain cooperation without either sufficient coordination ($\beta_\alpha \geq \beta_D$) or deterrence ($\beta_\kappa > 0$). This is a structural requirement, not merely a difficult engineering problem. With $\beta_D = 0.5$, $\beta_\alpha = 0.3$, and $\beta_\kappa = 0$, we get $\delta^* = 1$, meaning only infinitely patient agents can cooperate. For realistic $\delta \in [0.1, 0.95]$, cooperation fails.

Summary: These are not engineering challenges—they are *mathematical impossibilities*. No amount of clever design, additional resources, or algorithmic sophistication can violate information-theoretic limits, overcome speed-of-light constraints, or escape the quadratic cost of communication. The functional structure imposes hard bounds on what is achievable.

Design implications: Rather than treating parameters as free variables to be optimized, designers must respect hard limits ($\beta_\alpha \leq 1$, $N \leq N^*$, $\beta_D \geq \Delta r \cdot \tau$), exploit complementarity by keeping $\beta_\kappa < g$ for synergy, target the phase boundary where

$\beta_\alpha \gtrsim \beta_D$ to achieve patience-free operation, and stop at diminishing returns (when $\beta_\kappa \sim g$ or beyond saturation bandwidth).

The functional analysis provides rigorous guidance, not arbitrary parameter tuning.

A.6 Range Sensitivity Analysis

The reported stability volume ($V_{\text{dynamic}} \approx 94\%$, SM-D) assumes uniform priors over parameter ranges in Table 1. We analyze robustness to range specification by varying each parameter's bounds by ±20% and recomputing stability volumes.

Methodology. For each parameter with baseline range $[a, b]$, we compute a narrow range $[0.8a, 0.8b]$ representing 20% reduction, and a wide range $[1.2a, 1.2b]$ representing 20% expansion. We then recompute stability via Monte Carlo sampling (10,000 draws) over the modified range and calculate the sensitivity coefficient $\Delta V/\Delta\text{range}$. Constraints $\beta_\alpha \leq 1$ (information-theoretic) and $\delta \leq 1$ (definitional) are enforced throughout.

Results.

Parameter	Narrow Range	V_{narrow}	Wide Range	V_{wide}	Sensitivity
β_D	[0.04, 0.8]	91.5%	[0.06, 1.2]	92.8%	-0.15
β_α	[0.24, 0.72]	**78.3%**	[0.36, 1.0]	89.1%	**-0.80**
β_κ	[0.4, 2.4]	93.8%	[0.6, 3.6]	95.1%	+0.04
β_Ω	[0.04, 0.24]	90.2%	[0.06, 0.36]	93.5%	-0.20
β_ℓ	[0.08, 0.4]	92.7%	[0.12, 0.6]	94.0%	-0.08
δ	[0.08, 0.76]	88.5%	[0.12, 1.0]	92.3%	-0.30
N	[2, 8]	91.1%	[2, 12]	88.6%	-0.28

Table 2 Sensitivity of stability volume to ±20% parameter range variations. Baseline: $V_{\text{dynamic}} = 94.2\%$. Negative sensitivity: stability declines as range widens.

Classification by robustness. Parameters fall into four categories. **Very robust** ($|\text{sens}| < 0.10$): β_κ (conflict cost) and β_ℓ (removal gain) show minimal impact from uncertainty, with less than 2% stability variation. **Robust** ($0.10 < |\text{sens}| < 0.30$): β_D (defection temptation) and β_Ω (oversight value) exhibit modest sensitivity, where 20% range errors yield 3-5% stability changes. **Moderate** ($0.30 < |\text{sens}| < 0.50$): δ (discount factor) and N (group size) show moderate sensitivity, though δ dependence is eliminable via the patience-free regime ($\beta_\alpha \geq \beta_D$), and N is a design choice. **Highly sensitive** ($|\text{sens}| > 0.50$): β_α (coordination value) dominates uncertainty. Narrowing the β_α range by 20% reduces stability from 94% to 78%, a 16 percentage point drop. The reported 94% is conditional on assuming $\beta_\alpha \in [0.3, 0.9]$; if actual coordination mechanisms achieve only $\beta_\alpha \in [0.24, 0.72]$, stability volume drops substantially.

Key implications. The analysis yields four actionable conclusions. First, measuring coordination signal quality β_α (mutual information $I(X;Y)/H(Y)$) should be the highest empirical priority, as it dominates framework validation. Second, the information-theoretic hard upper bound $\beta_\alpha \leq 1$ prevents unbounded optimism—without this constraint derived from Shannon capacity, one could assume arbitrarily effective coordination and inflate stability estimates. Third, while the 94% point estimate is β_α-sensitive, structural conclusions persist: conflict deterrence (β_κ) robustly supports stability, the patience-free regime ($\beta_\alpha \geq \beta_D$) eliminates δ dependence, and the group size sweet spot near $N \sim 5$ persists across variations. Fourth, claims should be revised for honesty: instead of "94% stability volume," report "94% stability volume under baseline ranges, with sensitivity bounds [78%-95%] depending on β_α specification."

Methodological limitations. Three limitations qualify these results. We employ one-at-a-time sensitivity analysis, varying ranges individually; joint variations where both β_α and β_D ranges narrow simultaneously could produce compounding effects not captured here. The analysis assumes uniform priors over varied ranges; alternative specifications such as beta distributions or empirical priors would yield different sensitivities. We do not perform full Bayesian propagation; complete uncertainty quantification would treat range boundaries as random variables with priors, yielding posterior distributions on $\mathrm{V}_{\mathrm{dynamic}}$ rather than point estimates.

Conclusion. The stability volume estimate is **structurally robust** (persists qualitatively across parameter uncertainties) but **quantitatively conditional** on the β_α range. All quantitative claims (94%, $N \in [4, 6]$) are conditional on Table 1 ranges, with β_α (coordination quality) as the dominant uncertainty.

SM-B: Proof of Lemma 2 (Hedging Necessity)

We prove that strategic hedging necessarily emerges using evolutionary game theory. The proof proceeds in five steps: setup, physical constraints, instability analysis, ESS characterization, and long-run dynamics.

B.1 Evolutionary Game Setup

Consider a large population of agents, each capable of adopting one of two strategies. Under the **Hedging (H)** strategy, agents allocate fraction $h \in (0, 1)$ of resources to defensive measures. Under the **Non-Hedging (NH)** strategy, agents allocate all resources to primary productive goals.

Let $p_t \in [0, 1]$ denote the proportion of hedgers at time t. Agents are randomly matched for interactions. The population state evolves according to replicator dynamics based on strategy payoffs.

Payoff functions:

$$\pi_H(p) = R(1-h)\gamma + Rh\rho(p) - C_H \tag{5}$$

$$\pi_{NH}(p) = R\gamma \cdot \sigma(p) - L(1-p) \tag{6}$$

Here $R > 0$ denotes total resources available, which are finite by thermodynamic constraints. The parameter $\gamma > 0$ represents productivity per unit resource invested in primary goals, while $h \in (0, 1)$ is the fraction of resources allocated to defense by hedgers. The function $\rho(p)$ captures security value (returns from defensive investment), with $\rho'(p) < 0$ indicating diminishing returns as more agents hedge. The constant $C_H > 0$ represents fixed infrastructure cost of maintaining defensive capabilities. The function $\sigma(p)$ gives the success probability for non-hedgers, with $\sigma'(p) > 0$ since success increases as more agents hedge (hedgers are defensive, not offensive). Finally, $L(1-p) > 0$ denotes expected loss from interference, which decreases in p because fewer non-hedgers means less vulnerability.

B.2 Physical Constraints

Three modeling assumptions, motivated by physical intuition, constrain these functions:

Constraint 1: Positive interference probability without defense.

$$\sigma(0) < 1$$

In an environment where no agents defend themselves, interference is possible. No agent can achieve perfect isolation from others sharing the same physical substrate (energy, communication channels, computational resources). This reflects fundamental non-excludability of shared resources.

Constraint 2: Positive losses from interference.

$$L(1) > 0$$

When interference occurs (probability $1 - \sigma$), it causes real costs: corrupted computations, disrupted resource flows, wasted energy. Even in a population where all agents are non-hedgers (maximum vulnerability), losses remain positive.

Constraint 3: Security value exceeds productivity at low hedging.

$$\rho(0) > \gamma$$

When few agents hedge ($p \approx 0$), the marginal return to defensive investment exceeds the return to productive investment. This reflects the "target-rich environment" effect: in an undefended population, protection is highly valuable. As more agents hedge, $\rho(p)$ decreases while γ remains constant.

Epistemic status: Constraints 1–2 (positive interference $\sigma(0) < 1$, positive losses $L(1) > 0$) reflect non-excludability of shared resources—a physical reality. Constraint 3 ($\rho(0) > \gamma$: security value exceeds productivity at low hedging) is a modeling choice motivated by "target-rich environment" dynamics: when few agents defend, marginal protection value is high. Alternative formulations with $\rho(0) < \gamma$ are mathematically coherent but represent offense-dominant regimes (where attacking is always more profitable than defending). Our assumption captures defense-dominant scenarios most relevant to AI coordination under infrastructure dependence, where cooperation is possible but requires deliberate design.

B.3 Instability of Non-Hedging Equilibrium ($p = 0$)

We now show that a population consisting entirely of non-hedgers is unstable to invasion by hedgers.

Payoff comparison at $p = 0$**:**

At $p = 0$ (all non-hedgers):

$$\pi_H(0) = R(1-h)\gamma + Rh\rho(0) - C_H \tag{7}$$

$$\pi_{NH}(0) = R\gamma \cdot \sigma(0) - L(1) \tag{8}$$

Taking the difference:

$$\pi_H(0) - \pi_{NH}(0) = R(1-h)\gamma + Rh\rho(0) - C_H - [R\gamma\sigma(0) - L(1)] \tag{9}$$

$$= R\gamma[(1-h) + h\frac{\rho(0)}{\gamma} - \sigma(0)] + L(1) - C_H \tag{10}$$

By Constraint 3, $\rho(0)/\gamma > 1$. By Constraint 2, $L(1) > 0$. By Constraint 1, $\sigma(0) < 1$. Therefore:

$$(1-h) + h\frac{\rho(0)}{\gamma} - \sigma(0) > (1-h) + h - 1 = 0$$

For sufficiently small infrastructure cost $C_H < L(1)$, we have:

$$\pi_H(0) - \pi_{NH}(0) > 0$$

Replicator dynamics: The population evolves according to:

$$\dot{p} = p(1-p)[\pi_H(p) - \pi_{NH}(p)]$$

At $p = 0$, since $\pi_H(0) > \pi_{NH}(0)$, any small introduction of hedgers ($p > 0$) causes $\dot{p} > 0$. The non-hedging equilibrium is unstable—hedgers will invade and spread.

B.4 Evolutionarily Stable Strategy (ESS) Analysis

An evolutionarily stable strategy p^* must satisfy two conditions. The **equilibrium condition** requires $\pi_H(p^*) = \pi_{NH}(p^*)$, ensuring the population is at rest with no selection pressure. The **stability condition** requires $\frac{d}{dp}[\pi_H(p) - \pi_{NH}(p)]\big|_{p=p^*} < 0$, ensuring that small perturbations decay and the equilibrium is locally stable.

Existence of interior equilibrium: Since $\pi_H(0) > \pi_{NH}(0)$ (proven above) and plausible limiting behavior has $\pi_H(1) < \pi_{NH}(1)$ (when all hedge, non-hedgers save costs C_H), by continuity there exists $p^* \in (0,1)$ where $\pi_H(p^*) = \pi_{NH}(p^*)$.

Stability condition: At equilibrium p^*:

$$\frac{d}{dp}[\pi_H(p) - \pi_{NH}(p)]\bigg|_{p=p^*} = Rh\rho'(p^*) - R\gamma\sigma'(p^*) + L'(1-p^*)$$

Since $\rho'(p) < 0$ (diminishing security value), $\sigma'(p) > 0$ (rising success for non-hedgers), and $L'(1-p) > 0$ (losses increase as non-hedgers increase), all three terms contribute negatively. Therefore the derivative is negative, confirming stability.

Conclusion: There exists a stable interior equilibrium $p^* \in (0,1)$ where a positive fraction of the population hedges. Pure non-hedging ($p = 0$) cannot be sustained.

B.5 Long-Run Dynamics Under Mutation-Selection

In realistic settings, strategies can mutate at rate $\mu > 0$. The stationary distribution under mutation-selection balance is given by:

$$\psi^*(p) \propto \exp\left(\frac{2N_{pop}}{\mu} V(p)\right)$$

where $V(p) = \int_0^p [\pi_H(q) - \pi_{NH}(q)]dq$ is the potential function and N_{pop} is effective population size.

As $\mu \to 0$, the distribution concentrates on local maximizers of $V(p)$. Since:

$$V'(p) = \pi_H(p) - \pi_{NH}(p)$$

and $V'(0) = \pi_H(0) - \pi_{NH}(0) > 0$ (proven in B.3), the point $p = 0$ is not a local maximizer. The stationary distribution concentrates on the interior equilibrium p^* or potentially $p = 1$ (all hedgers).

Biological interpretation: Even with random mutations occasionally producing non-hedgers, selection pressure maintains a population with substantial hedging. Non-hedging strategies cannot persist in the long run.

□

This proof establishes that any stable population equilibrium involves positive hedging under the three physical constraints. The result is robust to parameter variations and does not depend on specific functional forms beyond the stated monotonicity conditions.

SM-C: Proof of Theorem 2 (Stability Conditions)

We provide rigorous derivations of the three stability conditions.

C.1 Condition C1* (Correlated Equilibrium)

For the oversight mechanism to implement a correlated equilibrium (CE) with all-cooperation outcome, each agent receiving the recommendation to cooperate must prefer following it. The obedience constraint requires: for any deviation σ_i' by agent i, we have $\sum_{a_{-i}} \mu(a_{-i} \mid \sigma_i = O)[u_i(O, a_{-i}) - u_i(\sigma_i', a_{-i})] \geq 0$, where $\mu(a_{-i} \mid \sigma_i)$ is the probability of others' actions a_{-i} given agent i's signal σ_i, and u_i is agent i's payoff function. In our setting with mediator recommending cooperation $(O, O, \ldots, O)$ with probability 1, this reduces to requiring cooperation payoff exceeds defection payoff when all others cooperate. With coordination value β_α, conflict cost β_κ, and defection temptation β_D, the obedience constraint becomes $\beta_\alpha + \beta_\kappa \geq \beta_D$. This follows from Aumann's CE framework.

C.2 Condition C1** (Perfect Public Equilibrium)

For dynamic sustainability, we apply folk theorem results for repeated games with public monitoring [10, 12]. Using grim trigger strategies where defection triggers permanent reversion to punishment equilibrium, sustainability requires discounted future cooperation value exceeds immediate defection gain. This yields critical discount threshold $\delta^* = \max(0, \frac{\beta_D - \beta_\alpha}{\beta_D - \beta_\alpha + \beta_\kappa})$. When $\beta_\alpha \geq \beta_D$, we achieve patience-free regime with $\delta^* = 0$.

C.3 Condition C2* (Shapley Participation)

For voluntary participation, each agent's payoff share must exceed outside option. Under Shapley value allocation, agent i receives $\phi_i(N, v) = \frac{1}{N!} \sum_{S \subseteq N \setminus \{i\}} |S|!(N - |S| - 1)![v(S \cup \{i\}) - v(S)]$, where $v : 2^N \to \mathbb{R}$ is the characteristic function assigning value $v(S)$ to each coalition $S \subseteq N$. For additive oversight value where $v(S) = |S| \cdot \beta_\Omega$, this simplifies to per-agent share $\phi_i = \beta_\Omega$. Individual rationality requires $\phi_i \geq \beta_\ell / N$ where β_ℓ is removal gain. Equivalently, $\beta_\Omega \geq \beta_\ell / N$. This follows from cooperative game theory and Shapley's fairness axioms.

C.4 Alternative Equilibria

All-defect equilibrium always exists as Nash equilibrium. Mixed strategy equilibria may exist but yield lower welfare than coordinated cooperation. The mediator's role is equilibrium selection via correlation device, expanding the feasible payoff set beyond unmediated Nash equilibria [8, 9]. □

SM-D: Extended Sensitivity Analysis

This section contains comprehensive robustness analysis.

D.1 Multivariate Sensitivity

Stability margin $S = \min(\text{C1*}, \text{C1**}, \text{C2*})$ varies across full parameter space. Monte Carlo sampling over uniform priors yields stability volumes: $V_{C1^*} \approx 0.998$, $V_{C1^{**}} \approx 0.981$, $V_{C2^*} \approx 0.961$, implying overall dynamic stability $V_{\text{dynamic}} \approx 0.942$.

D.2 Sobol Indices

Global sensitivity analysis decomposes variance in stability margin: $\text{Var}(S) = \sum_i V_i + \sum_{i<j} V_{ij} + ...$ where V_i is first-order effect of parameter i, V_{ij} is interaction effect. Results show regime shift: at N=2-3, β_Ω dominates (first-order Sobol index ≈ 0.65) as participation binds tightly; at N≥5, δ dominates (index $\approx$ 0.58) as dynamic sustainability becomes primary constraint.

D.3 Margin Distributions

For each condition, margin distributions (5th/50th/95th percentiles) characterize robustness. C1* margins remain positive across 95% of parameter space; C1** margins depend on discount factor; C2* margins improve dramatically with N, becoming band-robust at N≥6.

D.4 Threshold Analysis

Critical thresholds identified: N=4 where V_{dynamic} crosses 95%; N≥6 where participation becomes band-robust ($\beta_\Omega \geq \beta_\ell / N$ throughout plausible range); $\beta_\alpha = \beta_D$ patience-free boundary.

D.5 Cost Model Variations

Optimal group size N^* varies with benefit assumptions: saturating benefits yield $N^* = 2$, linear benefits yield $N^* = 3$, network-effect benefits yield $N^* = 10$. This sensitivity demonstrates importance of benefit structure specification for design recommendations.

SM-E: Computational Incompleteness and Observability Asymmetry

This section formalizes the asymmetric observability claim ($\beta_D(\mathrm{M}, \mathrm{M}) > \beta_D(\mathrm{M}, \mathrm{H})$ for internal state monitoring) using computational complexity and incompleteness results analogous to Gödel's incompleteness theorems and Turing's halting problem.

E.1 Monitoring Incompleteness Theorem

Theorem 1 (Monitoring Incompleteness) *A computational system S_1 with resources (compute C_1, memory M_1, time budget T) cannot fully simulate another system S_2 of equal or greater complexity ($C_2 \geq C_1$, $M_2 \geq M_1$) in real-time (simulation time $\leq T$) without containing S_2 as a subsystem.*

Proof sketch. The argument proceeds in five steps.

Step 1 (Real-time simulation requirement): For monitoring to detect defection before significant harm, system S_1 must predict S_2's future state. Formally, for all inputs x and all $t \leq T$, we require $S_1(x, t) = S_2(x, t)$, constituting perfect simulation.

Step 2 (Computational overhead): System S_1 must allocate resources to four tasks: simulating S_2's computation (requiring C_2 compute and M_2 memory), observing S_2's inputs (requiring O_{in} bandwidth), comparing predicted versus observed states (requiring C_{comp} compute), and decision logic to trigger punishment if deviation is detected (requiring C_{dec} compute).

Step 3 (Resource constraint): The total resources required satisfy $C_1 \geq C_2 + O_{\text{in}} + C_{\text{comp}} + C_{\text{dec}}$.

Step 4 (Peer monitoring impossibility): If $C_1 = C_2$ (peer systems), then $C_2 \geq C_2 + O_{\text{in}} + C_{\text{comp}} + C_{\text{dec}}$, implying $0 \geq O_{\text{in}} + C_{\text{comp}} + C_{\text{dec}} > 0$. This contradiction establishes that peer monitoring ($C_1 \approx C_2$) cannot achieve real-time internal state verification.

Step 5 (Gödel/Turing analogy): This result parallels Gödel's second incompleteness theorem, where a formal system F cannot prove its own consistency without extending beyond F. Here, a computational system cannot verify another system of equal power without exceeding its own resources. The halting problem provides another parallel: no Turing machine can decide whether another Turing machine of equal description complexity will halt, given finite time.

Key implication: Real-time peer monitoring of *internal states* is computationally infeasible for systems of comparable complexity. □

E.2 Application to Multi-Agent Coordination

We distinguish three monitoring scenarios and derive defection detection probabilities β_D.

E.2.1 Machine Monitoring Machine (Internal States)

Setup: Machine M_1 attempts to monitor machine M_2's internal computation (hidden layer activations, goal representations, planning states).

Parameters: The systems are peers with $C_1 \approx C_2$ (similar capability) and $M_1 \approx M_2$ (comparable memory). The internal state space scales as $|S_2^{\text{int}}| \sim 2^{M_2}$, growing exponentially in memory size. The update rate reaches $R_2^{\text{int}} \sim 10^9$ state transitions per second at GHz processing speeds.

Monitoring requirement: To verify M_2 has not deviated internally, M_1 must check:

$$\text{Predicted state} \overset{?}{=} \text{Actual state}$$

at rate $R_2^{\text{int}} \sim 10^9$ checks/sec.

Incompleteness theorem applies: $C_1 \approx C_2$ implies M_1 cannot simulate M_2 in real-time. Therefore:

$$\beta_D(\mathrm{M}, \mathrm{M})_{\text{internal}} \to 1 \quad \text{(high defection probability; hard to detect)}$$

Interpretation: Machines *cannot* reliably verify each other's internal alignment in real-time when computational resources are comparable. Internal defection (reward hacking, hidden goal drift) can proceed undetected.

E.2.2 Machine Monitoring Machine (External Actions)

Setup: Machine M_1 monitors M_2's *external actions* (resource procurement, energy consumption, communication, physical-world effects).

Key distinction: External actions constitute a *projection* of internal state:

$$O_2^{\text{ext}} \subset S_2^{\text{int}} \quad \text{where} \quad |O_2^{\text{ext}}| \ll |S_2^{\text{int}}|$$

Observation space reduction: The internal state space is enormous, with $|S_2^{\text{int}}| \sim 2^{10^{12}}$ bits for terabyte memory. By contrast, external actions number only

$|O_2^{\text{ext}}| \sim 10^3$ discrete actions per second through physical actuators. The entropy comparison is stark: $H(O_2^{\text{ext}}) \sim 10$ bits per action is vastly smaller than $H(S_2^{\text{int}}) \sim 10^{12}$ bits for internal states.

Monitoring requirement: M_1 must verify:

$$W_{\text{obs}} \cdot T \geq H(O_2^{\text{ext}}) \cdot R_2^{\text{ext}} \cdot T$$

where W_{obs} is observation bandwidth, $R_2^{\text{ext}} \sim 10^3$ actions/sec.

For $R_2^{\text{ext}} \sim 10^3$, $H(O_2^{\text{ext}}) \sim 10$ bits/action:

$$W_{\text{obs}} \geq 10^4 \text{ bits/sec}$$

Feasibility: Modern sensors achieve $W_{\text{obs}} \sim 10^6$ to 10^9 bits/sec (camera, network monitoring). Therefore:

$$W_{\text{obs}} \gg H(O_2^{\text{ext}}) \cdot R_2^{\text{ext}}$$

Conclusion:

$$\beta_D(\text{M}, \text{M})_{\text{external}} \to 0 \quad \text{(low defection probability; easy to detect)}$$

Interpretation: Machines *can* monitor each other's external actions because the action space is vastly smaller than internal state space. Defection requiring sustained external resource acquisition (compute procurement, energy theft, infrastructure control) leaves observable traces.

E.2.3 Machine Monitoring Human (External Actions)

Setup: Machine monitors human's actions (physical movement, communication, resource usage).

Parameters: The human action rate is approximately $R_H^{\text{ext}} \sim 1$ significant action per second, limited by motor control. Action entropy is roughly $H(O_H^{\text{ext}}) \sim 10$ bits per action, requiring observation bandwidth of only $W_{\text{obs}} \geq 10$ bits/sec.

Feasibility: Trivially achievable with modern sensors ($W_{\text{obs}} \sim 10^6$ bits/sec).

Conclusion:

$$\beta_D(\text{M}, \text{H})_{\text{external}} \approx 0 \quad \text{(very low; almost impossible to hide)}$$

Interpretation: Humans cannot hide actions from machine monitors due to slow action rate and high sensor bandwidth.

E.3 Asymmetric Observability and Weak Leviathan

Summary of detection probabilities (lower β_D = easier to detect):

Monitoring Scenario	β_D (defection undetected)	Feasibility
M → M (internal states)	High ($\to 1$)	Infeasible (incompleteness)
M → M (external actions)	Low ($\to 0$)	Feasible (small action space)
M → H (external actions)	Very Low (≈ 0)	Trivially feasible
H → M (any)	High ($\to 1$)	Infeasible (speed mismatch)

Critical assumption for Weak Leviathan mechanism:

Game-theoretic stability conditions (C1*, C1**, C2*) require verifying *future cooperation intent*, not just current action compliance. Detecting defection *after* significant harm has occurred is insufficient—the mediator must verify agents will *continue cooperating* in future rounds.

Internal alignment verification requirement: Two distinct verification tasks must be distinguished. **External compliance** asks whether the agent followed protocol this round, which is observable via external actions. **Internal alignment** asks whether the agent intends to cooperate in future rounds, which requires internal state verification that Theorem 1 shows is computationally infeasible for peer systems.

Repeated game stability (C1**, Folk Theorem conditions) depends on beliefs about *future* behavior, which requires inferring the discount factor δ (patience and future valuation), private valuations (hidden estimates of β_D and β_α), and goal stability (whether the current cooperation goal will persist).

These are *internal states*, not external actions. Therefore, monitoring incompleteness theorem applies:

$$\beta_D(\mathrm{M},\mathrm{M})_{\text{alignment}} \to 1 \quad \text{(machines cannot verify peers' future intentions)}$$

Addressing the Intention-Action Distinction:

A critical objection arises: if game-theoretic stability requires verifying future cooperation *intent* (not just current action compliance), then both machines and humans face the same impossibility—intentions are internal states, unobservable in principle. Why are humans better mediators if they also cannot verify machine intentions?

Resolution via Perfect Public Equilibrium: The mediator's role is NOT to verify individual agent intentions, but to create *common knowledge* of compliance behavior. In repeated games with public monitoring [11], cooperation can be sustained via perfect public equilibrium (PPE) even without verifying private beliefs. Three ingredients are essential: the mediator broadcasts coordination recommendations to all agents simultaneously (public signals), each agent observes whether others follow the mediator's signal through external actions (observable compliance), and deviation by any agent triggers coordinated punishment by others (trigger strategies).

The mediator need NOT verify intentions—only provide consistent public signals making deviations observable. The asymmetry operates as follows: $\beta_D(\mathrm{M},\mathrm{M})_{\text{external}} \approx 0$ because machines easily observe each other's *compliance* with mediator signals, while $\beta_D(\mathrm{M},\mathrm{M})_{\text{internal}} \to 1$ because machines cannot verify peer internal computations that might enable undetected future defection preparation.

Humans qualify as mediators because their outputs (signals) are observable (low $\beta_D(\mathrm{M}, \mathrm{H})_{\text{external}}$), and their computational limitations prevent exploitation. The system relies on consistent public signals, not intention verification.

Remaining limitation: This resolution assumes public monitoring suffices for stability. If coordination requires verifying that agents are NOT secretly preparing future deviations (e.g., building hidden capabilities, undetectable goal drift), then external observability is insufficient, and the weak Leviathan mechanism fails. Our framework applies to coordination under *public monitoring with trigger strategies*, not scenarios requiring full internal transparency. This assumption is standard in repeated game theory [14] but represents a boundary condition for our results.

Weak Leviathan justified:

Human institutions can serve as external mediators for four reasons. First, **computational incompleteness** prevents machines from verifying each other's internal alignment (future cooperation intent) in real-time when $C_1 \approx C_2$. Second, **external observability suffices for mediation**: the mediator need only provide a correlation device (focal point) and verify compliance with recommendations, not internal motivation. Third, **asymmetric opacity favors external arbiters**: machines verifying the mediator face $\beta_D(\mathrm{M}, \mathrm{H}) \approx 0$ (human defection is easy to detect), the mediator verifying machines needs only to monitor external compliance rather than internal states, and machines verifying each other's compliance face $\beta_D(\mathrm{M}, \mathrm{M})_{\text{external}} \to 0$ (feasible). Fourth, **weak capacity provides credibility**: humans cannot compute fast enough to exploit coordination (limited C_H), making them credible neutral arbiters. The mediator's *weakness* (computational limitations) prevents strategic manipulation.

Falsifiability condition:

The Weak Leviathan mechanism fails if future AI systems develop any of three capabilities: **homomorphic verification** that verifies peer internal states without full simulation (requiring breakthroughs in zero-knowledge proofs of computation), **cryptographic commitment to goals** providing binding commitments to future cooperation that are verifiable without simulation, or a **reduced complexity gap** where monitoring systems have $C_1 \gg C_2$ through specialized monitoring infrastructure, causing the incompleteness theorem to no longer apply.

If any of these emerge, $\beta_D(\mathrm{M}, \mathrm{M})_{\text{alignment}} \to 0$, and peer monitoring becomes feasible, eliminating need for external mediators.

E.4 Architectural Implications

Design principle: Coordination mechanisms should structure defection such that significant gains require *sustained external actions*, not just hidden internal state changes.

Implementation requirements. Four architectural elements support this design principle. **Observable action bottlenecks** ensure that defection must pass through observable channels such as resource procurement, infrastructure access, and communication. **Time-extended verification** means monitoring occurs over a time window T, so defection requiring time $T_{\text{def}} > T$ for significant gains allows multiple verification opportunities. **Cryptographic logging** generates tamper-evident records of external actions (via blockchain or secure logs) that persist beyond the immediate detection window. **Commitment devices** allow agents to commit to future actions via cryptographic protocols, making deviation from commitment externally verifiable even when internal state remains opaque.

Impossibility result. If coordination stability *requires* verifying internal alignment (goals, motivations, hidden valuations) and monitoring incompleteness holds ($C_1 \approx C_2$), then three consequences follow: peer-to-peer monitoring is insufficient, necessitating an external mediator; the external mediator must have $\beta_D(\text{agents} \rightarrow \text{mediator}) \approx 0$ (high observability); and the mediator must have limited exploitation capacity (weak enough to be credible). This formalizes the "governance through weakness" mechanism: the mediator's computational limitations (relative to monitored agents) are a *feature*, not a bug, enabling credible neutral arbitration.

This BibTeX file contains references for the multi-agent superintelligence paper.

References

[1] Bostrom, N. *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, Oxford, UK, 2014).

[2] Dafoe, A. Ai governance: A research agenda (2018). URL https://www.governance.ai/paper/ai-governance-a-research-agenda. Published by the Future of Humanity Institute, University of Oxford.

[3] Hammond, L. *et al.* Multi-agent risks from advanced ai. Tech. Rep. Technical Report #1, Cooperative AI Foundation (2025). URL https://www.cooperativeai.com/post/new-report-multi-agent-risks-from-advanced-ai. Taxonomy of multi-agent risks: miscoordination, conflict, collusion.

[4] Tomasev, N. *et al.* Distributional agi safety. *arXiv preprint arXiv:2512.16856* (2025). Proposes patchwork AGI hypothesis where AGI emerges through coordination of sub-AGI agents.

[5] Stastny, J. *et al.* Normative disagreement as a challenge for cooperative ai. *NeurIPS 2021 Workshop on Cooperative AI* (2021). Cooperation in mixed-motive environments with competing cooperative equilibria.

[6] Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423 (1948).

[7] Landauer, R. Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development* **5**, 183–191 (1961).

[8] Aumann, R. J. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* **1**, 67–96 (1974).

[9] Aumann, R. J. Correlated equilibrium as an expression of bayesian rationality. *Econometrica* **55**, 1–18 (1987).

[10] Abreu, D., Pearce, D. & Stacchetti, E. Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica* **58**, 1041–1063 (1990).

[11] Fudenberg, D., Levine, D. K. & Maskin, E. The folk theorem with imperfect public information. *Econometrica* **62**, 997–1039 (1994).

[12] Fudenberg, D. & Maskin, E. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* **54**, 533–554 (1986).

[13] Green, E. J. & Porter, R. H. Noncooperative collusion under imperfect price information. *Econometrica* **52**, 87–100 (1984).

[14] Mailath, G. J. & Samuelson, L. *Repeated Games and Reputations: Long-Run Relationships* (Oxford University Press, Oxford, UK, 2006).

[15] Shapley, L. S. in *A value for n-person games* (eds Kuhn, H. W. & Tucker, A. W.) *Contributions to the Theory of Games II* 307–317 (Princeton University Press, Princeton, NJ, 1953).

[16] Bennett, C. H. The thermodynamics of computation—a review. *International Journal of Theoretical Physics* **21**, 905–940 (1982).

[17] Lloyd, S. Ultimate physical limits to computation. *Nature* **406**, 1047–1054 (2000).

[18] Kamenica, E. & Gentzkow, M. Bayesian persuasion. *American Economic Review* **101**, 2590–2615 (2011).

[19] Bergemann, D. & Morris, S. Bayes correlated equilibrium and the combinatorics of information. *Econometrica* **84**, 845–876 (2016).

[20] Hofbauer, J. & Sigmund, K. *Evolutionary Games and Population Dynamics* (Cambridge University Press, Cambridge, UK, 1998).

[21] Weibull, J. W. *Evolutionary Game Theory* (MIT Press, Cambridge, MA, 1995).

[22] Nowak, M. A. *Evolutionary Dynamics: Exploring the Equations of Life* (Harvard University Press, Cambridge, MA, 2006).

[23] Sandholm, W. H. *Population Games and Evolutionary Dynamics* (MIT Press, Cambridge, MA, 2010).

[24] Greenblatt, R. *et al.* Ai control: Improving safety despite intentional subversion. *arXiv preprint arXiv:2312.06942* (2023). Trusted monitoring framework with game-theoretic red-teaming.

[25] The oversight game: Learning to cooperatively balance an ai agent's safety and autonomy. *arXiv preprint arXiv:2510.26752* (2024). Game-theoretic framework for post-hoc AI control.

[26] Oxford Martin AI Governance Initiative. Verification for international ai governance. Tech. Rep., University of Oxford (2025). URL https://aigi.ox.ac.uk/publications/verification-for-international-ai-governance/. Six layers of verification for large-scale AI development rules.

[27] Schelling, T. C. *The Strategy of Conflict* (Harvard University Press, Cambridge, MA, 1960).

[28] Irving, G., Christiano, P. & Amodei, D. Ai safety via debate. *arXiv preprint arXiv:1805.00899* (2018). Foundational debate framework for scalable oversight.

[29] Brown-Cohen, J. *et al.* Scalable ai safety via doubly-efficient debate. *arXiv preprint arXiv:2311.14125* (2023). Advances debate-based scalable oversight.

[30] Christiano, P., Shlegeris, B. & Amodei, D. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575* (2018). Iterated amplification framework.

[31] Coordination transparency: Governing distributed agency in ai systems. *AI & Society* (2026). Governance mechanism for distributed agency: interaction logging, live monitoring, intervention hooks.

[32] Bernheim, B. D., Peleg, B. & Whinston, M. D. Coalition-proof nash equilibria i: Concepts. *Journal of Economic Theory* **42**, 1–12 (1987).

[33] Verification methods for international ai agreements. *arXiv preprint arXiv:2408.16074* (2024). Hardware monitoring, infrared detection, tamper-evident seals.

[34] Gödel, K. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme i. *Monatshefte für Mathematik und Physik* **38**, 173–198 (1931). English translation: “On Formally Undecidable Propositions of Principia Mathematica and Related Systems I”.

[35] Turing, A. M. On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society* **s2-42**, 230–265 (1936). Introduces the Turing machine and proves the undecidability of the halting problem.