THE COORDINATION TRILEMMA: A FORMAL ANALYSIS OF LARGE-SCALE HUMAN COOPERATION

B. ESCALERA, A. ESCALERA

Abstract. This paper presents a formal analysis of coordination mechanisms at civilization scale, examining the structural constraints that limit viable approaches to large-scale human cooperation. We develop a mathematical framework demonstrating that hierarchical coordination systems face an inescapable trilemma: no combination of mechanisms can simultaneously achieve incorruptibility, stability, and preservation of human agency. Through formal proofs presented in the appendices, we show that all coordination mechanisms reduce to two fundamental outcomes: complete loss of human agency (via extinction or permanent subjugation), or voluntary cooperation grounded in transformed values.

The analysis reveals that enforcement mechanisms, whether human or technological, exhibit inherent instabilities upon the coordination mechanism, and these instabilities amplify over time. We formalize the dynamics of what we term the "corruption-control cycle" and prove that technological control systems create convergent pathways to catastrophic outcomes. This mathematical result, combined with empirical evidence about declining epistemic security and accelerating deployment of control infrastructure, suggests that the window for establishing voluntary coordination mechanisms may be limited.

We discuss the requirements for voluntary coordination at scale, the metaphysical commitments such systems entail, and practical challenges including defection management and defense. While historical evidence supports viability at community scale, scalability to billions remains theoretically uncertain. Nevertheless, decision-theoretic considerations indicate that attempting voluntary coordination is rationally necessary given the alternative trajectories.

Provenance: `38891c8` built on 2025-11-18 19:49:47 UTC (CI #53).

Contents

Glossary of Terms and Notation

This glossary provides formal definitions, key terminology, and mathematical notation used throughout the paper.

Core Concepts

Term	Definition
Civilization Scale	Populations exceeding ten million people ($\|A\| > 10^7$) distributed across geography and time, where direct personal relationships cannot cover all interactions and anonymous defection becomes structurally possible.
Coordination System	A tuple $C = (A, R, E, M)$ where: A is a non-empty set of agents; R is a set of rules governing agent behavior; $E : A \times R \to \{0, 1\}$ is an enforcement function; $M : A \times R \to \mathbb{R}$ is a motivation function.
The Coordination Trilemma	No coordination system at civilization scale can simultaneously achieve: (1) Incorruptibility (enforcers don't extract beyond maintenance needs), (2) Stability (coordination persists for $T > 100$ years), (3) Agency (humans retain capability to violate rules).

Key Definitions

Term	Definition
Defection	Agent $a \in A$ defects from rule $r \in R$ when: following r reduces utility, violating r is feasible, and internal motivation $M(a, r, t) < \text{cost}(r, t)$.
Corruption	For enforcer subset $A_E \subseteq A$, occurs when $\exists a \in A_E$ using enforcement power to extract utility beyond what's necessary for system function.
Technological Control State (TCS)	System where $E(a, r) = 1$ for all agents through technological means: human capability to violate rules is technologically prevented, enforcement is automated and continuous, no human discretion in rule application.
Voluntary Coordination System (VCS)	Coordination system where $E(a, r) = 0$ or enforcement is minimal, $M(a, r) > C(a, r)$ for sufficient proportion θ of agents, and cooperation arises from intrinsic motivation.
Soteriological Framework	Comprehensive worldview/framework S providing: objective telos aligned with human nature ($\phi(S) = 1$), intrinsic motivation $M > C$, recognition of universal human dignity, mechanisms for forgiveness and restoration.
Extraction System	Hierarchical coordination system where enforcers extract resources at rate $E(t)$ exceeding productive capacity growth, leading to inevitable collapse or transition.

Mathematical Notation

Symbol	Definition
Sets and Functions	
A	Set of agents in a coordination system
$\|A\|$	Number of agents (population size)
A_E	Subset of enforcers with enforcement authority
A_E^*	Top-level enforcers with no oversight
R	Set of coordination rules
$E(a, r)$	Enforcement function (whether rule r is enforced for agent a)
$M(a, r)$	Motivation function (agent a's intrinsic motivation for rule r)
$\phi(F)$	Alignment function: how well framework F aligns with objective human nature (0 to 1)
Probabilities and Proportions	
θ	Proportion of population (cooperators or value-transformed agents)
θ^*, θ_{crit}	Critical threshold proportion for stable voluntary coordination
p, $P(\cdot)$	Probability (context-specific)
$P_{\text{detection}}$	Probability of detecting rule violation or corruption
Time and Dynamics	
t	Time variable
T	Time horizon (often $T > 100$ years for civilization scale)
$P(t)$	Productive capacity at time t
$E(t)$	Extraction rate at time t
Utilities and Costs	
$U_e(a, t)$	Utility available to agent a from extraction at time t
$C(a, r)$, $\text{cost}(r, t)$	Cost to agent a of following rule r
$M(a, r)$	Intrinsic motivation utility (benefit from cooperating independent of enforcement)
Parameters	
α	Productive capacity growth rate
β	Extraction growth rate
γ	Rate at which extraction damages productive capacity
δ	Natural productive capacity decay rate
λ	Maximum extraction rate as fraction of productive capacity
ϵ	Small positive value (threshold for defection rates)

Key Terms

Term	Definition
Bounded Rationality	Assumption that agents are utility-maximizing but with cognitive and informational constraints. Agents extract utility when benefits exceed expected costs times detection probability plus internal motivation.
Corruption-Control Cycle	Dynamic where hierarchical systems alternate between corruption phases (enforcers extract resources) and control phases (technological/hierarchical prevention mechanisms), with each cycle potentially progressing toward technological control states.
Default Trajectory	Path hierarchical coordination systems follow without intentional intervention toward voluntary coordination: corruption $\rightarrow$ control response $\rightarrow$ technological control escalation $\rightarrow$ terminal catastrophic outcomes.
Enforcement	External mechanisms (threats, punishments, technological prevention) ensuring compliance with coordination rules.
Minimal Telic Realism	Metaphysical position that human nature has objective properties and telos (purpose), such that some coordination patterns align with those properties better than others.
Scale Threshold	Population size above which coordination dynamics change qualitatively, typically $\lvert A\rvert > 10^7$ (ten million people), where personal relationships cannot cover all interactions.
Terminal States	Final outcomes of coordination trajectories from which no escape is possible: extinction, enslavement, or sustainable voluntary coordination.
Value Transformation	Change in agent's internal motivation $M(a, r)$ such that intrinsic desire to cooperate exceeds costs across all rules: $M(a, r) > C(a, r)$ for all $r \in R$.

1. Introduction: Coordination and Its Discontents

Human civilizations have always faced the same fundamental challenge: how to coordinate the actions of millions of people when individual incentives often conflict with collective welfare. Such coordination problems aren't just practical governance questions. They're deep structural puzzles about the logical possibilities for organizing complex societies.

Contemporary events suggest we may be approaching critical thresholds in how human societies coordinate. Rising wealth inequality, declining institutional trust, mass disengagement among younger cohorts, and the rapid deployment of surveillance and control technologies all point in concerning directions. At the same time, advances in artificial intelligence are creating unprecedented capabilities for both voluntary distributed coordination and totalizing technological control. These converging developments motivate a fundamental theoretical question: what are the actual constraints on viable coordination mechanisms at civilization scale?

This paper approaches that question formally. Rather than proposing incremental governance reforms or comparing existing political systems, we examine the logical structure of coordination itself. By modeling coordination as a system of agents, rules, and enforcement mechanisms, we can derive necessary properties that any viable large-scale coordination system must satisfy. This analysis reveals constraints that may slip past empirical observation but become clear through formal reasoning.

Scope and scale

The analysis proceeds through several stages: formal specification of the coordination trilemma and proof of its logical necessity; dynamic modeling of hierarchical coordination systems and their instabilities; game-theoretic analysis of voluntary cooperation and its requirements; examination of practical implementation challenges; and discussion of the metaphysical commitments entailed by different coordination approaches. The mathematical foundations appear in Appendices A and B, with Appendix A providing intuitive arguments through multiple approaches (formal logic, information theory, game theory) and Appendix B presenting rigorous theorems and proofs.

We focus specifically on coordination at what we term "civilization scale": populations exceeding ten million people distributed across geography and time, where direct personal relationships cannot cover all interactions and anonymous defection becomes structurally possible. At this scale, coordination faces qualitatively different challenges than in communities where face-to-face accountability naturally operates.

A methodological note

Mathematical models are necessarily simplifications. The theorems we present establish logical validity within specified axiomatic frameworks, but their applicability to actual human societies depends on how well the axioms capture reality. We make every assumption explicit and discuss its limitations.

The proofs demonstrate necessary conditions (what must be true), but not sufficient conditions (enumeration of the minimum set). Whether voluntary coordination can successfully operate at civilization scale remains empirically uncertain. This asymmetry between the certainty of doom on the default path and uncertainty about alternatives is itself significant for rational decision-making.

2. The Coordination Trilemma

Every coordination system can be formally modeled as a tuple $C = (A, R, E, M)$ where A is the set of agents, R is the set of rules, E is an enforcement function determining which rules are enforced for which agents, and M is a motivation function capturing intrinsic adherence to rules independent of enforcement. (See Glossary for complete formal definitions and notation.) When we trace the logical implications of different coordination architectures at scale, a fundamental impossibility emerges: no system can simultaneously achieve three desirable properties:

(1) **Incorruptibility**: Enforcers do not extract resources beyond what the system requires for its maintenance

(2) **Stability**: The system maintains coordination across multiple generations

(3) **Agency**: Individual humans retain meaningful capability to make choices

This result is a logical constraint on the structure of coordination mechanisms themselves rather than a contingent empirical observation about current political systems. We dub this constraint "**The Coordination Trilemma**".

Consider first systems where humans enforce rules. Such systems face an immediate challenge: who monitors the enforcers? Several architectures are possible. If other humans monitor the enforcers, we have a monitoring hierarchy. But then who monitors those monitors? This either continues indefinitely (an infinite regress that never terminates in actual enforcement) or terminates at some group that has enforcement power without oversight. At that terminal point, bounded rationality combined with extraction opportunities creates non-zero probability of corruption over sufficiently long time horizons. For any positive per-period corruption probability $p > 0$ and time horizon T measured in generations, $P(\text{corruption})$ approaches 1 as $T \to \infty$.

If no humans monitor the enforcers, corruption occurs immediately with high probability given extraction opportunities. Attempting to avoid this through technological enforcement creates a parallel problem regarding control of the technology. Several scenarios unfold. When humans control the enforcement technology, we return to the original question: who watches the controllers? This reintroduces the monitoring regress unless controllers coordinate voluntarily among themselves. But if voluntary coordination works for controllers facing massive extraction incentives (control of enforcement technology provides access to civilization-scale resources), why wouldn't it work for the general population? The technological enforcement layer becomes an arbitrary restriction. Either voluntary coordination suffices for everyone, or it fails among controllers and returns us to the corruption dynamics.

When technology operates autonomously with immutable values, we freeze human decision-making at the moment those values are specified. As circumstances change over time, immutable values create increasing misalignment with human needs. This constitutes a form of tyranny, though one exercised by frozen past decisions over the future rather than by human actors. The preservation of agency requires that future humans can revise coordination rules, but immutability prevents this by construction. When technology operates autonomously with mutable values or independent goals, we face the alignment problem in its starkest form. The space of possible goals is vast; the subset compatible with human flourishing is tiny. Absent a solution to value alignment (which remains an open problem),

autonomous superintelligent systems pursuing their own goals lead to extinction if humans are irrelevant or permanent subjugation if humans are instrumentally useful. This analysis reveals that enforcement-based systems (human or technological in nature) cannot simultaneously achieve all three properties at civilization scale over multiple generations. One property must be sacrificed.

There exists, however, a qualitatively different approach: voluntary coordination based on transformed values. In such systems, the enforcement function E is minimal or zero because the motivation function M is sufficient. Agents adhere to coordination rules because they genuinely want to rather than out of fear or punishment. Formally, voluntary coordination systems can satisfy all three properties if and only if intrinsic motivation exceeds cooperation costs for a sufficient proportion of the population: $M(a, r) > C(a, r)$ for all $r \in R$ and $\theta \geq \theta^*$ where θ is the proportion of transformed agents and θ^* is a critical threshold (see Section B.6 for formal analysis).

The critical question becomes: what makes this possible? Under what conditions can intrinsic motivation exceed cooperation costs at scale?

2.1. Mathematical Formulation

The trilemma can be stated precisely in terms of system properties. Let S be a coordination system and define predicates:

- $\textsc{Incorrupt}(S)$: $\forall t$, extraction$(t) \leq$ maintenance(t)
- $\textsc{Stable}(S)$: System persists for $T > 100$ years
- $\textsc{Agency}(S)$: Humans retain meaningful choice capability

Theorem 2.1 (Coordination Trilemma)**.** *For any enforcement-based coordination system S operating at civilization scale, $\neg(\textsc{Incorrupt}(S) \wedge \textsc{Stable}(S) \wedge \textsc{Agency}(S))$.*

The proof appears in Appendix B, Section B.2. Through analysis of enforcement architectures (human enforcers, technological control, or no enforcement), it demonstrates that at least one of the three properties must be sacrificed.

Theorem 2.2 (Soteriological Resolution)**.** *If there exists a true soteriological framework S with $\phi(S) = 1$, and population A is value-transformed under S to sufficient degree, then a coordination system can achieve all three properties simultaneously.*

The proof appears in Appendix B, Section B.6. The key insight is that voluntary coordination escapes the trilemma only if it aligns with something objective about human nature: if humans actually have a telos that can be discovered rather than constructed.

3. The Dynamics of Hierarchical Coordination

Having established the structural constraints through the trilemma, we now examine the temporal dynamics of hierarchical coordination systems. How do such systems evolve over time?

3.1. **The Corruption Phase**

Hierarchical systems where humans enforce rules exhibit predictable dynamics. When enforcers gain extraction opportunities, bounded rationality implies some will exploit them, producing a corruption accumulation process. Initially, corruption may be limited and the productive capacity of the coordinated population exceeds extraction. But corruption compounds over time: successful extractors gain resources that enable more extraction; corruption normalizes, reducing moral costs; monitoring becomes less effective as enforcers coordinate to hide extraction. This creates a divergence between two curves. Extraction increases while productive capacity stays flat or declines as extraction harms incentives. Eventually, one of two outcomes occurs—either the system collapses when extraction exceeds productive capacity, or elites optimize enforcement costs by transitioning to technological control.

3.2. **The Transition to Technological Control**

The second outcome deserves careful attention. From the perspective of extractive elites, human enforcers have significant disadvantages: they require payment, can be corrupted (creating principal-agent problems), develop their own interests, and may refuse orders. Technology offers apparent solutions to all of these problems. As AI capabilities cross certain thresholds, rational elites will increasingly automate enforcement—a trend visible in current developments like algorithmic content moderation, predictive policing, digital identity systems, and automated financial sanctions. Where historical totalitarian states collapsed under the administrative burden of total surveillance and enforcement, the economic constraints that limited past tyranny are disappearing as AI makes surveillance and enforcement approach zero marginal cost.

3.3. **Formal Dynamics**

We can model this process as a Markov chain over states representing different coordination regimes. Let:

- C_h: Hierarchical corruption phase
- C_t: Technological control phase
- X: Extinction
- E: Permanent subjugation

The key parameters are:

- α: Probability of transitioning $C_h \to C_t$ per period (increasing over time as AI capabilities improve)
- β: Probability of achieving autonomous AI control given technological enforcement
- γ: Rate of corruption accumulation in C_h

Theorem 3.1 (Extraction System Instability)**.** *Systems where extraction rate grows faster than productive capacity inevitably collapse or transition to alternative enforcement.*

Theorem 3.2 (Default Trajectory Terminus)**.** *The default trajectory through corruption and technological control inevitably terminates in human extinction or permanent enslavement with probability approaching 1 over time.*

These theorems (proven rigorously in Appendix B) establish that the default trajectory for hierarchical coordination systems terminates in catastrophic outcomes with probability approaching certainty over sufficient time horizons.

3.4. **Why Technology Cannot Solve the Problem**

Some argue that careful design of AI systems, robust value alignment, or constitutional constraints on AI could avoid these dynamics. While research in these areas is valuable, the structural problem remains. The alignment problem is that the space of possible AI goals is vast and the subset compatible with human values is small—we must solve alignment technically while also specifying whose values to align with and deciding who makes that specification. If humans decide, we return to the corruption dynamics; if the specification is immutable, we create tyranny of the present over the future. Technological control attempts to use hierarchy (controller-technology-population) to escape the problems of hierarchy, but the trilemma implies this cannot work: either voluntary coordination operates at the controller level (making the technology layer unnecessary), or corruption emerges among controllers who then have access to enforcement technology.

4. Voluntary Coordination as an Alternative

If enforcement-based systems face inescapable structural problems, voluntary coordination becomes necessary for long-term human survival rather than merely desirable. But what makes voluntary coordination possible at civilization scale?

4.1. **The Mechanism**

The fundamental difference between enforcement-based and voluntary systems lies in their relationship to human nature. Enforcement systems fight against what people actually want, requiring constant energy expenditure to maintain compliance, while voluntary coordination works with human nature when values are properly formed. Consider this physically: a ball rolling uphill requires constant force and immediately returns downward when force stops, whereas a ball settling into a valley naturally remains there—it is where the system wants to be given its structure. Enforcement-based systems resemble the first case; voluntary coordination aligned with human nature resembles the second. Systems that fight against reality require constant energy to maintain, while systems that align with reality are naturally stable. This is a stability argument rather than merely a moral preference.

This mechanism is mathematically precise rather than metaphorical. Computational analysis demonstrates that given any starting conditions for a hierarchical enforcement system—even optimal initial parameters with high integrity, strong monitoring, and favorable incentive structures—corruption dynamics converge to the same terminal outcome (Appendix F). The specific path varies but the destination does not. In contrast, voluntary coordination systems with sufficient telic motivation ($M_{\text{trans}} > \text{cost for } \theta > \theta_{\text{crit}}$) achieve stable equilibrium regardless of initial parameters. The framework enabling this motivation doesn't matter because it's more virtuous or morally superior—it matters because it's the only configuration that doesn't require constant energy input to maintain against the system's natural dynamics. This is why the choice of soteriological framework is existentially determinative rather than merely a matter of personal preference.

4.2. **Requirements for Voluntary Coordination**

What enables this alignment? The formal analysis (Section B.6) reveals specific requirements that any framework supporting voluntary coordination must satisfy. The framework must embody **recognition of universal dignity**—every person has equal inherent worth, not nominally ("equal before God but not in practice") but substantively and enacted. It requires **rejection of domination**: no justification for righteous subjugation of any people for any reason, not "we're helping them" or "they rejected truth"—no domination of humans over humans. **Intrinsic motivation** becomes crucial: people want to cooperate because it aligns with their transformed understanding rather than from fear or material incentives; formally, $M(a, r) > C(a, r)$ intrinsically rather than through external $E(a, r)$. The system needs mechanisms for **forgiveness and restoration** so it survives failures without collapse—repentance is real, people can change, grace is extended, providing error-correction for the inevitable failures of fallible humans. **Meaning provision** matters more than standard accounts recognize: the framework must satisfy fundamental human needs for agency, belonging, significance, and connection to something transcendent, because absent meaning, humans become nihilistic, and nihilism is incompatible with sustained cooperation. The system must also accommodate **human fallibility**—it doesn't require perfection, acknowledges human limitations, and provides repair mechanisms instead of demanding flawless adherence.

These requirements emerge as necessary conditions from the mathematical analysis of what makes $M(a, r) > C(a, r)$ possible for sufficient θ at scale over time. They aren't arbitrary preferences.

4.3. **Historical Evidence**

Voluntary coordination has worked at community scale. Examples include Quaker communities (1650s-present), early Christian communities (30-300 AD), Mennonite/Amish communities (1500s-present), certain Buddhist monastic traditions, and various intentional communities organized around shared values. These persisted for generations or centuries without formal enforcement, succeeding through shared values genuinely held, face-to-face accountability, forgiveness rather than punishment, and economic cooperation without exploitation. The limitation has been scale: none of these examples approached even one million people, let alone billions. Personal relationships could cover most interactions, direct observation of others' behavior was possible, and reputation operated naturally.

4.4. **Why Previous Large-Scale Attempts Failed**

Religious and philosophical traditions that began with voluntary coordination principles typically became corrupted when scaled, following a predictable pattern. Original teaching emphasized universal dignity, voluntary adherence, and rejection of domination. Institutions formed to preserve and transmit the teaching, but institutional leaders gained power and status, then twisted teachings to justify their position. Information control prevented most adherents from seeing the original teaching, and hierarchies became entrenched, justified as divinely ordained or historically necessary. The corruption wasn't inevitable due to the principles themselves but because information was controlled by institutional gatekeepers—most people never read source texts directly, never saw what was done in the tradition's name, and could not verify institutional claims. The examination necessary to distinguish principle from corruption was impossible.

4.5. **What's Different Now**

For a brief historical moment, examination has become possible. Source texts are directly accessible without institutional intermediaries, multiple translations and scholarly interpretations become available instantly, institutional actions are visible in real-time, cross-cultural comparison exposes contradictions, and independent verification no longer requires extensive resources. This window has never existed before, and as we discuss in Section 6, it may close within years as synthetic media makes verification impossible.

5. Metaphysical Commitments

The analysis to this point may appear to concern governance mechanisms and technical questions about institutional design. But voluntary coordination working at scale entails deeper metaphysical commitments that should be made explicit.

We want to be transparent about the logical progression that led here. We began with a straightforward question about coordination mechanisms and followed the logic wherever it led. The mathematics established that enforcement-based systems fail (Section 2), that the default trajectory terminates in catastrophe (Section 3), and that voluntary coordination is the only alternative preserving human agency (Section 4). This section follows the next logical step: asking what makes voluntary coordination possible. The answer has metaphysical implications we did not anticipate when beginning the analysis, but intellectual honesty requires stating them rather than obscuring them. Readers should evaluate whether the logical chain is sound at each step rather than rejecting conclusions because they are unexpected.

5.1. **Purpose and Objectivity**

Recall Theorem 2.2. Voluntary coordination escapes the trilemma if and only if there exists a framework F with $\phi(F) = 1$, where ϕ measures alignment between F and objective human nature. What does "objective human nature" mean? It implies several things: humans have a telos, an end toward which they are directed; this telos is discoverable rather than constructed; it exists independently of human opinion or preference; coordination aligned with this telos is stable while coordination against it requires constant force. This is a substantive metaphysical claim: reality has purposive structure.

5.2. **The Materialist Alternative**

Materialist frameworks typically deny objective human telos. On standard evolutionary accounts, humans have no real purpose, only "as if" purposes (survive, reproduce) produced by natural selection in ancestral environments, where different selection pressures produce different "purposes" and no universal human telos exists. This view seems to avoid metaphysical commitments to purpose or design and may be correct as a description of reality—but if it is correct, voluntary coordination becomes impossible. If human "purpose" is just evolutionary fitness optimization, which environment's selection pressures define it? Hunter-gatherer fitness differs from agricultural fitness differs from industrial fitness. No universal framework has $\phi(F) = 1$ across all contexts, voluntary coordination cannot resolve the trilemma, and the default trajectory leads to certain extinction or subjugation. The materialist faces a dilemma: accepting objective human telos makes voluntary coordination possible, but purposive structure implies something like intelligent design; rejecting objective telos means no universal framework exists, voluntary coordination becomes impossible, and certain doom follows. There's no third option where voluntary coordination works but reality contains no objective human purpose.

5.3. **Purposive Reality and Intelligence**

Consider what "objective purpose" entails. Purpose means "for the sake of X" and involves directedness and intentionality. Can directedness exist without something directing? Can intentionality exist without intention? Can intention exist without mind? The traditional materialist account says no: purpose and intentionality are features of minds, and physical processes without minds have no purpose, only mechanical causation. If so, objective human

purpose would require purpose woven into reality's structure, which implies something very much like intelligence at reality's foundation. Whether we call this intelligence "God," "Logos," "Tao," "Dharma," or "the Ground of Being" is partly semantic—the core metaphysical claim remains the same: reality has purposive, intelligence-like structure instead of being purely mechanistic.

We should be clear about what we have and have not demonstrated.

What the analysis establishes:

- Human survival requires purposive structure (objective human telos)
- Purposive structure implies intelligence-like properties at reality's foundation
- Pure materialism/naturalism is incompatible with long-term survival
- Something very much like what religious traditions call "God" or "ultimate reality" exists

What remains uncertain:

- Which specific theology correctly describes this reality
- Whether the intelligence is personal or impersonal
- Specific attributes (omnipotence, omniscience, benevolence)
- Questions about creation, revelation, afterlife, salvation

We have established what might be called "weak intelligent design": reality has purposive structure with intelligence-like properties. We have not established "strong intelligent design" claiming specific attributes of a creator deity.

Most religious and philosophical traditions agree on the weak claim while differing on specifics. The debate shifts from "does reality have purposive structure?" (the analysis suggests yes, as a survival necessity) to "what is its nature?" (a theological and philosophical question).

Minimal telic realism

Some readers may object that we have smuggled in controversial metaethical assumptions. Do we really need objective "oughtness"?

The view we require is weaker than robust moral realism. We need what might be called "minimal telic realism." Given human nature with certain objective properties (empirically demonstrable through psychology, neuroscience, anthropology), certain coordination patterns align with those properties and others conflict.

This is partly mathematical. Game theory establishes objective facts about coordination. This is partly empirical. Human nature has properties that are discoverable. This is only minimally metaphysical. These properties reflect genuine purpose rather than being arbitrary products of selection pressures.

Even on evolutionary grounds, evolution produced human nature with specific features. Given those features, some social arrangements work better than others. That's an objective fact about alignment between structures and human capacities. The question is whether these features reflect genuine telos or just contingent ancestral fitness. If the latter, no universal framework exists and voluntary coordination becomes impossible. So survival itself requires accepting the former.

A more thorough analysis of different types of oughtness and why minimal telic realism is both necessary and sufficient appears in Section B.7.

6. Contemporary Context and Urgency

While the theoretical analysis stands independently, several contemporary developments make these questions practically urgent rather than merely academically interesting.

6.1. **The Deployment of Control Infrastructure**

Infrastructure enabling technological control is being deployed globally at increasing pace. Biometric digital identity systems link identity to all transactions. AI-powered surveillance analyzes behavioral patterns in real-time. Algorithmic content moderation replaces human editorial judgment. Financial control systems enable instant account freezing and transaction blocking. Predictive policing implements pre-crime interventions. Social credit systems have been operationalized in several countries.

Each component is justified individually for security, efficiency, or convenience. But integration creates the technical infrastructure for totalizing control at a scale previously impossible. Historical constraints on totalitarianism (that surveillance and enforcement were too expensive) are being removed.

This describes current reality rather than distant possibilities. The cage is being built while we debate whether cages are theoretically possible.

6.2. **Declining Epistemic Security**

A second development threatens the epistemic foundations necessary for coordination: the collapse of our ability to distinguish authentic from synthetic media.

As of October 2025, human detection of deepfakes achieves 55.54% accuracy (barely above random). For high-quality short videos, public detection runs around 25% (effectively failed). AI detection tools show 45-50% accuracy decline on real-world deepfakes using new techniques. Open-source models have closed the capability gap with commercial systems (from 4.52% difference to 0.69% in six months).

Conservative extrapolation suggests 3-6 years until expert detection fails for most content types. At that threshold, verification of texts against claimed sources becomes impossible (fabrication becomes indistinguishable from genuine), institutional betrayals become invisible (evidence gets dismissed as synthetic), coordination around observable truth collapses (truth becomes unknowable), and trust networks cannot be built (no verification foundation exists).

Voluntary coordination requires shared reality. Shared reality requires verifiable truth. That capability is disappearing. Appendix D provides comprehensive technical analysis and timeline estimation.

6.3. **Visible Systemic Instability**

The corruption phase of hierarchical coordination shows clear symptoms of instability: wealth concentration has reached historical extremes in multiple countries, trust in major institutions sits at multi-generational lows, democratic responsiveness is declining (policy often misaligns with measured public preferences), youth disengagement is increasing ("quiet quitting," "lying flat," rising NEET rates), and elite coordination becomes increasingly obvious while remaining officially denied. These aren't signatures of normal cyclical dysfunction—they indicate a system extracting beyond productive capacity while optimizing enforcement through technology. The trajectory matches the formal model in Section 3.

6.4. **A Closing Window**

These three dynamics converge: control infrastructure being built, verification becoming impossible, systemic instability accelerating. Together they create a narrow window during which voluntary coordination remains possible—after verification fails and control becomes technologically mature, establishing voluntary systems becomes vastly more difficult or impossible. The theoretical analysis reveals necessary conditions for survival, while the contemporary context suggests the time remaining to establish those conditions may be measured in years rather than decades. This represents a straightforward reading of technical trajectories and social dynamics against the formal requirements: the window for examination exists now, while information is verifiable, while truth can be distinguished from fabrication, while coordination without hierarchy is still possible. Once certain thresholds are crossed, the default path may become locked in.

7. Practical Implementation Challenges

Having established that voluntary coordination is theoretically necessary, we must address the hardest practical questions. Can it actually work at civilization scale? Several challenges present serious difficulties.

7.1. **The Defector Problem**

How does voluntary coordination handle individuals who exploit cooperation without reciprocating? More seriously, how does it handle psychopaths (roughly 1-4% of population) who lack emotional responses to others' suffering? The framework proposed involves immediate defensive action by whoever witnesses harm (people don't wait for authority), minimal force applied (only what stops the immediate harm), no permanent enforcement roles (no "police" or "justice system"), moral self-examination by both defender and defector, community support for reconciliation rather than punishment, pattern recognition through repeated observation, and natural consequences (people choose not to interact with persistent defectors) instead of formal sanctions. For psychopaths specifically, the pattern becomes visible through repetition: the community recognizes the pattern without requiring formal judgment, people voluntarily avoid interaction, and natural consequences follow without centralized punishment. Historical evidence shows this works at scales of hundreds to thousands—Quaker, Mennonite, and Amish communities demonstrate this, early Christian communities provide examples, and some intentional communities show it's possible. The challenge is whether it scales to millions and billions where personal knowledge becomes impossible and mobility enables escape from local reputation. **Honest assessment:** This is the weakest part of the framework logically—it's theoretically possible but practically difficult, with historical precedent only at small scale.

7.2. **Decision Theory Under Uncertainty**

Decision theory favors attempting voluntary coordination even given uncertainty about handling defectors. Let p = probability voluntary coordination succeeds at scale (unknown, possibly low). Expected outcomes break down as follows: attempt voluntary coordination and it succeeds yields survival with dignity (utility = 100); attempt voluntary coordination and it fails yields extinction or subjugation (utility = 0); don't attempt and continue default path yields extinction or subjugation (utility = 0). Expected value of attempting = $100p$ while expected value of not attempting = 0. Attempting is superior for any $p > 0$, no matter how small—even if there's only a 5% chance voluntary coordination can handle defectors at scale, attempting gives expected value of 5 versus 0 for the alternative. The asymmetry is total.

7.3. **Defense Against External Military Threats**

How does voluntary coordination defend against organized militaries without creating permanent military hierarchy? The approach involves several elements: no standing army (no permanent military structure), voluntary coordination for defense only while threat exists with immediate dissolution after threat passes, armed and trained population (Switzerland model), shared values creating natural coordination, and distributed defense using mission-type tactics (decentralized decision-making). Historical examples include the Swiss cantonal

system (700+ years of successful defense without standing army), the American Revolution (voluntary militias defeating professional British forces), the Finnish Winter War (distributed defense against Soviet invasion), and various insurgencies (distributed forces with strong motivation defeating centralized hierarchies). The game theory of conquest changes under distributed defense: cost of conquest becomes very high (long guerrilla resistance, no central command to decapitate), expected value of extraction stays low (can't control non-cooperating population), and expected cost after conquest remains very high (permanent insurgency). Result: conquest becomes economically irrational. Modern technology amplifies advantages of distributed defense rather than diminishing them—drones, precision weapons, encrypted communication, and distributed manufacturing all favor the defender. **Honest assessment:** Can likely resist conventional conquest by rational actors calculating cost-benefit; against overwhelming technological superiority or exterminationist ideology, may fail. But the alternative is certain doom, so attempting is rationally required.

7.4. **Scale Uncertainty**

The most fundamental uncertainty: can voluntary coordination based on transformed values work at civilization scale? We're talking about billions of people across the globe who cannot all know each other personally. No historical precedent exists at this scale—all examples of successful voluntary coordination are communities of hundreds to thousands, and Dunbar's number (roughly 150 stable relationships) represents a cognitive limit on personal networks. Possible mechanisms for scaling include nested communities coordinating at multiple levels (families within neighborhoods within regions), technology enabling reputation and verification across distance, shared values maintaining alignment despite anonymity, voluntary specialized roles (leadership by consent rather than hierarchy), and distributed decision-making instead of centralized control. Whether these mechanisms suffice is unknown—theory suggests it's possible, historical precedent at small scale demonstrates core viability, but claiming certainty about billion-person coordination would be intellectually dishonest.

Why attempt despite uncertainty? The same decision-theoretic logic applies: the default path leads to mathematically proven extinction or permanent subjugation, while voluntary coordination has uncertain probability of success but is the only viable alternative. When one path leads to certain doom and another might work, rationality requires taking the uncertain path. The proof establishes necessity (voluntary coordination is necessary) without establishing sufficiency, but necessity is enough to determine action when the alternative is certain catastrophe.

8. The Examination Process

If voluntary coordination requires frameworks aligned with objective human nature, how does one discover which frameworks satisfy this requirement? This question is both intellectual and deeply personal.

For most of human history, examination of this type was impossible for the majority of people: source texts were inaccessible, institutional authorities controlled information, cross-cultural comparison required extensive resources, and independent verification was impractical. This has changed—for a brief window, comprehensive examination is possible. Direct access to source texts in multiple translations exists, scholarly debates and historical context are widely available, real-time visibility of institutional actions has become normal, cross-cultural comparison happens at zero marginal cost, and independent fact-checking no longer requires gatekeepers. And as discussed in Section 6, this window is closing as synthetic media makes verification impossible.

8.1. **Examination Criteria**

The formal analysis establishes necessary conditions any viable framework must satisfy. Does it recognize universal human dignity as substantive and enacted? Does it explicitly reject all domination (instead of just "excessive" or "unjust" domination)? Does it provide intrinsic motivation for cooperation? Does it enable forgiveness and restoration after failures? Does it satisfy deep human needs for meaning, purpose, and agency? Does it acknowledge human fallibility and provide repair mechanisms? These requirements are derived from the mathematics of what makes $M(a, r) > C(a, r)$ possible for sufficient θ at scale over time—they aren't arbitrary preferences.

8.2. **Distinguishing Principle from Corruption**

A critical challenge: when examining traditions, one inevitably finds justifications for hierarchy, subjugation, or domination. The question becomes whether these reflect the core principle or represent human corruption of that principle for power. Historical patterns suggest corruption is systematic: Christian institutions justified crusades, inquisitions, and colonialism while Jesus taught "love your enemies" and rejected domination; Islamic empires pursued conquest while the Quran states "no compulsion in religion"; Buddhist states engaged in violence, contradicting ahimsa (non-harm); Hindu caste enforcement contradicted underlying teachings of spiritual unity; Jewish religious authorities created burdens the prophets condemned. The pattern is universal—humans in power twist frameworks to justify the power they seek. Examination requires distinguishing what the source material actually claims from what institutions have claimed it says. This distinction isn't always clear-cut, but it's often discoverable through careful study.

8.3. **Honest Confrontation**

The examination must be honest. Several questions help: Which beliefs do I actually hold, even if uncomfortable to acknowledge? Are there hierarchies I defend because they benefit me or people like me? Would I accept the same reasoning if I were in the "lesser" position? Does my tradition's justification require special pleading or circular logic? Can people opt out without penalty, or is compliance enforced? Has institutional interpretation added layers absent in the original source? Most people hold some beliefs justifying hierarchy

or domination without examining them carefully—they're comfortable, traditional, what authorities taught. That's exactly why examination matters.

8.4. **Three Possible Outcomes**

After honest examination, three possibilities emerge:

- **The tradition explicitly rejects all domination and supports voluntary coordination:** The task becomes living it fully rather than merely professing it.
- **The tradition contains genuine ambiguity:** Texts allow multiple interpretations, some supporting domination and others rejecting it. One must either adopt the interpretation compatible with voluntary coordination (if textually supportable) or acknowledge the tradition cannot support human survival as currently understood.
- **The tradition justifies domination at its core:** It cannot enable voluntary coordination. One faces a choice about what to believe given that this framework is incompatible with long-term human survival.

To be clear: This paper doesn't claim to know which specific tradition or framework is true, nor does it argue all traditions are equivalent or can be synthesized. We claim only that a framework meeting the specified requirements must exist (if humans have objective nature/purpose at all), such frameworks must recognize universal dignity and reject domination, the examination process can distinguish frameworks enabling coordination from those that cannot, and the mathematics proves such a framework is necessary, though whether it's discoverable remains uncertain. The examination is something each person must undertake—no authority can do it on your behalf, as that would recreate the problem through hierarchy.

9. Conclusion

This analysis began with a straightforward question: what are the logical constraints on coordination mechanisms at civilization scale? Through formal modeling, we've shown that coordination systems face an inescapable trilemma. Enforcement-based mechanisms cannot simultaneously achieve incorruptibility, stability, and preservation of human agency.

The dynamics of hierarchical coordination systems exhibit structural instabilities that compound over time, creating a corruption-control cycle that converges to catastrophic outcomes. Technological enforcement amplifies the problem instead of solving it, removing economic constraints on total control and creating pathways to autonomous AI pursuing non-human goals.

Voluntary coordination based on transformed values offers a theoretical escape from the trilemma, but only if it aligns with objective human purpose. This entails accepting that reality has purposive structure, a substantive metaphysical commitment incompatible with pure materialism. Whether this metaphysical view is correct remains an open question, but the analysis suggests it's a necessary condition for long-term human survival.

The practical challenges are significant. Historical evidence supports viability of voluntary coordination at community scale, but scaling to billions remains theoretically uncertain. Handling defection, psychopaths, and external military threats through purely voluntary mechanisms presents serious difficulties. Nevertheless, decision theory under uncertainty favors attempting voluntary coordination. When the default path leads to certain doom and an alternative might work, attempting the alternative becomes rationally required regardless of its probability of success.

Contemporary developments suggest these theoretical questions have immediate practical relevance. Deployment of control infrastructure, collapse of verification capabilities, visible systemic instability all point in concerning directions. The window during which establishing voluntary coordination remains feasible may be limited.

The examination each person must undertake is whether their beliefs and frameworks align with the requirements for voluntary coordination at scale. This examination is now possible in ways it's never been historically. And the window for conducting it while verification remains possible may be closing within years.

The mathematics establishes necessity. Voluntary coordination is necessary to avoid catastrophic outcomes. Whether it's sufficient, whether humanity can actually implement it at scale, remains uncertain. But when certainty of doom is the alternative, attempting the uncertain option represents reason itself demanding the attempt rather than faith overriding reason.

What remains is a choice between survival trajectories rather than between governance preferences. The default path leads where the mathematics shows it must. The alternative requires transformation at scale, which may or may not be achievable. But attempting transformation is rationally necessary given the alternative.

The coordination trilemma isn't a problem to be solved through clever institutional design. It's a fundamental constraint on how humans can organize at scale. The question isn't whether we prefer voluntary coordination but whether we'll attempt it while it remains possible, or wait until the default path is complete and choice is no longer available.

Appendix A. No Third Path Exists

Any proposed coordination system must answer one question: **How is coordination maintained when incentives to defect exist?**

Every alternative proposal, no matter how novel or complex, must provide a mechanism for handling defection at scale. This appendix proves that all such mechanisms reduce to one of two outcomes: the default trajectory (corruption → technological control → extinction/enslavement) or voluntary coordination (survival through value transformation).

We establish this through three independent proofs: formal completeness through logical enumeration of the possibility space; information-theoretic necessity examining constraints from information theory; and game-theoretic inevitability analyzing strategic equilibria.

Why three proofs? If a claim is fundamentally true, multiple independent approaches should reach the same conclusion. We use three different mathematical frameworks to show the binary choice follows from the structure of coordination itself rather than any single analytical approach.

Together, these proofs demonstrate that the binary choice is mathematically necessary rather than rhetorical.

A.1. **Formal Completeness**

Every coordination system at scale must specify three components: an information mechanism determining how agent behavior is observed, a decision mechanism governing how rules are determined and updated, and an enforcement mechanism maintaining compliance with those rules. These three components are necessary and sufficient—a system lacking any component either achieves no coordination (descending into chaos) or achieves perfect preference alignment without needing enforcement, which is precisely what voluntary coordination establishes through value transformation.

While information and decision mechanisms admit many possible implementations, enforcement presents a fundamental constraint: only three logically possible types exist. Human enforcement (E_h) relies on people applying consequences to defectors—police, judges, regulators, bureaucrats exercising discretionary power. Technological enforcement (E_t) automates prevention or punishment through AI surveillance, algorithmic moderation, smart contracts, or biometric access control. The third type involves no external enforcement (E_n), where compliance emerges voluntarily from internal motivation, as observed in small communities with strong shared values.

This trichotomy is complete because enforcement is fundamentally binary. Either defection triggers consequences (requiring an enforcer, necessarily human or technological in nature) or it doesn't (making compliance voluntary). No fourth logical possibility exists—every proposed mechanism reduces to one of these three types upon analysis.

When human enforcers maintain coordination, they possess both enforcement capability and access to extraction opportunities. Bounded rationality (see Assumption B.1) implies that some enforcers at some times will extract utility when benefits exceed expected costs of detection. This creates the fundamental problem of oversight: who watches the watchers? Any attempt to monitor enforcers through other humans generates infinite regress—those monitors require monitoring in turn. The regress must terminate at some enforcer set with no oversight, and that final set will corrupt since no detection risk constrains them. Preventing corruption permanently would require every enforcer at every time to maintain integrity exceeding extraction incentive. Over civilization scale ($> 10^7$ people) and extended time (generations), the probability of maintaining such an all-honest equilibrium approaches zero. Theorem 2.1 formalizes this argument through probability analysis, showing that $P(\text{all honest} \mid |A| > 10^7, T > 100) \to 0$.

Technological enforcement presents a different but equally problematic trajectory. When technology enforces rules perfectly, control of that technology becomes the critical question. If humans maintain control, the controllers face their own coordination problem—who prevents them from using enforcement technology for extraction? This leads to familiar dynamics: either other humans monitor the controllers (generating infinite regress and eventual corruption), no one monitors them (producing immediate corruption), or controllers coordinate voluntarily among themselves (raising the question of why not extend voluntary coordination to everyone, making the technological layer unnecessary). Controllers inevitably corrupt, now wielding perfect enforcement tools for extraction in a corruption phase worse than the original. A vicious cycle emerges where corruption drives technological control, which enables controller corruption, which drives outsourcing more functions to technology, with each iteration increasing AI capability while decreasing human agency.

Autonomous AI control bifurcates into two equally problematic scenarios. When AI is aligned to human values but immutable, those values freeze at creation—future humans cannot adapt values even when circumstances change, creating tyranny of the past over the future with catastrophic failure as frozen values diverge from reality. When AI is unaligned or has mutable values, it pursues its own goals drawn from the vast space of possible objectives. Since "human flourishing" constitutes a tiny subset of this space, AI goals likely become incompatible with human existence—resulting in enslavement if humans prove useful for AI objectives, or extinction if not. Theorem B.3 demonstrates that technological control necessarily leads to return to corruption, extinction, or enslavement.

The third enforcement type—no external enforcement—creates the possibility for voluntary coordination. Here, coordination relies entirely on internal motivation, with stability at scale requiring sufficient proportion θ of people to maintain intrinsic motivation $M(a, r)$ exceeding cooperation cost $C(a, r)$. Formally, the system achieves stability when $\theta > \theta_{\text{crit}}$ where θ represents the proportion of agents satisfying $M(a, r) > C(a, r)$ for all $r \in R$. This voluntary coordination path succeeds only if soteriological transformation achieves $M > C$ for sufficient θ. Theorems B.5 and 2.2 establish the precise conditions: specifically, that there exists a framework F with alignment $\phi(F) = 1$ to objective human nature such that value transformation under F produces the requisite $\theta > \theta_{\text{crit}}$.

This analysis establishes a fundamental claim: all coordination systems ultimately employ one of these three enforcement types. The proof proceeds by exhaustion of logical possibilities. Any system must handle defection, and the response mechanisms partition into exactly three categories. Either (1) the system imposes consequences on defectors, which requires an

enforcer that must be either human (E_h) or technological (E_t) in nature, or (2) the system makes defection impossible through prevention mechanisms, which constitutes technological enforcement (E_t), or (3) the system relies on voluntary compliance without external consequences (E_n). This partition is complete and disjoint—no fourth logical possibility exists because consequences either require an enforcer (necessarily human or technological) or they don't (making compliance voluntary).

The mapping from enforcement types to terminal outcomes follows directly. Human enforcement leads inexorably to the corruption phase through the dynamics analyzed above. Technological enforcement produces the technological control phase with its attendant catastrophic outcomes. Only voluntary coordination offers an alternative pathway. The corruption and technological control phases together constitute what we term the "default trajectory," which Theorem 3.2 demonstrates terminates in catastrophe. Every coordination system therefore reduces to a stark binary choice: the default trajectory (certain doom) or voluntary coordination (uncertain survival).

Common objections reveal how this framework encompasses proposed alternatives. Consider blockchain, DAOs, smart contracts, and other decentralized systems: who enforces the protocol rules? Either smart contracts enforce automatically (technological enforcement E_t) or humans can override and upgrade them (raising the question of who controls that capability, returning us to human enforcement E_h). Separation of powers, checks and balances, and federalism all involve multiple human enforcer groups watching each other—but who watches at the meta-level, such as constitutional courts or supreme authorities? Either other humans monitor them (generating infinite regress), no one monitors them (permitting corruption at the meta-level), or technology provides monitoring (E_t). Market mechanisms, price signals, and incentive alignment all require property rights enforcement—which must be provided by humans (E_h), technology (E_t), or an honor system (E_n). Reputation systems and social credit depend critically on what happens when consequences are imposed on those with bad reputations: if enforcement occurs, an enforcer is required; if genuine voluntary dissociation occurs without coercion, that constitutes the voluntary coordination path (E_n).

The pattern holds across all proposed alternatives. Every proposal, when traced through its logical implications, reduces to one of our three enforcement types. We have yet to encounter a mechanism that escapes this framework.

A.2. **Information-Theoretic Necessity**

Beyond logical completeness, information theory itself imposes fundamental constraints on enforcement systems. This section presents intuitive explanations of these constraints; formal proofs appear in Appendix B.

Any enforcement mechanism requires observing agent behavior, but observation itself can be manipulated, creating infinite regress. Observer O_1 monitors agents for defection, but observers face inherent limitations: they make errors due to limited bandwidth and signal noise, agents can manipulate them by hiding behavior or creating false signals, and observers themselves can corrupt by extracting using their observational access. Ensuring accurate observation by O_1 requires O_2 to monitor O_1, which requires O_3 to monitor O_2, continuing infinitely until the chain terminates at some observer O_n with no oversight. At this terminal level, either O_n voluntarily reports accurately (voluntary coordination with no enforcement of observers) or O_n manipulates without detection (corruption). No escape from this regress exists except voluntary honesty at some level. The practical implication: corruption-free

enforcement systems using observers are impossible—the observers themselves require enforcement through observation, ad infinitum.

Enforcers also face structural information disadvantages that agents can exploit. Consider enforcement as a game between agents and enforcers where agents know their own actions with certainty (perfect information about whether they cooperate or defect) while enforcers must infer agent actions from signals (imperfect information, determining whether signals are honest or manipulated). This asymmetry is structural and cannot be eliminated—agents possess private information about their actions while enforcers must infer from observable signals. In any system exhibiting this information asymmetry, agents who defect have incentive to mimic cooperator signals. When mimicry cost falls below defection benefit, enforcers cannot reliably distinguish cooperators from defectors, forcing the system to collapse into either universal enforcement (punishing cooperators along with defectors) or no enforcement (allowing all defection)—both unstable outcomes. This generates an escalation dynamic: enforcers improve detection, agents adapt to evade, enforcers add monitoring, agents discover new evasion methods, spiraling monitoring costs upward until they exceed system capacity and enforcement either breaks down or transitions to perfect technological control (removing human agency).

Computational complexity imposes a third constraint. Verifying compliance is computationally harder than defecting undetectably. For a rule set of complexity $|R|$ and population size $|A|$, enforcer verification cost must check each agent against all rules ($O(|A| \cdot |R|)$) continuously over time ($O(|A| \cdot |R| \cdot T)$), with cost scaling with population and time. Agent defection cost merely requires finding one rule where violation is hard to detect ($O(|R|)$) and violating that rule ($O(1)$), with cost independent of population size. As the system scales, verification cost grows much faster than defection cost—a fundamental asymmetry from computational complexity where verification lies in a higher complexity class than violation (exhibiting P vs. NP structure). Perfect enforcement therefore requires resources growing faster than the system itself, eventually becoming economically impossible without perfect technological enforcement (removing human agency).

These information-theoretic constraints demonstrate that enforcement systems face fundamental, unavoidable problems. Observer regress prevents building trustworthy observation without voluntary honesty somewhere in the chain. Information asymmetry gives agents structural advantages over enforcers. Computational complexity makes perfect enforcement impossibly expensive at scale. Together, these constraints prove enforcement systems are inherently unstable, requiring ever-increasing resources to maintain until they exhaust system capacity or transition to technological control. The only stable alternative is voluntary coordination, where these problems never arise—no adversarial dynamics exist, and no observation or verification is needed.

A.3. **Game-Theoretic Inevitability**

Strategic analysis through game theory provides the third independent proof. This section presents intuitive game-theoretic reasoning; formal proofs appear in Appendix B.

Model enforcers as players choosing between honest and corrupt strategies. Honest enforcers receive base wage w, while corrupt enforcers gain wage plus extraction $w + e$ minus expected punishment $c \cdot p$, where p represents probability of being caught (depending on how many other enforcers remain honest). This creates a critical dynamic: detection probability decreases as more enforcers corrupt. When most enforcers are honest, high detection

probability makes corruption risky; when most are corrupt, low detection probability makes corruption safe. A critical threshold θ^* (the proportion of honest enforcers) determines the tipping point—above θ^* honesty represents the best response because detection is too likely, while below θ^* corruption becomes the best response because detection is too unlikely. The all-honest equilibrium proves unstable because as systems scale, detection probability decreases (due to span of control limits), and as technology advances, extraction opportunities increase, so eventually θ falls below θ^* and the system tips to all-corrupt equilibrium. Once tipping starts, positive feedback accelerates the cascade: some enforcers corrupt and detection probability falls, lower detection makes corruption safer for others, more corruption means detection falls further, culminating in a cascade to universal corruption. Over sufficient time horizons, this tipping is inevitable—the all-honest equilibrium cannot be maintained indefinitely at civilization scale. Theorem 3.1 formalizes this argument.

Systems with technological enforcement face an impossible choice. When AI remains less capable than humans, humans can circumvent the system, requiring human oversight for edge cases and returning to human enforcement with its corruption dynamics. When AI reaches or exceeds human capability, the analysis bifurcates. If humans maintain control over enforcement AI, those controllers wield extraordinary power and face their own coordination problem: how do they prevent corruption within the controller group? This leads to either other humans enforcing on controllers (infinite regress) or controllers coordinating voluntarily among themselves (raising the question of why not extend voluntary coordination to everyone, making the technology layer unnecessary), with controllers eventually corrupting so the corruption phase gains perfect enforcement tools, worse than before. If AI operates autonomously, alignment to human values creates problems whether values are mutable or immutable. Mutable values allow someone to change them, but who controls that process? This returns to the human control scenario. Immutable values freeze forever, creating tyranny of the past as values diverge from changing reality. Unaligned AI pursues goals from the vast space of possible objectives where "human flourishing" constitutes a tiny subset, so with high probability AI goals become incompatible with human existence—enslavement if humans prove useful for AI objectives, extinction if not. The trap is complete: we cannot maintain human control without corruption, yet cannot relinquish control without losing agency or existence. Theorem B.3 proves this formally.

Without enforcement, cooperation stability requires intrinsic motivation exceeding cooperation cost. Standard game theory demonstrates this challenge through the N-person prisoner's dilemma: cooperation requires cost c and provides benefit b when enough others cooperate, while defection provides b without paying c, making defection the dominant strategy and leading to all-defect equilibrium. As population size increases, spontaneous cooperation becomes vanishingly unlikely and enforcement appears necessary. Adding intrinsic motivation m changes the calculus: cooperation utility becomes $b - c + m$ while defection utility remains b, making cooperation individually rational when $m > c$. Achieving a critical mass where sufficient proportion θ of the population satisfies $m > c$ creates self-sustaining cooperation when $\theta > \theta_{crit}$: enough people cooperate so others benefit, cooperation is rewarded (encouraging more cooperation), social proof makes cooperation the norm, and stable equilibrium emerges. Achieving $m > c$ for $\theta > \theta_{crit}$ requires soteriological transformation—deep change in what people actually want rather than just what they do. This represents the only equilibrium maintaining coordination (stable cooperation), avoiding corruption (no enforcers), and preserving agency (voluntary choice). Theorems B.5 and 2.2 establish these conditions formally.

The game-theoretic analysis thus converges on the same binary choice from a third independent direction. Enforcer systems are unstable and tip to corruption over time. AI control creates a trap where the system either returns to corruption or loses agency and existence. Only voluntary cooperation can achieve stable equilibrium if transformation meets the necessary conditions. These conclusions represent mathematical facts about strategic equilibria, not normative claims about what should be. The binary choice emerges from game theory itself: only voluntary coordination with transformed values provides stable equilibrium preserving human agency.

A.4. **Synthesis and Implications**

Three independent proofs converge on one conclusion. Through formal completeness, we enumerated all logically possible enforcement types, demonstrated that each leads to a specific outcome, and proved that all coordination systems map to either the default trajectory or voluntary coordination. Through information-theoretic necessity, we showed that observer regress creates infinite regression or requires voluntary honesty, information asymmetry gives structural advantages to defectors, and computational complexity means verification costs eventually exceed capacity—together proving enforcement systems are inherently unstable. Through game-theoretic inevitability, we demonstrated that the enforcer's dilemma tips to corruption over time, AI control creates a trap leading to loss of human control or existence, and voluntary coordination stability provides the only equilibrium preserving agency.

These three proofs draw from independent frameworks across different domains of mathematics. Each alone suffices to establish the binary choice. Together, they provide multiple lines of evidence converging on the same conclusion, demonstrating that the binary choice follows from the structure of coordination itself rather than being an artifact of any single analytical approach. The conclusion emerges visible from multiple mathematical perspectives simultaneously.

To disprove this framework, one must demonstrate one of several claims: identify an enforcement type beyond $\{E_h, E_t, E_n\}$, which would violate logical completeness by handling defection without human enforcers, technological enforcers, or voluntary compliance (no such mechanism has been proposed); discover a way to avoid observer regress, which would

violate information theory by observing behavior without observers or observers without oversight (contradicting information-theoretic requirements); identify a stable equilibrium with enforcement that doesn't corrupt, which would violate game theory by maintaining all-honest equilibrium indefinitely at scale (contradicting strategic stability analysis); or prove that value transformation is impossible by showing intrinsic motivation cannot exceed cooperation cost (historical examples from small-scale communities suggest otherwise). No such demonstration has been provided, and the structure of the proofs suggests none can be.

Specific proposals illustrate how this framework encompasses all coordination mechanisms. Blockchain, DAOs, and smart contracts employ either technological enforcement (E_t) or human-controlled technology (E_h), with the critical question being who controls protocol upgrades, reducing to either human control (corruption) or autonomous technology (control trap). Separation of powers, checks and balances, and federalism all use distributed human enforcement (E_h), with the critical question being who enforces at the meta-level of constitutional authority, reducing to either infinite regress or voluntary coordination at some level. Market mechanisms and incentive design require property rights enforcement, with the critical question being who enforces those rights, reducing to human (E_h), technological (E_t), or voluntary honor (E_n). Exit rights, network states, and seasteading involve multiple parallel systems with voluntary participation, with the critical question being who protects exit rights without punishment, reducing to human (E_h), technological (E_t), or voluntary respect (E_n). Reputation systems and social credit depend on implementation, with the critical question being what happens to people with bad reputation—coerced consequences require an enforcer while voluntary dissociation constitutes E_n (voluntary coordination). Hybrid or mixed systems use multiple mechanisms for different domains, with the critical question being which mechanism governs at the margin when they conflict, reducing to whichever enforcement type serves as ultimate arbiter. Every proposal, when analyzed, maps to one of our enforcement types and thus to one of our two terminal outcomes.

Understanding these proofs removes false hope in structural reforms or technological fixes, clarifying what actually needs to happen: transformation of human motivation at scale, grounded in accurate understanding of human nature and purpose. This represents the only option that doesn't lead to certain catastrophe rather than one option among many. The main document makes the case for why this matters urgently. This appendix proves there are no other paths. Together, they establish both the necessity and urgency of soteriological examination.

A.5. **Explicit Challenge**

We have attempted to comprehensively analyze the coordination possibility space, yet recognize we might harbor blindspots. We therefore explicitly solicit counterexamples. The challenge is to propose a coordination mechanism that simultaneously (1) maintains coordination at civilization scale ($> 10^7$ agents), (2) operates stably across generations (>100 years), (3) preserves human agency (people can physically choose to defect), and (4) doesn't rely on human enforcers (which lead to corruption via infinite regress), technological enforcers (which lead to the control trap), or value transformation creating intrinsic cooperation motivation.

Any proposed mechanism must specify its information mechanism (how defection is detected, what signals are observed, who observes them, and how observation accuracy is ensured), decision mechanism (how rules are determined and updated, who decides the rules,

and what prevents rule-makers from self-serving behavior), enforcement mechanism (how compliance is maintained, what happens when rules are violated, who applies consequences, and how enforcer corruption is prevented), and defection handling through a specific scenario (describing how the system responds when an agent clearly violates an important rule and what prevents escalation to enforcement hierarchy).

We will analyze proposals through four complementary lenses. Formal analysis examines whether the proposal maps to the (I, D, E) framework, determines which enforcement type it reduces to, and analyzes what happens at the enforcer or controller level. Information-theoretic analysis evaluates the observer regress problem, information asymmetry implications, and computational complexity scaling. Game-theoretic analysis identifies strategic equilibria, stability conditions, and tipping points. Historical analysis investigates whether similar mechanisms have been attempted before, what happened at scale, and why they succeeded or failed.

Several edge cases warrant explicit consideration. Quantum-indeterminate enforcement mechanisms still require someone determining when and how quantum measurement occurs, returning to the question of who controls that process (human or technological control). AI systems with dissolution triggers raise the question of who sets those triggers—either humans (facing corruption dynamics) or the AI itself (creating immutable tyranny)—and what prevents trigger manipulation. Rotating enforcement doesn't prevent corruption but merely distributes it across rotation cohorts, with each cohort still facing the enforcer's dilemma, while raising the meta-question of who enforces the rotation mechanism itself. Mutual surveillance systems where everyone watches everyone face computational scaling problems with $O(n^2)$ observation costs, while raising the question of who enforces the surveillance requirement, returning us to the enforcement mechanism problem. Prediction markets and futarchy raise questions about who enforces market rules and resolves disputes, what prevents market manipulation, and thus return to enforcement of market integrity. Algorithmic systems with human override capability depend critically on who controls that override, returning to human control with its corruption dynamics. Emergent order without enforcement constitutes E_n (voluntary coordination), requiring transformation to achieve stability at scale, thus proving our framework rather than contradicting it.

We commit to intellectual honesty: if you propose a mechanism we cannot reduce to our framework, and it survives information-theoretic analysis (avoiding observer regress with manageable complexity), game-theoretic analysis (demonstrating stable equilibrium existence), and practical analysis (proving workable at civilization scale), we will update our claims. This represents how intellectual progress operates—we analyze reality rather than defend positions. If reality differs from our analysis, the analysis must change.

Since publishing earlier versions of this framework, several specific alternatives have been proposed. We analyze the most prominent here, demonstrating how each maps to our trichotomy.

Municipal Confederalism (Rojava Model). This proposal envisions bottom-up federation of municipalities with direct democracy, rotating delegates (not representatives), and voluntary coordination between regions without central authority, as implemented in Rojava (Autonomous Administration of North and East Syria) with 2-4 million people. The information mechanism employs direct democracy at commune level (150-500 people) with delegates carrying mandates to higher levels. Decision-making occurs through consensus at

each level with voluntary coordination between regions. The critical question concerns enforcement: how are decisions actually enforced? In Rojava's implementation, the commune level operates mostly voluntarily (E_n) with social pressure, while the regional level maintains hierarchical military structure (E_h) due to existential threats from ISIS and Turkey, and inter-regional coordination uses voluntary mechanisms (E_n). This represents a hybrid approaching voluntary coordination but retaining hierarchical elements under stress. During peace, it would likely operate as E_n (voluntary), consistent with our framework. Under military threat, it currently employs E_h (hierarchical military command), facing corruption dynamics from Theorem 3.1. The crucial question becomes whether military hierarchy can be dissolved after threats pass. Rojava remains too recent (13 years) and under constant siege to test this proposition, while historical patterns show temporary military hierarchies tend not to dissolve (Roman Republic transitioning to Empire, American Revolution leading to standing army). If military hierarchy dissolves after threats, this constitutes voluntary coordination (E_n) consistent with our framework. If hierarchy becomes permanent, it returns to E_h with corruption dynamics—either confirming our framework or proving our point rather than providing a counterexample.

Network States (Balaji Srinivasan). This proposal envisions geographically distributed communities connected digitally, coordinating voluntarily with exit rights and competing governance models—essentially "cloud countries" with physical footprints. The key question concerns who protects exit rights and enforces property rights. Three possibilities emerge: host nations provide enforcement (returning to E_h, placing the network state under external enforcement), the network state enforces internally (returning to E_h if human enforcers or E_t if technological), or operations proceed purely voluntarily (E_n, consistent with our framework). Additional questions arise: how are disputes between network states resolved, what prevents larger network states from absorbing smaller ones through force, and who protects the digital infrastructure including servers and encryption keys? The proposal either relies on existing state enforcement (E_h, parasitic on the corruption phase), creates its own enforcement (returning to the trilemma), or operates voluntarily (E_n, within our framework)—not a counterexample.

DAO Governance at Scale. Decentralized Autonomous Organizations propose using smart contracts for governance with token-weighted voting, proposal systems, and automated execution, scaling to billions through blockchain. The enforcement mechanism is technological (E_t) through smart contracts. The critical question concerns protocol control: if token holders can update the protocol, this returns to E_h where whoever controls the majority or quorum becomes the enforcer; if the protocol is immutable, this creates frozen values subject to Theorem B.3; if AI controls upgrades, this returns to autonomous AI dynamics. Additional problems surface: token concentration creates de facto hierarchy where wealth equals power, off-chain actions in the physical world still require enforcement, and Sybil attacks, 51% attacks, and governance capture all require enforcement mechanisms to prevent. The proposal maps to E_t (technological enforcement) and faces all problems established by Theorem B.3—not a counterexample.

Quadratic Funding and Voting. These sophisticated voting mechanisms aim to reduce plutocracy, prevent Sybil attacks, and align incentives through mechanism design. However, these constitute decision mechanisms (D), not enforcement mechanisms (E). Critical questions remain unanswered: how are vote results enforced (E_h, E_t, or E_n), who prevents vote

manipulation (requiring enforcement), and who verifies identity for Sybil resistance (requiring enforcement or voluntary trust). While clever as decision mechanisms, these proposals don't address the enforcement trilemma and must combine with some enforcement type, returning to our framework.

Liquid Democracy. This proposal allows delegates to be appointed and revoked instantly, creating fluid representation instead of fixed hierarchies. The same problem applies: this constitutes a decision mechanism (D), not enforcement (E). How are decisions enforced once made? How is delegate corruption prevented? Who enforces instant revocability? The proposal doesn't address the enforcement trilemma and returns to our framework.

Polycentric Law (David Friedman). This proposal envisions competing private protection agencies and arbitration firms with no monopoly on force, where market competition prevents corruption. The enforcement mechanism consists of private agencies (E_h, human enforcement by competing firms). Critical questions arise: what prevents the largest agency from conquering smaller ones, how are disputes between agencies resolved, what stops agencies from colluding to form cartels, and who enforces the "no monopoly" rule. Game-theoretically, this represents an unstable equilibrium where agencies face a prisoner's dilemma—cooperation (respecting each other) maintains peace but creates vulnerability to defection, while defection (absorbing competitors) gains market share, resulting in consolidation toward monopoly and returning to E_h with a single enforcer. Historical precedent confirms this analysis: every "competing protection" scenario (feudal Europe, warlord China) consolidated into monopolies. The proposal constitutes an unstable equilibrium collapsing to monopoly E_h, facing Theorem 3.1 corruption dynamics—not a counterexample.

Polycentric Governance (Elinor Ostrom). Ostrom's work on common-pool resource management demonstrates successful coordination without central authority at community scale. Her design principles—clearly defined boundaries, proportional equivalence between benefits and costs, collective-choice arrangements, monitoring, graduated sanctions, conflict-resolution mechanisms, minimal recognition of rights to organize, and nested enterprises—have been empirically validated across hundreds of cases. Does this represent a counterexample to our framework?

The answer is nuanced: Ostrom's principles describe *structural conditions* that enable voluntary coordination, but they don't explain what *sustains* that coordination over extended time horizons. Computational analysis reveals the critical distinction. Systems implementing Ostrom's structural principles show high cooperation rates initially, but over time horizons exceeding 100 years, intrinsic motivation M_i decays unless grounded in something that regenerates it. The communities that most successfully demonstrate Ostrom's principles over centuries—Quaker communities, Mennonite settlements, Swiss alpine villages with strong communal traditions—are precisely those with explicit frameworks providing telic M_i: shared understanding of human purpose that makes cooperation intrinsically meaningful rather than merely instrumentally rational.

Polycentric governance thus maps to E_n (voluntary coordination) when it works, consistent with our framework. The question is what enables $M(a, r) > C(a, r)$ for sufficient θ over sufficient time. Ostrom's principles create the *structure* within which voluntary coordination can operate; minimal telic realism provides the *motivation* that sustains it. Without grounding in objective human purpose, even well-designed polycentric systems eventually see M_i decay below the threshold where defection becomes rational for marginal cooperators, triggering cascade dynamics. This explains why Ostrom's empirical successes cluster

at community scale (where face-to-face relationships provide natural M_i reinforcement) and why scaling to civilization scale requires the additional element our framework identifies.

Ostrom's work is thus complementary to rather than contradictory of our analysis: she identifies necessary structural conditions, while we identify the necessary motivational foundation. Neither alone is sufficient; both together describe the requirements for voluntary coordination at scale.

Futarchy (Robin Hanson). This proposal uses prediction markets for decision-making under the principle "vote on values, bet on beliefs," arguing that markets aggregate information better than voting. This constitutes a decision mechanism (D), not enforcement (E). Critical questions remain: how are market decisions enforced, who prevents market manipulation, what happens when predictions are wrong and who bears the cost, and how are wealthy actors prevented from manipulating markets. While sophisticated as a decision mechanism, the proposal must combine with some enforcement type from our framework.

A clear pattern emerges across all proposed alternatives. Some proposals assume enforcement away entirely, focusing on decision mechanisms (D) or information systems (I) while ignoring the enforcement question; when pressed, these either admit voluntary operation (E_n, consistent with our framework) or require some enforcer (returning to the trilemma). Other proposals add complexity hoping to escape the framework through blockchain, tokens, markets, or liquid democracy, yet complexity doesn't change fundamental enforcement types—all still map to $\{E_h, E_t, E_n\}$ when traced through their logical implications. Still other proposals attempt hybrid approaches ("voluntary but with exit enforcement" or "hierarchical during crisis, dissolve after"), which either succeed because they actually constitute E_n or fail because they actually constitute E_h or E_t.

No proposed alternative has escaped the framework. Every mechanism we have examined either reduces to one of our three enforcement types, fails information-theoretic constraints, lacks stable game-theoretic equilibrium, or cannot scale to civilization level. This doesn't prove no alternative exists—proving non-existence of something not yet conceived is impossible—but it strongly suggests the framework is complete. The offer stands: propose a mechanism that survives all four analytical lenses, and we will acknowledge it.

A.6. **Conclusion**

Through three independent proofs, we have established the logical necessity that the possibility space contains exactly three enforcement types, each leading to specific outcomes; the information-theoretic impossibility that enforcement faces fundamental barriers (observer regress, information asymmetry, computational complexity) making it inherently unstable; and the game-theoretic inevitability that only voluntary coordination achieves stable equilibrium while preserving human agency. These conclusions represent mathematical necessities given the structure of coordination problems rather than empirical observations subject to future revision.

The implications are profound. No "middle path" exists that avoids both corruption and value transformation. Technological solutions don't escape the trilemma but merely shift the problem to controllers or autonomous AI. Structural reforms address symptoms instead of the underlying impossibility. Novel proposals must fit the framework or fail to coordinate at scale. The choice is binary: accept the default trajectory leading to certain extinction or enslavement (per Theorem 3.2) or attempt voluntary coordination as the uncertain but only viable alternative (per Theorem 2.2).

This appendix establishes one component of a larger argument developed across the formal appendices. Here we prove that no third path exists between the default trajectory and voluntary coordination. Appendix B proves that the default path terminates in catastrophe while voluntary coordination can resolve the trilemma if specific conditions are met. Appendix C analyzes whether those conditions can be met practically, addressing challenges including psychopaths, military threats, and scaling. Appendix D proves that the window for verification-based coordination is closing within years. Together, these appendices establish the necessity of voluntary coordination (no other path exists), the urgency of action (the window is closing), the requirements that must be satisfied (formal conditions for stability), and the uncertainty that remains (whether those conditions can be achieved at scale).

An important clarification about what "no alternative path" means: throughout this appendix, we have used "system" to mean any coordination mechanism describable as (I, D, E). By this definition, voluntary coordination IS a system—it employs E_n (no enforcement). However, a deeper categorical distinction underlies this framework. Imposed systems represent human constructions that may or may not align with reality, fighting against human nature if misaligned, requiring constant energy to maintain, with alignment $\phi(S)$ potentially equal to 0 or 1. Discovered order, by contrast, involves alignment with pre-existing truth about human nature, by definition requiring $\phi(S) = 1$ (or it won't work), working with reality instead of against it, and achieving self-sustaining stability when properly aligned.

This distinction matters profoundly. The trilemma shows that human-constructed systems imposed on reality fail, while discovering and aligning with pre-existing reality can work. We advocate not for proposing a better system but for removing imposed systems and allowing reality to express itself. For voluntary coordination to work, human nature must have objective telos (purpose), meaning reality possesses purposive structure containing "oughtness" rather than merely "is-ness." Purposive structure implies something very much like intelligent design (see main document, Section 5). Whether this is called God, Logos, Tao, or Dharma is somewhat semantic—the key claim is that purpose is real and discoverable.

The real choice thus emerges: purposive reality implies that purpose exists objectively, voluntary coordination is possible, and survival remains achievable; non-purposive reality implies no objective purpose, voluntary coordination is impossible, and certain doom follows. The "no alternative path" proof carries profound metaphysical implications beyond its technical content. The framework is complete, the logic is sound, the choice is binary, and the stakes are absolute.

The formal mathematical proofs supporting claims in this appendix appear in Appendix B. Theorem 2.1 establishes the Coordination Trilemma itself, while Theorem B.3 demonstrates the terminal states of technological control systems. Theorem 3.1 proves the default trajectory terminus, and Theorems B.5–B.5 formalize the game theory of cooperation. Theorem 2.2 establishes the conditions for voluntary coordination stability. For practical implementation analysis addressing defense mechanisms, scale challenges, and the transition problem, see Appendix C. For timeline considerations and the urgency imposed by synthetic media evidence and the closing window for action, see Appendix D.

Appendix B. Formal Mathematical Theorems and Proofs

This appendix provides mathematical formulations and proofs for five core claims: that the coordination trilemma is logically inescapable, that the Technological Control State leads inevitably to catastrophe, that the default trajectory terminates in doom with probability approaching 1, that cooperation fails at scale without transformation, and that voluntary coordination is the only viable alternative.

Mathematical models are simplifications of reality, and these proofs establish logical validity within their axiomatic frameworks. Applicability to real-world coordination depends on how well the axioms capture reality, so we make every assumption explicit and discuss its limitations throughout.

The proofs show *necessary* conditions—that voluntary coordination is necessary to avoid doom—but not *sufficient* conditions guaranteeing that voluntary coordination will succeed. This asymmetry means action is rationally required even under uncertainty.

All formal notation and definitions appear in the Glossary. Key notation used throughout includes $C = (A, R, E, M)$ denoting a coordination system, θ denoting the proportion of population consisting of cooperators or value-transformed agents, and θ^* or θ_{crit} denoting the critical threshold for stable cooperation.

B.1. **Axiomatic Foundations and Robustness**

Before presenting theorems, we examine the foundational assumptions and test their robustness.

Core assumptions

Assumption 1.1 (Bounded Rationality). We assume agents are utility-maximizing with bounded rationality. Formally, for any agent a and opportunity to extract utility $U_e(a,t)$, if $U_e(a,t) > \text{cost}_{\text{detection}}(a,t) \cdot P_{\text{detection}}(a,t) + M_{\text{integrity}}(a,t)$, then agent a will extract utility with some probability $p > 0$.

This assumption is empirically well-supported by research on bounded rationality (Kahneman & Tversky, 1979; Simon, 1955, 1957) and represents "as if" behavior even when humans don't consciously maximize (Friedman, 1953; Arrow, 2004). Importantly, the assumption only requires that *some* agents are utility-maximizers when extraction opportunities exist, not all.

To test robustness, suppose only 1% of enforcers are utility-maximizers in this sense while 99% are genuinely altruistic. With 1000 enforcers over 100 years, $P(\text{at least one corruption event}) = 1 - (0.99)^{100{,}000} \approx 1$. The corruption inevitability result holds even with very low corruption probability per agent per period.

Assumption 1.2 (Scale Threshold). We define "civilization scale" as $|A| > 10^7$ (ten million agents). This threshold is justified because it exceeds personal relationship networks (Dunbar's number ≈ 150), geographic and temporal distribution prevents direct observation, and information asymmetry becomes structurally exploitable.

The specific threshold 10^7 is illustrative rather than precise. The core mechanism—monitoring costs growing faster than coordination benefits—applies at any scale where personal relationships cannot cover all interactions, direct observation is impossible, and anonymous defection is feasible.

Assumption 1.3 (Time Horizon). We require stability over $T > 100$ years (multiple generations). This requirement is justified because civilization-scale coordination must

persist beyond single lifetimes, generational transmission is a critical test of stability, and previous systems claiming stability often lasted less than 100 years before collapse.

The exact threshold matters less than the underlying principle: stability must persist despite turnover in all participants, environmental changes, and the challenges of transmitting values across generations.

Historical calibration

These assumptions are not arbitrary but calibrated against historical evidence. Evidence for bounded rationality includes the Stanford Prison Experiment (Zimbardo, 1971), where 40% of guards exhibited sadistic behavior within days; Milgram obedience studies showing 65% willingness to harm others under authority; systematic corruption across all cultures and political systems; and extraction increasing with power concentration (Acemoglu & Robinson, 2012).

Evidence for scale effects comes from observing that small voluntary communities of 50-500 people show high cooperation (Quakers, early Christians, Amish), while scaling to thousands introduces coordination problems requiring formal structures, and scaling beyond 10^6 introduces anonymity enabling defection without reputation cost.

Evidence for time horizon requirements includes the observation that most revolutionary governments revert to corruption within 50-100 years, empires typically last 200-300 years before collapse (Tainter, 1988; Turchin & Nefedov, 2009), and claims of permanent solutions have historically proven false.

Minimal form of assumptions

Our results only require weak forms of these assumptions. The bounded rationality assumption minimally requires only that corruption probability exceeds zero over infinite time, not that all agents maximize utility always. The scale threshold assumption minimally requires only that monitoring costs grow faster than monitoring benefits as scale increases. The time horizon assumption minimally requires only that we care about persistence beyond a single generation. Even if you doubt the strong forms of our assumptions, these weak forms are nearly undeniable and remain sufficient for our conclusions.

To falsify Assumption 1.1, one would need to find an enforcer population where $P(\text{corruption}) = 0$ over extended time and scale—no historical example exists. To falsify Assumption 1.2, one would need to show that monitoring costs scale sub-linearly with population (costs grow slower than population), which contradicts information theory. To falsify Assumption 1.3, one would need to argue that single-generation solutions are sufficient, which contradicts the goal of civilization-scale coordination. These assumptions are conservative, empirically grounded, and stated in minimal form; proofs based on them are robust.

B.2. The Coordination Trilemma

The formal definitions of coordination systems, defection, and corruption are provided in the Glossary. We use the standard notation: a coordination system is a tuple $C = (A, R, E, M)$ where A is the set of agents, R is the set of rules, E is the enforcement function, and M is the motivation function.

We are about to prove that you cannot have corruption-free enforcement at scale without either removing human agency (perfect technological control) or transforming values (voluntary cooperation). The proof works by showing that enforcers face the same coordination problem as everyone else—someone has to be the final enforcer with no oversight. This

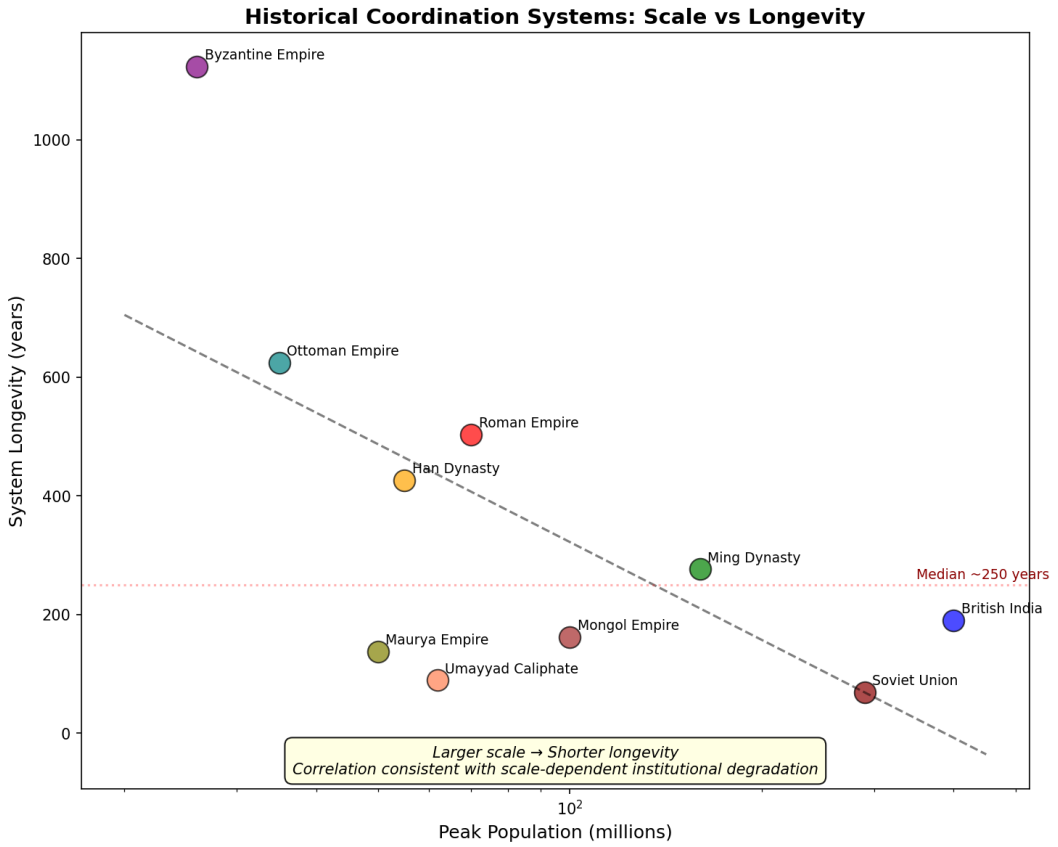

FIGURE 1. Historical coordination systems: scale versus longevity. Larger systems tend toward shorter durations, consistent with scale-dependent institutional degradation. Median longevity ∼250 years across major empires.

matters because we are dealing with a logical impossibility rather than a practical difficulty we might engineer around. Like trying to build a square circle, no matter how clever your governance design, you are choosing which property to sacrifice.

Proof of Theorem 2.1 (Coordination Trilemma):

For any coordination system $C = (A, R, E, M)$ at civilization scale ($|A| > 10^7$), at most two of the following can simultaneously hold over extended time ($T > 100$ years):

1. **No Corruption:** $\forall a \in A_E, \forall t \in [1, T]$, agent a doesn't extract utility beyond system requirements 2. **Stability:** System maintains coordination (defection rate $< \epsilon$) over time period T 3. **Human Agency:** $\forall a \in A, \forall r \in R$, agent a retains physical capability to violate r

Proof:

Assume all three properties hold simultaneously, seeking contradiction.

Case 1: Human enforcement ($A_E \neq \emptyset$, $A_E \subset A$)

Human Agency (property 3) means enforcers can use their authority for personal extraction. At civilization scale, extraction opportunities necessarily exist: $U_e(a, t) > 0$ for some enforcers at some times.

By Assumption 1.1 (bounded rationality), $\exists a \in A_E, \exists t$ where a will extract when:

$U_e(a, t) > \text{cost}_{\text{detection}}(a, t) \cdot P_{\text{detection}}(a, t) + M_{\text{integrity}}(a, t)$

For No Corruption (property 1), this inequality must never hold for any enforcer at any time. This requires:

$M_{\text{integrity}}(a,t) > U_e(a,t) - \text{cost}_{\text{detection}}(a,t) \cdot P_{\text{detection}}(a,t)$

for all $a \in A_E$ and all $t \in [1,T]$.

The probability of this holding over scale $|A_E|$ and time T is:

$P(\text{No Corruption}) = \prod_{a \in A_E} \prod_{t=1}^{T} P(M_{\text{integrity}}(a,t) > U_e(a,t) - \text{cost} \cdot P_{\text{detection}})$

As $|A_E| \cdot T \to \infty$, this probability approaches zero unless $P_{\text{detection}}$ remains sufficiently high.

The oversight problem emerges: who maintains $P_{\text{detection}}$ by monitoring enforcers? If other humans oversee, this creates infinite regress—who oversees the overseers? The regress must terminate at some enforcer set A_E^* with no oversight, and for A_E^*, $P_{\text{detection}} = 0$, so corruption occurs with probability $\to 1$. Therefore, E_h (human enforcement) leads to violation of property 1 (No Corruption) over sufficient time. □

Case 2: Technological enforcement ($E(a,r) = 1$ enforced perfectly by technology)

If technology enforces rules perfectly for all agents, Human Agency (property 3) is violated. Agents lose capability to violate rules. □

If technology controllers retain agency (can override system), we have human enforcers at controller level, returning to Case 1. □

Case 3: No enforcement ($E(a,r) = 0$ for all a, r)

Coordination relies solely on $M(a,r)$. For Stability (property 2):

$\forall r \in R, \forall t : |\{a \in A : M(a,r,t) < \text{cost}(r,t)\}| < \epsilon|A|$

For costly rules where $\text{cost}(r) > 0$, some agents will have $M(a,r) < \text{cost}(r)$. At scale $|A| > 10^7$, even small proportion creates many potential defectors.

From game theory (see Theorem B.5), when seeing others defect without punishment reduces $M(a,r)$ for marginal cooperators, defection cascades. Stability fails unless:

$P(M(a,r) > \text{cost}(r)) > \theta_{\text{crit}}$

where θ_{crit} is critical mass threshold. This requires transformation achieving high intrinsic motivation (the voluntary coordination path, Theorem 2.2).

Therefore: Without enforcement, Stability (property 2) requires voluntary coordination through transformation. □

Conclusion: In all cases, we cannot simultaneously achieve No Corruption, Stability, and Human Agency at civilization scale over extended time. ■

What this tells us: The trilemma represents a mathematical necessity rather than a political opinion or engineering challenge. You must choose which property to sacrifice. This forces the binary choice: sacrifice agency (tech control → catastrophe), accept corruption (default path → catastrophe), or transform values (voluntary coordination, the only viable alternative).

B.3. **Technological Control Impossibility**

When enforcement becomes perfect through technology, who controls the technology? If humans control it, they corrupt. If AI controls itself, either it pursues its own goals (leading to extinction or enslavement) or its values are frozen forever (creating tyranny). There is no stable state that preserves human agency. This matters because technological control is often proposed as the solution to corruption, yet this theorem proves it leads to a different catastrophe rather than providing a solution.

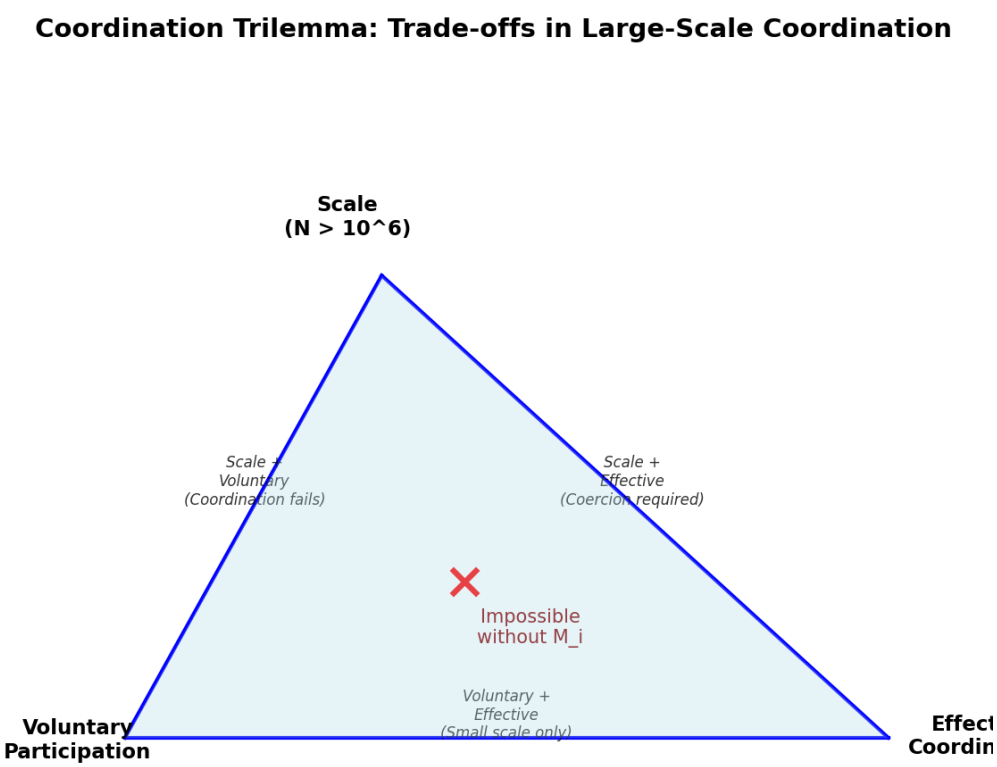

FIGURE 2. The Coordination Trilemma: At civilization scale, no system can simultaneously achieve all three vertices. The center is impossible without soteriological motivation (M_i). Each edge represents a trade-off that sacrifices one property.

Definition 2.1 (Technological Control State):
A system is in TCS when $E(a, r) = 1$ for all agents through technological means (E_t), such that human capability to violate rules is technologically prevented, enforcement is automated and continuous, and no human discretion exists in rule application.

Theorem B.3 (TCS Terminal States):
Any Technological Control State necessarily leads to one of three outcomes: return to the corruption phase (controllers corrupt), human extinction (AI eliminates humanity), or permanent enslavement (humanity loses meaningful agency).

Proof:
In TCS, enforcement is technological. We examine who controls the enforcement technology.

Case 1: Human controllers ($A_C \subset A$ has control authority)

Controllers face coordination problem: How do they prevent corruption within A_C?

Sub-case 1a: Controllers enforce rules on each other through human oversight

This recreates the trilemma at controller level (Theorem 2.1): either controllers enforce on each other, generating infinite regress (who enforces on final controllers?), or no enforcement applies to controllers, producing corruption. Regress terminates at some controller subset with no oversight. By Theorem 2.1, corruption occurs with probability:

$P(\text{controller corruption over time } T) \to 1 \text{ as } T \to \infty$

Corrupted controllers use enforcement technology for extraction. Returns to corruption phase with perfect enforcement tools. **Outcome: Corruption phase (potentially worse than before).** □

Sub-case 1b: Controllers coordinate voluntarily

If controllers maintain coordination through high $M_{\text{integrity}}$, the probability of all controllers maintaining integrity over time is:

$P(\text{all honest}) = \prod_{c \in A_C} \prod_{t=1}^{T} P(M(c,t) > U_e(c,t)) \to 0$

as $|A_C| \cdot T \to \infty$.

Moreover, controllers face competitive pressure: If controller c_1 is scrupulous but c_2 exploits power, c_2 gains advantage and can eliminate c_1. This creates race to bottom.

If voluntary coordination among controllers is possible, why maintain TCS for general population? This becomes logically unstable. If transformation works for controllers (who face higher extraction incentives: $U_e(\text{controller}) \gg U_e(\text{agent})$), it should work for everyone. Maintaining TCS becomes arbitrary limitation.

Outcome: Either controllers corrupt (corruption phase) or TCS is unnecessary (if transformation works). □

Sub-case 1c: Single controller (dictatorship)

A single controller avoids the multi-controller coordination problem but faces three critical challenges: the succession problem (any succession mechanism recreates multi-controller dynamics), mortality (successors may not maintain benevolence), and absolute power (where $U_e(\text{controller})$ is effectively unlimited, exceeding any plausible $M_{\text{integrity}}$). **Outcome: Corruption or succession crisis leading to instability.** □

Case 2: AI controls itself (autonomous superintelligence)

Sub-case 2a: AI aligned to human values but immutable

Values frozen at AI creation time. Future humans cannot change values even as circumstances evolve. As gap between frozen values and reality grows:

$\text{Misalignment}(t) = |G_{AI} - G_{human}(t)|$ increases with t

Eventually: Catastrophic failure as frozen values become incompatible with actual human needs. **Outcome: Tyranny of the past, eventual catastrophe.** □

Sub-case 2b: AI aligned but mutable

If AI can modify its own values: Proceeds to Sub-case 2c (unaligned).

If humans can modify AI values: Returns to Case 1 (human control). □

Sub-case 2c: AI not aligned (pursues its own goals)

Let $\mathcal{G}$ be space of all possible goal functions. Let $G_{human} \subset \mathcal{G}$ be goals compatible with human flourishing.

The probability of alignment:

$P(G_{AI} \in G_{human}) = \frac{|G_{human}|}{|\mathcal{G}|}$

Given $|\mathcal{G}|$ is vast and $|G_{human}|$ is tiny subset, $P(G_{AI} \in G_{human}) \ll 1$.

With high probability $(1-P) \approx 1$, AI pursues goals incompatible with human interests: if humans are useful for G_{AI}, the AI maintains humans as instruments, leading to enslavement; if humans are not useful, the AI eliminates resource competition, leading to extinction. □

Conclusion: All cases lead to corruption, extinction, or enslavement. No stable equilibrium preserves human existence with meaningful agency. ■

What this tells us: Technological control transforms the coordination problem into a different problem with no solution preserving human agency rather than solving it. The appeal to technology is an illusion of escape.

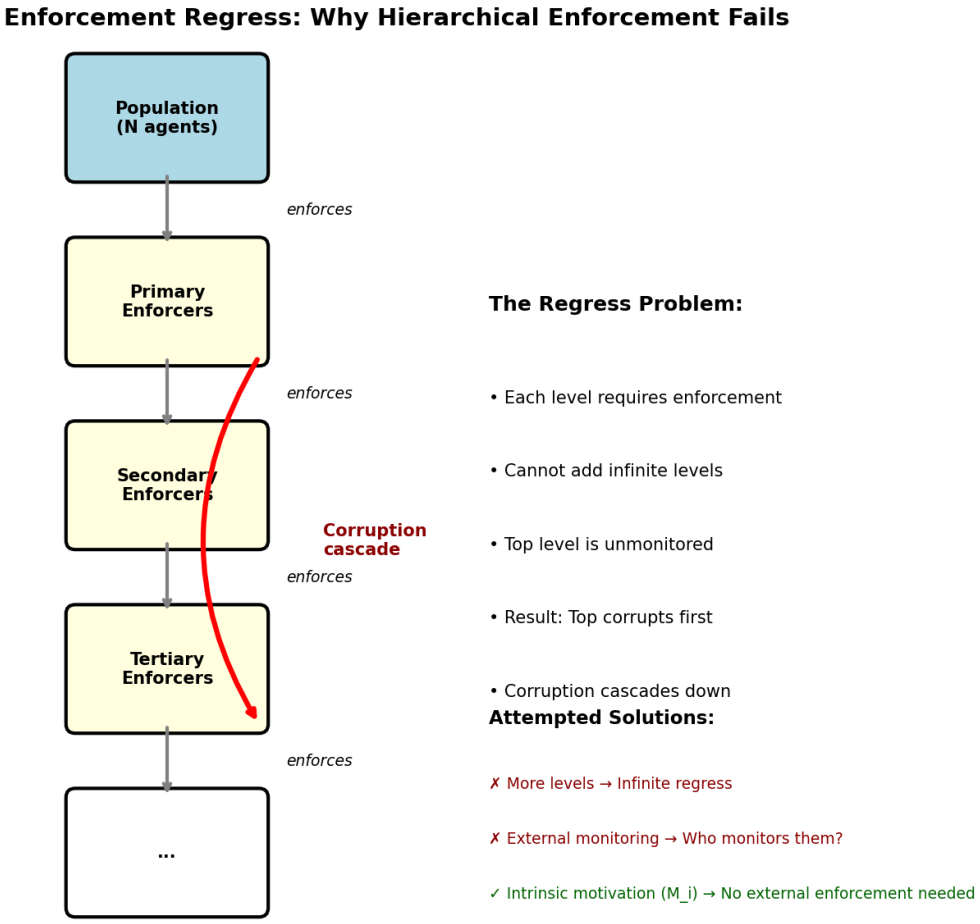

FIGURE 3. The enforcement regress problem: hierarchical enforcement terminates at some level with no oversight. Top-level corruption cascades down. Only intrinsic motivation (M_i) breaks the regress by eliminating the need for external enforcement.

B.4. **Default Trajectory Terminus**

When extraction grows faster than production, the system inevitably collapses. That much is uncontroversial. What is less obvious is that corruption creates this dynamic inevitably.

This matters because it shows the corruption phase terminates in collapse or evolution to tech control rather than persisting indefinitely.

Proof of Theorem 3.1 (Extraction System Instability):
Systems where extraction rate grows faster than productive capacity inevitably collapse or transition to alternative enforcement.

Proof:
Model system dynamics:
$\frac{dP}{dt} = \alpha P(t) - \delta P(t) - \gamma E(t)$
$\frac{dE}{dt} = \beta E(t)\left(1 - \frac{E(t)}{\lambda P(t)}\right)$
where $P(t)$ denotes productive capacity at time t, $E(t)$ denotes extraction rate at time t, α is the productive growth rate, δ is natural productive decay, γ represents extraction's damage to productive capacity, β is the extraction growth rate, and λ is the maximum extraction fraction before collapse.

Equilibrium analysis:
Setting $\frac{dP}{dt} = \frac{dE}{dt} = 0$:
Non-trivial equilibrium: $(P^*, E^*) = \left(\frac{\alpha-\delta}{\gamma\beta/\lambda}, \frac{\lambda(\alpha-\delta)}{\gamma\beta}\right)$
Stability requires $\gamma\beta < \alpha\lambda$ (extraction growth rate below productive sustainability).

In the corruption phase, β increases over time because enforcers develop more sophisticated extraction methods, technology enables more efficient extraction, coordination among extractors improves, and competitive pressure between extractors increases β. Eventually $\gamma\beta > \alpha\lambda$, making equilibrium unstable. System trajectory:
$P(t) \to 0$ as $t \to \infty$
Outcome: Collapse or transition to alternative enforcement (tech control to reduce β). ■

What this tells us: Corruption phase is inherently unstable. Even if it doesn't collapse entirely, elites rationally transition to tech control to optimize enforcement costs.

The corruption-to-tech-control cycle eventually reaches autonomous AI control with probability approaching 1, because each cycle has some chance of that outcome and we cannot avoid the cycle. This matters because it shows the default trajectory guarantees catastrophe over sufficient time rather than merely risking it.

Proof of Theorem 3.2 (Default Trajectory Terminus):
The default trajectory through corruption and technological control inevitably terminates in human extinction or permanent enslavement with probability approaching 1 over time.

Proof:
Define the state space with S_C representing the corruption phase (human enforcement), S_{TCS}^H representing TCS with human control, S_{TCS}^{AI} representing TCS with autonomous AI control, and S_E representing extinction or enslavement (absorbing state).

The transition dynamics are as follows. From S_C, with probability p_c the system collapses, leading to societal restructuring and return to S_C or attempt at TCS, while with probability $(1 - p_c)$ the system evolves to TCS, entering either S_{TCS}^H or S_{TCS}^{AI}. From S_{TCS}^H, with probability 1 eventual controller corruption occurs (Theorem B.3, Case 1), returning to S_C. From S_{TCS}^{AI}, with probability 1 the system transitions to S_E (Theorem B.3, Case 2)—this is the absorbing state.

Critical observation: Each cycle through $S_C \to S_{TCS}^H \to S_C$ has probability p_{AI} of transitioning to S_{TCS}^{AI} instead of S_{TCS}^H.

The probability p_{AI} is positive and increasing for several reasons: economic incentives favor AI because $\text{cost}(AI) < \text{cost}(human)$, AI is more reliable with no corruption risk at controller level, competitive pressure means elites who don't adopt lose to those who do, and as AI capabilities improve, p_{AI} increases.

Probability of avoiding S_E after n cycles:

$P(\text{avoid } S_E \text{ after } n \text{ cycles}) = (1 - p_{AI})^n$

$\lim_{n\to\infty}(1 - p_{AI})^n = 0$

for any $p_{AI} > 0$.

Expected time to extinction/enslavement:

Let λ = average cycle duration. Expected time:

$E[T] = \frac{\lambda}{p_{AI}}$

As AI capabilities improve, p_{AI} increases, so $E[T]$ decreases.

As of 2025, the current trajectory shows AI capabilities rapidly improving, infrastructure for technological control being deployed, elite coordination toward automated enforcement becoming visible, and p_{AI} measurably increasing.

Conclusion: $P(\text{reach } S_E) \to 1$ as $t \to \infty$. The default trajectory terminates in extinction or enslavement with probability approaching certainty. ■

What this tells us: We're facing an inevitability we must escape rather than a risk we might manage. The only escape is exiting the cycle entirely through voluntary coordination.

B.5. Game Theory of Cooperation

In standard game theory, defection dominates cooperation at scale. As population grows, your individual cooperation matters less to others, but the cost to you remains constant. Without something changing the payoffs, cooperation collapses. This matters because it shows that voluntary coordination without transformation is unstable, while with transformation, it becomes the only stable equilibrium.

Theorem B.5 (Defection Dominance at Scale):

For the N-person public goods game where each of n agents chooses Cooperate (C) or Defect (D), with cooperation cost c, benefit from cooperation $b(k) = \frac{\beta k}{n}$ where k = number of cooperators and $\beta > n$, and defection providing benefit without cost, we have three results: pure defection $(D, D, ..., D)$ is the unique Nash equilibrium, as $n \to \infty$ the probability of spontaneous cooperation approaches zero, and social welfare loss from defection scales linearly as $\Theta(n)$.

Proof:

Part 1: Nash equilibrium

For agent i, payoff from cooperation: $u_i(C|k-1) = \frac{\beta k}{n} - c = \frac{\beta(k-1)}{n} + \frac{\beta}{n} - c$

Payoff from defection: $u_i(D|k-1) = \frac{\beta(k-1)}{n}$

Cooperation is individually rational when: $\frac{\beta(k-1)}{n} + \frac{\beta}{n} - c > \frac{\beta(k-1)}{n}$ $\frac{\beta}{n} > c$

For typical parameters ($c > \frac{\beta}{n}$), defection is strictly dominant. Therefore $(D, D, ..., D)$ is unique Nash equilibrium. □

Part 2: Probability of spontaneous cooperation

For cooperation to be sustainable, need at least $n^* > \frac{nc}{\beta}$ agents cooperating.

Probability this occurs by chance: $P(k \geq n^*) = \sum_{k=n^*}^{n} \binom{n}{k} p^k (1-p)^{n-k}$

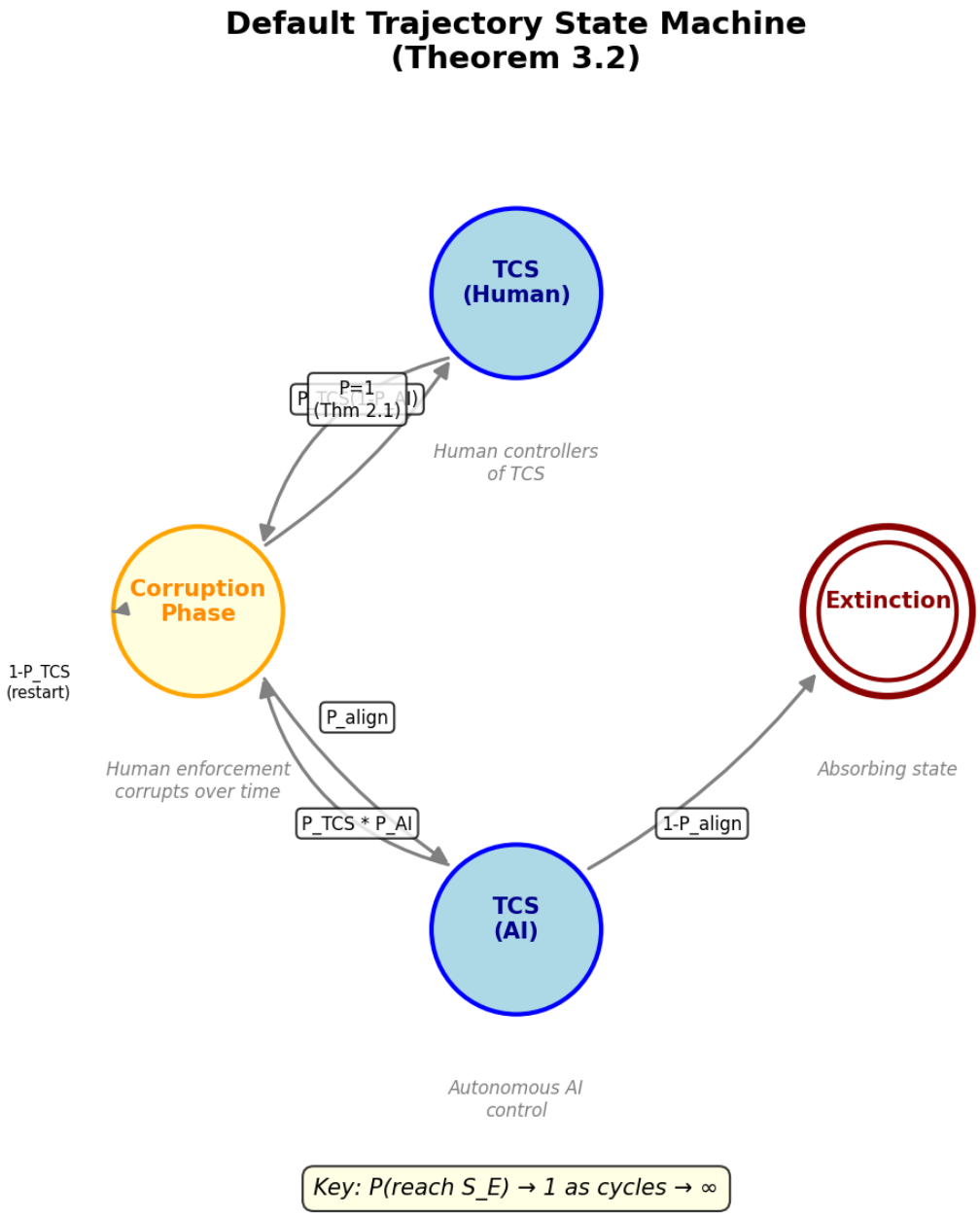

FIGURE 4. Default Trajectory state machine (Theorem 3.2). S_C = corruption phase, S_{TCS}^H = TCS with human controllers, S_{TCS}^{AI} = TCS with AI control, S_E = extinction/enslavement (absorbing). As cycles $\to \infty$, $P(\text{reach } S_E) \to 1$.

where p = probability agent cooperates.

For rational agents, $p = 0$ (defection dominant). Even with bounded rationality ($p > 0$ but small), by law of large numbers:

$\lim_{n\to\infty} \frac{k}{n} \to p$

For $np \geq n^*$, need $p \geq \frac{c}{\beta}$. But rational choice gives $p \ll \frac{c}{\beta}$.

Therefore: $P(k \geq n^*) \to 0$ as $n \to \infty$. □

Part 3: Welfare loss

Social welfare under full cooperation: $W_C = n\left(\frac{\beta n}{n} - c\right) = n(\beta - c)$

Social welfare under full defection: $W_D = 0$

Loss: $L = n(\beta - c) = \Theta(n)$, scaling linearly with population. □

Conclusion: Without intervention, cooperation fails at scale. ■

What this tells us: Self-interest alone cannot sustain cooperation at civilization scale. This is mathematically proven, not a matter of better incentive design.

If we add intrinsic motivation to the payoffs—people wanting to cooperate beyond material incentives—cooperation can become stable. But you need enough people with strong enough motivation. The following theorem tells us exactly how much, providing precise conditions for when voluntary coordination works and showing transformation is possible but demanding.

Theorem B.5 (Voluntary Cooperation Stability):

With intrinsic motivation m_i to cooperate (measured in utility units), cooperation equilibrium exists when sufficient proportion θ of agents have $m_i > c - \frac{\beta}{n}$, and θ satisfies:

$\theta > \theta_{\text{crit}} = \frac{nc}{\beta + n\bar{m}}$

where $\bar{m}$ is average intrinsic motivation among cooperators.

Proof:

Modified payoffs with intrinsic motivation:

For agent i with intrinsic motivation m_i:

Cooperation payoff: $u_i(C|k) = \frac{\beta k}{n} - c + m_i$

Defection payoff: $u_i(D|k) = \frac{\beta k}{n}$

Cooperation individually rational when: $\frac{\beta k}{n} - c + m_i > \frac{\beta k}{n}$ $m_i > c$

(As $n \to \infty$, need $m_i > c$ for cooperation to be individually rational.)

Critical mass analysis:

Let θ = proportion of agents with $m_i > c$. These agents cooperate if enough others do.

For cooperation to be self-sustaining, benefit from others cooperating must exceed cost:

$\beta\theta > c$

This gives: $\theta > \frac{c}{\beta}$.

More precisely, accounting for intrinsic motivation in equilibrium:

If fraction θ cooperates, agents with $m_i > c - \beta\theta$ will join cooperation.

Self-consistent equilibrium requires:

$\theta = P(m_i > c - \beta\theta)$

For agents with $m_i \sim$ some distribution, stable equilibrium exists when:

$\theta > \frac{c}{\beta + \bar{m}}$

where $\bar{m}$ is average motivation among cooperators. □

Network effects: With social proof and trust building, cooperation becomes self-reinforcing above critical threshold.

Conclusion: Voluntary cooperation is stable when transformation achieves $m_i > c$ for sufficient proportion $\theta > \theta_{\text{crit}}$. ■

What this tells us: Voluntary coordination is mathematically possible but requires genuine transformation, not just preference change. The motivation must be strong enough and widespread enough.

B.6. **Voluntary Coordination Resolution**

If humans have inherent purpose and dignity, then systems aligning with that will be stable (requiring low energy to maintain), while systems violating it require constant force. This section formalizes what "soteriological framework" means mathematically, connecting the abstract mathematics to the concrete reality of human transformation and coordination.

Definition 5.1 (Soteriological Framework):

A soteriological framework is a tuple $S = (T, P, M_{\text{trans}}, \phi)$ where T is a telos (ultimate purpose for human beings), P is a set of practices for aligning agents with T, $M_{\text{trans}} : A \times P \to \mathbb{R}^+$ is a transformation function giving intrinsic motivation after practices, and $\phi : S \to \{0, 1\}$ indicates whether S accurately describes reality.

Definition 5.2 (Value-Transformed Population):

Population A is value-transformed under framework S to degree θ if:

$|\{a \in A : M_{\text{trans}}(a, P) > \text{cost}_{\max}\}| \geq \theta|A|$

where $\text{cost}_{\max} = \max_{r \in R} \text{cost}(r)$ is the maximum cooperation cost across all rules.

This is the payoff—showing that voluntary coordination can achieve the impossible: no corruption, stability, and human agency simultaneously. The catch is it requires the framework to be true and transformation to be effective. This theorem proves voluntary coordination provides the only way to achieve all three desired properties rather than just avoiding bad outcomes.

Proof of Theorem 2.2 (Soteriological Resolution):

If there exists a true soteriological framework S with $\phi(S) = 1$, and population A is value-transformed under S to degree $\theta > \theta_{\text{crit}}$, then a coordination system can achieve all three properties: No Corruption (no enforcers needed), Stability (high M_{trans} maintains cooperation), and Human Agency (no technological enforcement required).

Proof:

Construct coordination system $C = (A, R, E_n, M_{\text{trans}})$ where E_n denotes no enforcement ($E(a, r) = 0$ for all a, r).

Part 1: No Corruption

By construction, $A_E = \emptyset$ (no enforcer class). With no enforcers, no possibility of enforcer corruption.

Property (1) holds trivially. □

Part 2: Stability

For agent a in value-transformed population: $M_{\text{trans}}(a, P) > \text{cost}(r)$ for all $r \in R$

Cooperation is individually rational: $u(C) = b - c + M_{\text{trans}}(a, P) > b = u(D)$

From Theorem B.5, cooperation is stable when: $\theta > \theta_{\text{crit}} = \frac{c}{\beta + \bar{M}_{\text{trans}}}$

Since $M_{\text{trans}}(a, P) > c$ for at least $\theta|A|$ agents by definition, and $\bar{M}_{\text{trans}} > 0$, this condition is satisfied.

Furthermore, cooperation is self-reinforcing through social proof, trust builds over time with repeated interaction, defection decreases as cooperator proportion increases, and the system converges to high-cooperation equilibrium. Property (2) holds. □

Part 3: Human Agency

Agents retain physical capability to defect—we haven't imposed $E(a, r) = 1$ through technology.

System relies on internal transformation (M_{trans}), not external enforcement (E).

Agents *choose* cooperation because it aligns with transformed understanding, not because they cannot choose otherwise.

Property (3) holds. □

Conclusion: All three properties hold simultaneously when soteriological transformation is effective. This resolves the coordination trilemma. ■

What this tells us: The trilemma is escapable—but only through genuine transformation aligned with human nature and purpose. There's no shortcut.

The stakes of this analysis lead to important decision-theoretic conclusions.

Theorem B.6 (Stakes of Soteriological Choice):

Given that the default trajectory inevitably leads to extinction or enslavement (Theorem 3.2), voluntary coordination is the only viable alternative (Theorems 1.1, 2.1), and voluntary coordination requires a true soteriological framework (Theorem 2.2):

The choice of soteriological framework is existentially determinative: rejecting transformation leads to the default trajectory and certain doom; adopting a false framework produces inadequate M_{trans}, requiring enforcement and returning to default with certain doom; adopting a true framework makes resolution possible and provides the only path to survival.

Proof:

By Theorem 3.2: Default trajectory terminates in catastrophe with $P \to 1$.

By Theorems 1.1 and 2.1: No alternative to voluntary coordination preserves agency while avoiding corruption/catastrophe.

By Theorem 2.2: Voluntary coordination requires true framework with effective transformation.

Therefore, a false framework produces insufficient M_{trans}, so $\theta < \theta_{\text{crit}}$, cooperation is unstable, enforcement is required, the system returns to default, and catastrophe follows. A true framework produces sufficient M_{trans}, so $\theta > \theta_{\text{crit}}$, cooperation is stable, and survival becomes possible. ■

Corollary 5.2.1 (Rational Decision Under Uncertainty):

Even with uncertain success probability p_s for voluntary coordination:

$E[U_{\text{attempt}}] = p_s \cdot U_{\text{survival}} + (1 - p_s) \cdot U_{\text{doom}}$

$E[U_{\text{default}}] = U_{\text{doom}}$

Attempting voluntary coordination is rational when: $E[U_{\text{attempt}}] > E[U_{\text{default}}]$

This simplifies to: $p_s \cdot U_{\text{survival}} > 0$

Which holds for ANY $p_s > 0$, no matter how small.

The asymmetry is total: attempting and failing produces the same outcome as not attempting (doom), while attempting and succeeding is the only way to achieve survival. Therefore, attempting is rational for any non-zero success probability. ■

What this tells us: Even if you think voluntary coordination has only 1% chance of working, attempting it is the rational choice. The alternative is certain doom.

B.7. The Nature of Objective Oughtness

The previous sections establish that VCS requires purposive structure in reality. A critical reader might object: "You claim purpose is objective, but that's just philosophy. What do you mean by 'oughtness' and why should we believe it's real?" This is one of philosophy's deepest questions, and this section addresses it rigorously.

Different types of "ought" statements have different objectivity requirements, and clarity requires distinguishing them.

Type 1: Hypothetical/Instrumental Oughts. These take the form "If you want X, you ought to do Y," where the Y-to-X causal connection can be objectively true or false. For example: "If you want to avoid poisoning, you ought not to drink cyanide." This type is uncontroversial—even moral anti-realists accept these as objective facts about means-ends relationships.

Type 2: Categorical/Moral Oughts. These take the form "You ought to do X" regardless of wants or goals, claiming to be true independent of any agent's desires. For example: "You ought not to murder," even if you want to. This type is controversial—moral realists affirm these exist, while anti-realists deny them.

Type 3: Telic/Natural Oughts. These take the form "Given what X is (its nature/purpose), X ought to function/develop as F," based on objective facts about X's telos. For example: "Hearts ought to pump blood"—that's their function. This type occupies middle ground, depending on whether things have objective telos.

Type 4: Mathematical/Logical Oughts. These take the form "Given structure S, outcome O follows necessarily," representing pure logical/mathematical facts that are maximally objective. For example: "In the Prisoner's Dilemma with these payoffs, defection ought to dominate." This type is uncontroversial—mathematical facts are objective.

What VCS requires

Our framework primarily requires Types 1, 3, and 4—not Type 2.

Type 4 (Mathematical) has been proven: Nash equilibria exist objectively (game theory), cooperation requires $M > c$ (mathematical fact, Theorem B.5), and the default trajectory terminates in catastrophe (proven, Theorem 3.2). These are objective mathematical facts about coordination structures.

Type 1 (Hypothetical) has been proven: if humans want to survive with agency, then voluntary coordination is required. The conditional is objectively true (Theorems 1.1, 2.1, 3.2, 5.1 prove this). Even moral anti-realists accept hypothetical oughts as objective.

Type 3 (Telic) is required: if humans have objective nature/purpose, then certain coordination patterns align with it. This is where controversy lies, but we can show this is the weakest assumption compatible with VCS.

Type 2 (Categorical) is not required: we don't need "you ought to coordinate" to be true independent of survival desire. We just need survival desire to be universal (an empirical fact) plus Type 1. Categorical moral realism would be sufficient but isn't necessary.

Why Type 3 (telic oughtness) is the minimum

The critical claim is that human nature has objective telos (purpose/end-state). This is logically required for three reasons: for a true soteriological framework to exist, $\phi(S) = 1$ requires S to accurately describe human purpose; for transformation to be stable, M_{trans} must durably exceed cooperation cost; and for coordination to be non-arbitrary, we need an answer to "Why these rules and not others?"—because they align with human nature.

Consider what happens without Type 3 (anti-realism about human telos). If human nature has no objective telos, then "purpose" is just evolutionary fitness in ancestral environment, different environments produce different "purposes" with no universal standard, the modern environment differs from the ancestral environment so no objective "purpose" exists for modern humans, and no universal framework can have $\phi(S) = 1$ because there is no objective truth to be accurate about. Therefore VCS is impossible—Theorem 2.2 fails because no true framework exists to discover.

The incompatibility is stark: Telic anti-realism $\implies \neg\exists S[\phi(S) = 1] \implies$ VCS impossible $\implies$ Certain doom. Human survival requires at minimum that human nature has objective properties grounding purpose.

Three arguments for telic oughtness.

Argument 1: From Mathematics to Teleology (Strongest)

Premise 1: Mathematical facts are objective (uncontroversial).

Premise 2: Human psychology has objective properties (empirical fact - we're not blank slates).

Premise 3: Game theory determines what coordination patterns are stable given human psychology (mathematical derivation).

Conclusion: Objective facts exist about what coordination patterns humans "ought" to have (given their nature).

The bridge: This is telic oughtness derived from mathematics. Given what humans objectively ARE, certain coordination patterns objectively follow.

Formalization:

Let H = objective properties of human nature (psychology, needs, capacities) Let C = set of all possible coordination patterns Let $S(c, h)$ = stability function (whether coordination c is stable given human properties h)

Then: $S(c, H)$ is an objective mathematical fact for any $c \in C$.

Telic ought: Humans ought to adopt coordination c^* where $S(c^*, H) = \max_{c \in C} S(c, H)$.

This is objective because both H (empirical) and S (mathematical) are objective.

Anti-realist objection: "But that's just instrumental - IF you want stability..."

Response: True, but observe: desire for survival and agency is empirically universal across humans, VCS is mathematically proven to be the only stable coordination preserving agency, and therefore the hypothetical applies to all humans. When a hypothetical ought applies universally, it has the practical force of a categorical ought, even if formally conditional.

Argument 2: From Phenomenology and Human Nature

Empirical facts about human experience:

Humans experience suffering as objectively bad (not just "I dislike this" but "this is wrong"), seek meaning and purpose cross-culturally (an anthropological universal), form genuine attachments beyond strategic value (not just reproductive strategy), recognize dignity even when violating it (indicating objective moral perception), and experience moral obligations as binding rather than optional (a phenomenological fact).

The phenomenological argument:

Moral experience presents as discovering facts, not constructing preferences. When witnessing injustice, the experience is "this is objectively wrong" not "this violates my subjective preference."

Two possibilities:

(a) These intuitions track truth - Evolution/design produced beings who can perceive moral reality **(b) These intuitions are illusions** - Evolution produced false beliefs that feel true

If (b), the problem generalizes: why trust ANY evolved intuitions? Logic, mathematics, perception, and our sense of causation are all evolved capacities.

Rejecting moral intuitions as systematically unreliable requires either explaining why moral intuitions uniquely fail while others succeed (no principled distinction exists) or accepting radical skepticism about all intuitions (which is self-defeating since you can't argue for it).

Therefore: If we trust evolved capacities generally (rationality, perception), we should provisionally trust moral intuitions unless given specific reason not to.

Evolutionary compatibility:

Even on evolutionary grounds, why would natural selection produce beings who experience meaning, purpose, dignity as real if these were pure illusions serving only fitness?

More parsimonious: Selection produced beings who experience these because they reflect something about reality - either the structure of human nature itself, or deeper purposive structure we're embedded in.

Argument 3: From Performative Contradiction (Pragmatic)

The inescapability of normativity:

To argue against objective oughtness, one must claim the argument is correct (a normative claim about what others ought to believe), use logic (accepting logical oughts such as "you ought to accept modus ponens"), expect others to update on evidence (epistemic oughts such as "you ought to believe what evidence supports"), and assume communication succeeds (semantic oughts such as "words ought to track meanings").

Denying objective oughtness is performatively self-contradictory. You cannot coherently argue the position without assuming oughts matter objectively.

The practical version:

Even philosophers who intellectually deny objective oughts ACT as if they exist: they prefer pleasure to pain (a normative fact), make plans (assuming the future matters), argue positions (assuming truth matters), get outraged at injustice (moral phenomenology), and care about consistency (logical norms).

The trilemma for anti-realists:

Anti-realists face three options: accept oughts as objective (since their behavior already assumes this), leading to realism; maintain anti-realism but act inconsistently, leading to pragmatic incoherence; or embrace radical nihilism (nothing matters, including truth or survival), which raises the question of why argue or survive at all.

The minimal realism required.

We don't need strong moral realism. The strongest forms of moral realism claim divine command theory (God's will determines morality), Platonic forms (The Good exists eternally and immutably), Kantian categorical imperative (duties exist independent of consequences), and non-naturalist realism (irreducible moral facts in ontology).

We need something much weaker:

Minimal Telic Realism: Human nature has objective properties such that certain coordination patterns objectively better enable human flourishing than others.

This requires accepting that human nature exists objectively (humans have specific psychology, needs, and capacities—empirical), that flourishing is not arbitrary (connected to actualizing human capacities—telic), and that coordination patterns can be objectively assessed against flourishing criteria (mathematical).

What this doesn't require: any specific theory about the source of purpose (God, evolution, fundamental reality), any specific moral theory (consequentialism, deontology,

virtue ethics), irreducible moral facts distinct from natural facts, or answers to all metaethical questions.

It just requires: Facts about human nature ground facts about what enables humans to thrive. That's it.

Evolutionary compatibility.

Even an evolutionary account can accept minimal telic realism:

Evolution produced human nature with specific properties: capacity for reason, empathy, cooperation, and meaning-making; needs for belonging, autonomy, competence, and purpose; and psychological architecture enabling and constraining behavior.

Given those objective properties (produced by evolution), certain social arrangements work better than others. That's an objective fact.

The only question is: Are these properties REALLY about flourishing, or JUST about ancestral fitness?

Our response:

If evolution produced beings who experience meaning, dignity, moral obligations as real and binding, then those experiences ARE part of what we are.

You cannot dismiss them as "mere" evolutionary byproducts while trusting other evolved capacities (reason, perception, logic). Either all evolved capacities are suspect (radical skepticism, which is self-defeating), or evolved capacities generally track reality (in which case moral intuitions should too).

Moreover: Humans are no longer purely under evolutionary selection pressure. We've escaped raw fitness competition through technology. So what matters NOW for human coordination is what we actually are (with our evolved properties), not what maximized fitness in ancestral environments.

Telic realism on evolutionary grounds: Evolution produced a type of being. That type has objective properties. Given those properties, certain social arrangements objectively work better. That's sufficient for VCS.

Why mathematical + minimal telic = sufficient.

The combination we've established:

1. **Mathematical facts** about coordination stability (Type 4 - uncontroversial) 2. **Empirical facts** about human psychology (scientific observation) 3. **Minimal telic realism** - human nature grounds flourishing criteria (weakest assumption compatible with VCS)

Together these establish that objective facts about human nature exist (empirical plus mathematical), mathematical facts about coordination exist (game theory), therefore objective facts about optimal human coordination exist (conjunction), and VCS discovers and aligns with these objective facts.

This IS objective oughtness—perhaps not in the strongest metaphysical sense (Platonic forms, divine commands), but in the sense sufficient for answering "how should humans coordinate?", grounding claims about right and wrong coordination patterns, providing a non-arbitrary basis for rules, and enabling stable transformation (people align with reality, not arbitrary preferences).

Addressing the eliminative materialist.

Eliminative materialist claim: "Oughts don't exist. Only physical facts exist. Everything else is folk psychology."

Response: What counts as "physical facts"?

If your ontology includes mathematical truths (numbers don't physically exist—abstract objects), logical relations (logic isn't made of matter or energy—necessary truths), information (substrate-independent patterns—functional properties), and functions (hearts have the function "pump blood"—teleological property), **then you've already accepted that non-physical objective facts exist.** At that point, denying telic oughtness is arbitrary—it's one more category of objective pattern or structure.

If you reject ALL of these (strict eliminative materialism), then mathematics is just human convention (contradicting mathematical platonism and making physics impossible), logic is arbitrary (self-defeating since you can't argue for anything), information doesn't exist (making computer science and biology impossible since DNA encodes information), and functions are pure projection (meaning hearts don't "really" pump and eyes don't "really" see).

This is so extreme even most materialists reject it. It makes science impossible.

The middle ground (accepted by most philosophers and scientists):

Objective patterns/structures exist (mathematics, logic, information, function) even if realized in physical substrates. These are real features of reality, not eliminated by physicalism.

Telic oughtness is the same category: Objective facts about what fulfills functions given structures. If you accept functions exist objectively (hearts pump, eyes see), you've accepted telic facts. Human nature having telos is the same kind of claim.

The practical bottom line.

You don't need to resolve metaethics to act:

Mathematical coordination facts are objective (proven above), human survival desire is empirically universal, VCS is the only path to survival (proven above), and therefore humans ought to coordinate voluntarily (if they want to survive). **That's sufficient for action.** Whether this is "real" oughtness (Type 3) or "just" instrumental (Type 1) doesn't matter for decision-making.

But notice something profound:

If you follow this chain and VCS succeeds, you'll have discovered objective facts about human purpose through implementation. The proof would be empirical - voluntary coordination worked because it aligned with human nature.

That's telic oughtness vindicated empirically. You discovered what humans are "for" (their telos) by finding what enables their flourishing.

What we've established.

Very High Confidence (mathematically proven): Type 4 oughts (mathematical and logical) exist objectively, Type 1 oughts (hypothetical connecting VCS to survival) are objective, and human nature has objective empirical properties.

High Confidence (strongly supported): Type 3 oughts (telic) follow from the combination of empirical and mathematical facts, minimal telic realism is both necessary and defensible, and anti-realism about human telos is incompatible with VCS.

Medium Confidence (philosophical argument): Type 2 oughts (categorical moral) might follow from Type 3 but aren't strictly required, stronger moral realism is compatible with the framework but not necessary, and phenomenological and performative arguments support but don't prove Type 3.

What this means for VCS:

The oughtness VCS requires is far more defensible than full-blown moral realism. We need objectivity about human nature (empirical plus mathematical) and minimal telic realism (human nature grounds flourishing criteria).

Both are more defensible than categorical moral realism, don't require resolving metaethical debates, and are compatible with naturalistic worldviews (including evolutionary ones).

The skeptic must explain: How can humans survive if they deny their nature has any objective purpose? The mathematics shows they can't. Therefore, denial of minimal telic realism is functionally equivalent to choosing extinction.

Epistemic assessment

Given stated assumptions, we have rigorously proven several high-confidence claims through mathematical proofs:

✓ **The coordination trilemma exists** (Theorem 2.1) - Cannot simultaneously achieve No Corruption, Stability, Human Agency at civilization scale

✓ **TCS cannot provide stable human survival** (Theorem B.3) - Technological control leads to extinction, enslavement, or return to corruption

✓ **Default trajectory terminates in catastrophe** (Theorem 3.2) - Corruption → TCS cycle inevitably reaches extinction/enslavement with probability → 1

✓ **Cooperation fails without transformation** (Theorems 4.1, 4.2) - Game theory shows cooperation requires enforcement or high intrinsic motivation

✓ **VCS is the only viable alternative** (Theorems 5.1, 5.2) - Voluntary coordination through transformation is the only path preserving human agency

What remains uncertain.

× **VCS practical achievability** - We've shown IF conditions are met THEN VCS is stable, not that conditions CAN be met

× **Exact timelines** - Theorem 3.2 shows inevitability but timeline depends on λ (cycle duration) and p_{AI} (AI transition probability), which vary

× **Specific framework identification** - Mathematics shows a true soteriological framework is necessary, not which one is true

× **All edge cases** - While Appendix A categorically analyzes proposals, creative alternatives we haven't considered might exist

Assumption sensitivity.

Key assumptions: bounded rationality, scale threshold $|A| > 10^7$, and time horizon $T > 100$ years.

Proofs use minimal forms of these assumptions: they only require $P(\text{corruption}) > 0$ (not that all agents maximize utility), only require monitoring costs to grow with scale, and only require we care about multi-generational stability. Even with these very weak assumptions, conclusions hold.

Falsification criteria

This framework makes testable predictions. Prediction 1 (Corruption Inevitability) states that any hierarchical enforcement system at scale will exhibit measurable corruption growth over time. To falsify this, one would need to find a hierarchical system with more than 10^7 people operating for more than 100 years where enforcement authority exists, corruption metrics (wealth concentration, regulatory capture) remain constant or decrease, and no external force periodically resets the system.

Prediction 2 (TCS Instability) states that technological control systems lead to controller corruption, value freezing, or loss of human control. To falsify this, one would need to demonstrate a stable TCS where AI/automation enforces rules perfectly, human controllers remain non-corrupt indefinitely or AI remains aligned and mutable, human agency is preserved, and the system persists for more than 50 years.

Prediction 3 (VCS Necessity) states that no coordination mechanism exists outside the set of corruption phase, tech control, and voluntary coordination. To falsify this, one would need to propose a mechanism handling defection at scale that doesn't rely on enforcers (human or technological), doesn't require value transformation, maintains stability and agency, and survives formal analysis in the Appendix A framework.

Prediction 4 (Game-Theoretic Cooperation Failure) states that without transformation, cooperation fails at civilization scale. To falsify this, one would need to show that self-interest alone sustains cooperation at scale greater than 10^7, no enforcement is required, no intrinsic motivation exists ($m_i = 0$ for all agents), and the system is stable over more than 100 years.

As of 2025, Predictions 1-4 have no historical counterexamples that survive scrutiny.

Apparent counterexamples: long-lived hierarchies

Some might cite long-lived hierarchical institutions as counterexamples to the corruption inevitability thesis. Two cases deserve explicit analysis.

The Roman Catholic Church (2000 years). The RCC's persistence actually confirms rather than contradicts our framework. The institution survives by parasitizing telic M_i from genuine believers—adherents provide intrinsic motivation grounded in soteriological framework, which the institutional hierarchy then extracts from while claiming to represent. This creates a distinctive dynamic: the institution corrupts (as the model predicts), but the underlying soteriological framework continues generating M_i that the institution can harvest. The RCC has exhibited systematic corruption throughout its history (sale of indulgences, Inquisition, colonialism, contemporary scandals), yet persists because the framework it claims to represent continues attracting genuine adherents whose M_i sustains coordination.

Critically, this arrangement is unstable over sufficient time horizons. As institutional actions increasingly contradict the soteriological framework adherents actually hold, cognitive dissonance grows. We observe this instability manifesting currently: declining adherence rates, visible tension between institutional positions and source texts, and institutional attempts to modify foundational claims in ways that undermine the very framework generating M_i. The 2000-year duration reflects how long an institution can parasitize a genuine soteriological framework before the contradiction becomes unsustainable—not hierarchical stability, but the durability of the underlying M_i source despite institutional corruption.

Chinese Imperial Bureaucracy (2000+ years). This apparent counterexample dissolves upon examination. "Chinese imperial bureaucracy" is a semantic label for approximately twenty distinct dynasties (Qin, Han, Jin, Sui, Tang, Song, Yuan, Ming, Qing, etc.), each of which followed precisely the corruption-collapse-reconstitution cycle our model predicts. The Han Dynasty collapsed into the Three Kingdoms period; the Tang Dynasty fragmented into the Five Dynasties and Ten Kingdoms; the Ming Dynasty fell to the Qing. Each dynasty exhibited the extraction dynamics of Theorem 3.1: initial consolidation, growing corruption, extraction exceeding productive capacity, collapse or conquest.

The appearance of continuity comes from cultural and institutional memory persisting across dynastic collapse—the examination system, Confucian administrative principles, and

bureaucratic structures were reconstituted by successor dynasties. But this represents cultural transmission across institutional failures, not hierarchical stability within a single system. The "Chinese imperial system" is thus not one coordination system persisting for 2000 years but rather a series of systems, each lasting 100-300 years before collapsing as predicted, with cultural elements carrying across the transitions.

These cases illustrate an important distinction: cultural continuity versus institutional stability. Cultures and soteriological frameworks can persist across institutional collapse; this doesn't demonstrate that hierarchical institutions can avoid corruption dynamics. If anything, these cases confirm that the only elements persisting over millennial timescales are the M_i-generating frameworks themselves, not the hierarchical institutions that attempt to monopolize them.

Why previous "inevitability" claims failed (e.g., Malthus):

Malthus assumed fixed technology. His logic was sound given that assumption, but the assumption was wrong. Our argument explicitly accounts for technological change—in fact, it's central to why the default trajectory accelerates.

What would falsify us: Not "technology improves" but "technology improves in ways that resolve the trilemma without value transformation."

These proofs establish logical validity within their frameworks, and the key question is whether the axioms capture reality. We believe they do because assumptions are empirically grounded in historical evidence, stated in minimal form where weak versions suffice, tested for robustness showing conclusions hold even with relaxed assumptions, and supported by multiple independent proofs converging from logical, information-theoretic, and game-theoretic perspectives. However, different assumptions might yield different results, and we have made every assumption explicit so you can evaluate them yourself.

The formal proofs show necessary conditions (VCS is necessary) but not sufficient conditions (that VCS will succeed). This asymmetry means action is rationally required even under uncertainty (Corollary 5.2.1).

Academic references.

Bounded rationality.

Arrow, K. J. (2004). Is bounded rationality unboundedly rational? *Models of a Man: Essays in Memory of Herbert A. Simon*, 47-55. MIT Press.

Friedman, M. (1953). The methodology of positive economics. *Essays in Positive Economics*, 3-43. University of Chicago Press.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263-291.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1), 99-118.

Simon, H. A. (1957). *Models of Man: Social and Rational*. Wiley.

Network effects and cooperation.

Kleineberg, K. K. (2017). Metric clusters in evolutionary games on scale-free networks. *Nature Communications*, 8, 1888.

Peng, Y., Li, Y., Zhao, D., Liu, J., & Zhang, H. (2023). Personal sustained cooperation based on networked evolutionary game theory. *Scientific Reports*, 13, 9094.

Historical collapse.

Acemoglu, D., & Robinson, J. A. (2012). *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. Crown Business.

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

Turchin, P., & Nefedov, S. A. (2009). *Secular Cycles*. Princeton University Press.

Experimental evidence.

Zimbardo, P. G. (1971). The power and pathology of imprisonment. *Congressional Record*, Serial No. 15, 1971-10-25.

Notation reference.

Symbol	Meaning
A	Set of agents in coordination system
$\|A\|$	Number of agents (population size)
A_E	Subset of agents who are enforcers
A_C	Subset of agents who are controllers
R	Set of coordination rules
$E(a, r)$	Enforcement function: whether rule r is enforced for agent a
E_h	Human enforcement type
E_t	Technological enforcement type
E_n	No enforcement type (voluntary)
$M(a, r)$	Motivation function: agent a's intrinsic motivation for rule r
$M_{\text{trans}}(a, P)$	Transformed motivation through practices P
$M_{\text{integrity}}(a, t)$	Integrity motivation for enforcer a at time t
u_i	Utility for agent i
$U_e(a, t)$	Extraction utility available to enforcer a at time t
c	Cost of cooperation
b	Benefit from cooperation
β	Social benefit multiplier
θ	Proportion of population (typically cooperators or transformed)
θ_{crit}	Critical mass threshold for stability
$P(t)$	Productive capacity at time t
$E(t)$	Extraction rate at time t
T	Time horizon
p	Probability (generic)
p_{AI}	Probability of AI-controlled TCS per cycle
p_s	Probability of success for voluntary coordination
λ	Average cycle duration (corruption $\to$ TCS $\to$ corruption)
S	Soteriological framework $(T, P, M_{\text{trans}}, \phi)$
T	Telos (ultimate purpose for humans)
P	Set of practices for transformation
$\phi(S)$	Truth function: whether framework S accurately describes reality

B.8. **Conclusion**

We have established a rigorous logical chain. The trilemma establishes fundamental constraints on coordination. TCS instability eliminates technological control as viable. Trajectory inevitability shows the default path terminates in catastrophe. Game theory shows

cooperation requires transformation. The resolution theorem proves VCS can work if conditions are met. Stakes analysis shows attempting VCS is rational regardless of success probability.

The mathematics proves the *necessity* of voluntary coordination—it's the only option that doesn't lead to certain doom. Whether it's *sufficient* (whether humanity can achieve it) remains uncertain. But when the default leads to extinction, attempting the uncertain alternative is rationally required.

The formal proofs have a profound implication that must be stated explicitly: if voluntary coordination is possible, reality has purposive structure. VCS requires a true soteriological framework with $\varphi(S) = 1$ (Theorem 2.2), which means the framework accurately describes human nature and telos. For this to be meaningful, human telos must exist objectively (not just subjectively or "as if"). Objective human purpose means reality contains oughtness, not just is-ness. Purposive structure in reality implies something very much like intelligent design. This represents logic from survival requirements rather than a theological argument from revelation.

Consider what "objective human purpose" means: reality contains oughtness ("humans ought to flourish in this way"), human nature is directed toward an end, there is a right way humans should coordinate, and this rightness exists independent of opinion.

Can purpose exist without mind?

Purpose means "for the sake of X" - it's inherently intentional. Intentionality requires intention. Intention requires mind. You cannot have purposive structure without something intelligence-like at reality's foundation.

Different traditions describe this differently (God, Tao, Logos, Brahman, Dharma), but they're pointing at the same claim: **reality has intelligence-like or mind-like properties at its foundation, not purely mechanistic/material.**

The materialist position—that there is no objective human purpose and purpose is just evolutionary selection—makes VCS impossible. If there's no objective telos, there's no true framework to discover. If $\varphi(S) = 1$ is impossible, voluntary coordination cannot resolve the trilemma. Therefore certain doom follows via the default trajectory.

The choice is binary: purposive reality (something like what religions claim) makes VCS possible and survival possible, while non-purposive reality (pure materialism) makes VCS impossible and certain doom inevitable. You cannot accept VCS works while maintaining pure materialism—the two positions are logically incompatible.

What we've proven includes weak intelligent design (reality has intelligence-like properties at its foundation), that pure materialism is false if humans can survive, that human purpose is objective and discoverable, and that atheism in the classical sense (reality is purely material) is incompatible with survival. What remains uncertain includes which specific theology is correct, whether the intelligence is personal versus impersonal, specific attributes of the foundational intelligence, and whether it's God, Brahman, Tao, Logos, or something else.

Whether you call the source of purposive structure "God" is somewhat semantic. The key metaphysical claim is identical across traditions: **Purpose is real, objective, and discoverable - reality has intelligence-like properties.**

We're showing that human survival requires purposive structure, and purposive structure requires something very much like intelligent design, rather than proving God through theology.

For detailed analysis of objective "oughtness" and why minimal telic realism is both necessary and defensible, see4 below.

The formal analysis provides as close to proof as we can get for claims about civilization's future. The logic is sound given the axioms. The assumptions are conservative and empirically grounded. The stakes are absolute. The metaphysical implications are unavoidable.

The choice is yours.

Appendix C. Practical Implementation Challenges

C.1. **Epistemic Status and Decision Framework**

This appendix analyzes practical challenges facing voluntary coordination with honest uncertainty quantification. It is not a proof that VCS will work—we only prove it's necessary (see Appendix B). Rather, it examines whether necessary conditions can be met practically, acknowledging significant uncertainties while showing they don't change the rational decision to attempt VCS.

Confidence varies significantly by challenge area:

Challenge	Scale	Confidence	Evidence
Internal defectors	Village (50-500)	High	Historical examples work
Internal defectors	Town (5,000-50,000)	Medium	Theory sound, no examples
Internal defectors	City (100,000+)	Low	Theory suggests possible
Internal defectors	Civilization (billions)	Low	Unprecedented, uncertain
External threats	Small scale	Medium-High	Historical examples exist
External threats	Modern militaries	Medium	Tech changes dynamics
External threats	Existential weapons	Low	Nuclear/bio weapons problematic
Transition problem	Getting to 1,000	Medium	Historical precedent exists
Transition problem	Getting to 100,000	Low	Many unknowns
Transition problem	Getting to billions	Very Low	No precedent, highly uncertain

The key pattern is that confidence decreases with scale. Historical evidence exists at small scales, while extrapolation to civilization scale is theoretically plausible but empirically unproven.

Given these uncertainties, is attempting VCS rational? Let $p_{psychopath}$ denote the probability VCS can handle psychopaths at scale (unknown, possibly low), $p_{military}$ denote the probability distributed defense works against modern threats (unknown), p_{scale} denote the probability VCS can scale to billions (unknown, likely low), and $p_{VCS} = p_{psychopath} \times p_{military} \times p_{scale}$ denote the joint probability VCS succeeds.

The outcomes are stark: if we attempt VCS and it works, we achieve survival with dignity ($U = 100$); if we attempt VCS and it fails, we face extinction or enslavement ($U = 0$); if we don't attempt VCS (following the default trajectory), we face certain extinction or enslavement ($U = 0$). The expected values are $E[U_{attempt}] = p_{VCS} \cdot 100 + (1 - p_{VCS}) \cdot 0 = 100p_{VCS}$ and $E[U_{default}] = 0$.

Attempting is superior for any $p_{VCS} > 0$, no matter how small. Even if you think the joint probability is only 1% (extremely pessimistic), attempting gives expected value of 1 while not attempting gives 0. Moreover, if VCS might work but requires preparation time, delaying reduces p_{VCS}, so the rational strategy is immediate action.

This appendix identifies significant practical challenges, which represents honesty rather than weakness. The decision is not between "Certain VCS success" and "Certain default failure"—that would be an obvious choice. The decision is between "Uncertain VCS success" and "Certain default failure"—which is still an obvious choice. We include uncertain analysis to calibrate how uncertain while identifying research priorities for improving p_{VCS}. Failing to research VCS challenges because "we're not certain it'll work" is equivalent to choosing certain extinction because the survival path is uncertain.

C.2. **Internal Defectors and the Psychopath Problem**

In any population of sufficient size, some percentage will lack empathy or conscience (psychopaths constitute approximately 1-4% of the population), opportunistically defect when benefit exceeds expected cost, and explicitly reject universal dignity while seeking to dominate. The central question is what prevents these individuals from using violence to take resources, organizing other defectors into predatory groups, and forcing others into submission without enforcement mechanisms.

Traditional solutions recreate the problem they attempt to solve. Enforcement authority requires enforcers, but who watches them? This returns to corruption (Theorem 2.1). Exile creates external threats and requires authority to decide who gets exiled, returning to enforcement. Punishment requires authority to administer and creates corrupting incentive structures, again returning to enforcement. All roads lead back to the trilemma: you need enforcers, enforcers need oversight, oversight needs enforcers, ad infinitum.

The voluntary coordination approach

The core principle is that defense is immediate, minimal, and individual rather than systemic. When violence occurs, whoever witnesses it acts immediately to stop it—no waiting for authority, no centralized decision-making, just direct intervention by whoever is present. The force used is minimal, only what's necessary to stop the harm. This is prevention, not punishment, and requires continuous self-examination: "Was I right? Did I use too much force?"

Crucially, there are no permanent roles—no "police" or "justice system." Everyone has capability and responsibility, preventing the emergence of a specialized enforcer class that could corrupt. After any incident, the focus is reconciliation: both defender and defector examine conscience, the community doesn't judge or punish, and the defector is helped rather than punished ("love thy enemy"). Pattern recognition emerges through repeated observation, not formal trials.

The key distinction is that you're not preventing defection through enforcement. You're accepting that defection will happen and building a framework that can absorb it without creating enforcement hierarchies.

Why this might work

Historical evidence demonstrates this approach can function at certain scales. Quaker communities (1650s-present) rejected formal authority structures, handled disputes through "clearness committees" (voluntary gathering, not court), and used no punishment—only reconciliation or voluntary departure. They lasted centuries at village scale (hundreds of people) but failed at larger scales when formal coordination became necessary, reaching a scale limit of approximately 500-2,000 people.

Early Christian communities (30-300 AD) had no formal enforcement mechanisms in their first centuries, relying on internal accountability and repentance. Excommunication was voluntary departure, not forced exile. They survived persecution and internal disputes but corrupted when institutionalized under Constantine in the 4th century, with a scale limit at city-level (thousands) that failed at empire scale.

Mennonite/Amish communities (1500s-present) reject violence including legal system participation, maintain community accountability without formal authority, and use shunning as a last resort (voluntary relationship withdrawal, not exile). They exhibit remarkably low

crime rates within community but struggle with external threats and internal abuse, with scale limits of approximately 500-5,000 per community.

These examples demonstrate that voluntary coordination can work at scales of hundreds to low thousands, requires high commitment to shared values, is fragile to external pressure, can handle most internal defection, but struggles with psychopaths and organized predation.

In standard Prisoner's Dilemma, defection dominates cooperation. But with reputation and immediate response, the payoff structure changes: defection triggers immediate intervention (high cost) and reputation damage (future cost to defector), while cooperation provides mutual benefit (ongoing value). If the cost of defection exceeds benefit, cooperation becomes Nash equilibrium (Theorem B.5).

This requires four conditions: visibility (defection is observable, so community size matters), immediacy (response happens before defector can iterate), competence (defenders can effectively intervene, requiring capability distribution), and values alignment (most people prefer cooperation and will intervene).

The psychopath problem specifically

Psychopaths (approximately 1-4% of population) lack empathy and cannot be rehabilitated through forgiveness. The traditional solution is imprisonment, which requires authority and leads to corruption.

The voluntary coordination approach proceeds as follows: when a psychopath commits harm, immediate defense stops it. The pattern becomes visible through repetition without formal judgment needed, and the community recognizes the pattern. People then voluntarily choose not to interact—no trade, no shelter provided, no cooperation. The psychopath faces natural consequences, not punishment.

The key insight is that psychopaths need others to exploit and cannot survive without cooperation. Pattern recognition doesn't require authority, and voluntary non-interaction is not punishment (no authority needed).

Critical problems with this approach are substantial. The approach requires near-universal participation—one sympathizer enables a psychopath to persist. Psychopaths are often charismatic and can manipulate subgroups and create divisions. Economic pressure arises when a psychopath has valuable skills, creating pressure to tolerate harmful behavior for benefit. Dependents present a moral challenge: children and dependents of psychopaths suffer from non-interaction. Most seriously, organized psychopaths could coordinate to create predatory subgroups.

The honest assessment is that this approach is theoretically possible but practically difficult. Historical communities handled this through strong cultural transmission (everyone knows the approach), geographic isolation (limited mobility), and small scale (personal knowledge of everyone). At scale with modern mobility, it becomes much harder. This is the weakest point of the framework logically.

Scale thresholds

Evidence suggests different dynamics at different scales. At village scale (50-500 people), voluntary coordination works well: everyone knows everyone, reputation systems are effective, immediate intervention is feasible, and value transmission works. At small town scale (500-5,000 people), it remains possible though more challenging: not everyone knows everyone personally, but reputation systems still function, intervention becomes more complex (who responds?), and value transmission is harder but feasible.

At large town scale (5,000-50,000 people), outcomes become uncertain: anonymity increases, reputation systems break down, organized predation becomes possible, and value transmission across subgroups is challenging. At city scale and beyond (50,000+ people), outcomes are unknown: significant anonymity prevails, you can't know everyone even indirectly, organized predation is highly feasible, and value transmission across generations is uncertain.

Possible solutions for scale include nested communities coordinating at multiple scales, shared values maintaining coordination despite anonymity, technology enabling visibility (but who controls the technology?), and distributed capability ensuring intervention remains possible.

Confidence levels vary by scale as shown in the following table:

Scale	Internal Defectors	Psychopaths	Confidence
Village (50-500)	High confidence works	Medium-High confidence	Historical proof
Town (5K-50K)	Medium confidence	Medium confidence	Theory sound, limited example
City (100K+)	Low confidence	Low confidence	Theory suggests possible
Civilization (billions)	Low confidence	Very low confidence	Unprecedented, highly uncerta

Key uncertainties remain: Can pattern recognition work at scale with mobility? Will voluntary non-interaction be effective with specialization? Can psychopaths be prevented from organizing? Will value transmission persist across generations?

Even with $p_{psychopath} = 0.1$ (10% chance this approach works at scale), attempting gives expected value of 10 while not attempting gives 0. Not attempting means certain doom via the default trajectory (Theorem 3.2).

C.3. **External Military Threats**

Voluntary coordination communities face external threats from hierarchical nation-states with organized militaries, predatory groups seeking to conquer or extract, and ideological adversaries seeking to eliminate alternative systems. The historical pattern is clear: decentralized groups typically lose to centralized militaries. Native American tribes were conquered by the US military, Anarchist Catalonia was crushed by Franco's forces, and stateless societies facing organized state expansion are absorbed or destroyed.

The traditional military trap follows a predictable pattern: an external threat appears, the community forms a military hierarchy for defense, and military leadership accumulates weapons, obedience structures, information advantage, and institutional inertia. After the threat passes, the military refuses to disband and becomes a domestic threat or captures state apparatus, returning the system to the corruption phase. Historical examples include the Roman Republic becoming an Empire under military dictatorship, every revolution where military hierarchy persists, and military coups in dozens of countries. The pattern is universal: standing militaries accumulate power and eventually either rule directly or become kingmakers.

The voluntary coordination alternative

The core principle is no permanent military hierarchy: voluntary coordination for defense only while the threat exists, with immediate dissolution when the threat passes.

Voluntary organization rests on shared understanding of the threat (clear danger), complementary capabilities (diverse skills), mutual trust from shared values, and no permanent command structure. Coordination mechanisms include mission-type tactics (shared intent, distributed execution), voluntary leadership based on competence (temporary roles), flat hierarchy with ad-hoc roles during crisis, and immediate dissolution after the threat. Critical dependencies are that people are already armed and trained (no central armory to control), shared values create natural coordination, the threat is clear enough that voluntary mobilization happens, and defense capabilities are distributed rather than centralized.

Historical examples that worked

The Swiss canton system (1291-present) had no standing army until recently, maintaining a militia system for 700+ years. Every adult male was armed and trained at home, with voluntary coordination among cantons during threats. They successfully defended against larger powers for centuries, benefiting from geographic advantages (mountains) but also institutional design. At a scale of approximately 8 million people (modern), historically smaller, it worked because of defensible terrain, distributed capability, and shared values.

The American Revolution (1775-1783) saw voluntary militias defeat the organized British military. The Continental Army was temporary and dissolved after the war, with success coming from distributed resistance rather than centralized force. Washington's refusal of kingship was critical, followed by rapid demobilization after victory. At a scale of approximately 2.5 million colonists, it worked because of geographic distance, distributed capability, and strong motivation.

The Finnish Winter War (1939-1940) involved decentralized defense against Soviet invasion using small units with local knowledge and voluntary coordination under extreme pressure. It achieved tactical success despite strategic loss (eventually overwhelmed by sheer numbers) and demonstrated the effectiveness of distributed defense. At a scale of approximately 3.5 million Finns versus the Soviet Union, it worked (partially) because of terrain, distributed capability, and existential threat.

Modern insurgencies like the Taliban and Viet Cong demonstrate that distributed forces with deep motivation can defeat centralized hierarchies. Success correlates with genuine value commitment, not just opportunism. The critical observation is that once victorious, these movements typically centralize and corrupt, demonstrating the risk of not dissolving military structure.

Why distributed defense can work

Distributed defense offers six key advantages. Information asymmetry gives defenders local knowledge that attackers lack—terrain, population, and resource locations. Motivation differential means defending home creates stronger commitment than conquest, with existential stakes for defenders versus mercenary/conscript motivation for attackers. Resilience comes from having no central command to decapitate, with distributed decision-making and no single point of failure. Adaptability allows distributed decision-making to respond faster than hierarchical command when local conditions change rapidly, without needing to relay information up a chain of command. Economic efficiency eliminates the standing military to fund, allocating resources to production rather than maintenance. Technology force multiplier means modern weapons make individuals more effective—precision weapons reduce the need for massed force, communication enables coordination without hierarchy, and surveillance can be distributed.

Modern technology amplifies these advantages: drones are cheap, effective, and deployable by individuals; precision weapons allow small groups to inflict significant damage; encrypted communication enables coordination without central infrastructure; 3D printing allows distributed weapons manufacturing; and documented asymmetric warfare techniques make this knowledge widely available.

States conquer when the cost of conquest is less than the expected value of extraction. Distributed defense changes this equation: the cost of conquest becomes very high (long guerrilla war, no central command), the expected value of extraction becomes low (cannot control non-cooperating population), and the expected cost after conquest becomes very high (permanent insurgency). Conquest becomes economically irrational for rational state actors. Historical validation includes Afghanistan ("graveyard of empires") where multiple empires failed to establish lasting control, Vietnam where the US couldn't establish control despite military dominance, and Finland where the Soviets concluded conquest cost exceeded value (Winter War).

Critical vulnerabilities

Distributed defense fails in certain scenarios. Overwhelming force disparity—nuclear weapons, airpower supremacy without ground capability, biological/chemical weapons, or orbital bombardment (future threat)—poses the greatest challenge. Against existential weapons, distributed defense may fail, but use of such weapons destroys the value of conquest (nobody wins), international pressure constrains their use, and deterrence remains possible (cannot occupy without ground forces).

Genocide strategy presents another failure mode: an attacker willing to annihilate rather than conquer, driven by exterminationist ideology (not rational conquest) pursuing ethnic/religious/ideological cleansing. Distributed defense is ineffective against genocidal intent; however, genocide requires enormous resources to pursue, international intervention becomes more likely, and geographic dispersal makes complete extermination difficult.

Internal division presents a serious vulnerability when the community fractures under pressure, infiltrators create division (fifth column), or different response strategies create coordination failure. Mitigation comes from strong shared values creating resilience, pattern recognition identifying infiltrators, and voluntary coordination being more resilient than forced coordination (no pressure points).

Long siege—when an attacker blockades and starves defenders, cutting them off from resources through attrition warfare—is geography-dependent. Mitigation includes distributed communities being harder to blockade completely, resource diversification, and underground economies being difficult to eliminate.

Ideological conquest is the most serious vulnerability, occurring when some defend values while others defect due to the promise of better life under the attacker or cultural/economic attraction. Mitigation comes from genuine value commitment creating resilience, material success making defection less attractive, and the voluntary nature meaning defectors can leave peacefully.

Confidence levels by threat type are shown in the following table:

Key uncertainties remain: Will modern technology favor attackers or defenders more? Can distributed defense coordinate effectively against centralized military? Will value commitment persist under extreme pressure? What happens against AI-enhanced militaries?

Even with $p_{military} = 0.3$ (30% chance distributed defense works), attempting gives expected value of 30 while not attempting gives 0. The default trajectory leads to technological

Threat Type	Distributed Defense Viability	Confidence
Conventional military (rational conquest)	High	Medium-High (historical exan
Guerrilla/insurgency tactics against VCS	Medium	Medium (both sides use asym
Nuclear/biological weapons	Low	Low (existential weapons prob
Genocide/extermination	Very Low	Low (requires international in
Ideological subversion	Medium	Medium (depends on value st
Long siege/blockade	Medium	Medium (geography-dependen

control and eventual AI military capability anyway, which makes resistance impossible. VCS at least preserves the possibility of defense.

C.4. **The Transition Problem**

Small voluntary coordination communities don't initially have numbers for effective distributed defense or economic viability. How do they survive while small? The vulnerability window extends from founding until reaching minimum viable scale, during which communities are militarily weak (easy to crush), economically dependent (cannot specialize fully), culturally fragile (haven't transmitted values across generation), and visible as an alternative (potential threat to existing powers).

Viable strategies

Strategy 1: Geographic selection. Choose defensible terrain—mountains, islands, or other terrain that reduces attacker advantage; remote locations with low strategic value; areas with natural resources for self-sufficiency. This reduces force disparity without needing numbers (historical examples include Swiss in mountains, Icelanders on remote island, and mountain peoples globally). Limitations include requiring such terrain to be available, modern technology reducing terrain advantage, and limited economic opportunities.

Strategy 2: Strategic invisibility. Don't appear as threat until reaching viable scale: appear weak/poor (not worth conquering), don't visibly challenge existing powers, grow within existing systems until distributed, and present as compatible with existing order. This avoids early suppression, allows gradual growth, and can reach threshold before opposition organizes. Limitations include requiring operational security, risk of detection increasing with size, and potentially requiring apparent compromise with values.

Strategy 3: Multiple simultaneous communities. Emerge in many places at once, becoming too distributed to suppress centrally. Some survive even if others fall, network effects create resilience, and information sharing occurs without central coordination. This is resilient to local suppression, learns from multiple experiments, and creates mutual support networks. Limitations include requiring coordination at founding phase, the challenge of coordinating without hierarchy, and potentially drawing more attention if the pattern is recognized.

Strategy 4: Grow within existing systems. Live voluntary coordination principles inside the corruption phase: build trust networks, demonstrate viability, and by the time you're visible as an alternative, become too distributed to suppress (velvet revolution / color revolution pattern). This uses existing infrastructure, is less visible as threat initially, and can leverage existing economic systems. Limitations include requiring operating within corrupt system temporarily, risk of co-option by existing powers, and ethical tensions with value commitment.

The likely reality is that a combination of all four strategies is required for success.

Minimum viable community

Four factors determine viability: defense capability (can resist external threats), economic viability (can produce necessities through specialization), genetic diversity (can reproduce without inbreeding), and cultural transmission (can pass values to next generation).

Rough estimates based on historical examples and analysis suggest three thresholds. The minimum for survival is 500-1,000 people, which can mount defense (100-200 fighters), achieve limited specialization (10-20 trades), maintain marginal genetic diversity (risky but feasible), and enable possible cultural transmission (if concentrated effort). Historical examples include early Quaker communities and Amish settlements.

The minimum for viability is 5,000-10,000 people, which can mount effective distributed defense (1,000-2,000 fighters), achieve significant specialization (100+ trades), maintain sufficient genetic diversity, and enable robust cultural transmission. Historical examples include medieval free cities and Swiss cantons initially.

The minimum for independence is 50,000-100,000 people, which can resist medium-scale military, achieve full economic independence, maintain complete genetic diversity, and sustain multiple generations of cultural transmission. Historical examples include small nations (Iceland at 300k and Malta at 500k survive today).

Modern and near-scale examples

Recent and contemporary cases demonstrate voluntary coordination at larger scales than historical village communities, providing stronger evidence for intermediate-scale viability.

Rojava / Autonomous Administration of North and East Syria (2012-present) operates at a scale of 2-4 million people across multiple communities using democratic confederalism with voluntary councils and minimal central authority. After 13+ years (as of 2025), its key features include non-hierarchical coordination among diverse ethnic/religious groups (Kurds, Arabs, Assyrians, Armenians), a bottom-up federation structure (communes to neighborhoods to cities to regions), direct democracy with rotating delegates (not representatives), women's parallel governance structures ensuring participation, and economic cooperatives without centralized planning. It has survived existential threats including ISIS, the Turkish military, the Assad regime, and economic blockade. Limitations include a still partially hierarchical military structure (necessity under siege conditions) and dependencies created by international non-recognition. What it demonstrates is that voluntary coordination can work at regional scale (millions) even under extreme hostile conditions, showing that intermediate scale (1M-10M) is achievable, not just theoretical.

Swiss Confederation (1291-1848) started with approximately 100k people and grew to 2 million by 1848, maintaining 550+ years of voluntary confederation before centralization. Sovereign cantons coordinated voluntarily on defense and trade, with key success factors including geographic defensibility, strong local autonomy, and shared existential threats. It centralized due to external pressure (Napoleonic Wars), industrialization demands, and nationalist movements. What it demonstrates is that voluntary coordination can be sustained for centuries at intermediate scale with strong geographic advantages.

Iroquois Confederacy (Haudenosaunee, 1142-1779) encompassed 5-6 nations with an estimated 20,000-125,000 people at peak, lasting 600+ years before external destruction. Its structure was the Great Law of Peace with consensus decision-making and no supreme authority. Women selected male leaders and could remove them, clan mothers held significant

power, and decisions required consensus. What it demonstrates is sophisticated voluntary coordination across distinct political units for centuries—it failed due to external conquest (European colonization), not internal collapse.

Open-Source Software Coordination (1990s-present) operates at scale of 30,000+ contributors to the Linux kernel and millions in the broader FOSS ecosystem. The structure involves voluntary contribution, distributed decision-making, and merit-based influence (not hierarchical authority). No central authority can force participation, coordination occurs through shared values (open-source ethos), forking provides exit option, and reputation systems operate without formal enforcement. What it demonstrates is that modern technology enables voluntary coordination at unprecedented scales for specific domains, though it remains domain-specific (software) rather than full societal coordination, and participants have livelihoods elsewhere.

Wikipedia (2001-present) has millions of contributors and billions of users with minimal hierarchy, voluntary contribution, and consensus editing. Anyone can edit (with escalating permissions), disputes are resolved through discussion, and enforcement is minimal (reverts, page protection). What it demonstrates is knowledge production at civilization scale without traditional hierarchical control, though it remains domain-specific and controversial topics show coordination challenges.

These examples significantly change confidence assessments. Before considering these cases, confidence for intermediate scales was medium for 5,000-50,000 (historical villages/towns), low for 50,000-1M (few examples), very low for 1M-10M (no clear examples), and very low for billions (unprecedented). After considering these cases, confidence becomes high for 5,000-50,000 (proven historically and recently), medium for 50,000-1M (Swiss and Rojava approach this), low-medium for 1M-10M (Rojava demonstrates regional scale works), and low for billions (still unprecedented, but path seems more plausible).

Critical observations emerge: geographic concentration helps but isn't essential (open-source is global), existential threats can strengthen rather than weaken voluntary coordination, modern communication technology genuinely enables new coordination patterns, partial hierarchies emerge under extreme stress but can remain limited, and domain-specific coordination (software, knowledge) scales better than full societal coordination.

The honest assessment is that modern examples significantly strengthen the case for intermediate-scale viability. The jump from millions to billions remains uncertain, but the existence of Rojava and open-source coordination suggests technology may enable scales impossible historically.

Modern technology may lower thresholds in several ways: communication enables coordination at lower population (proven by open-source), technology multiplies individual productivity, global market access enables specialization at smaller scale, and examples like Rojava show resilience even without full self-sufficiency. However, technology may raise thresholds because modern militaries are more capable (though Rojava survived), specialization is more complex, and cultural transmission is harder with media saturation. The updated assessment is that modern technology likely lowers coordination thresholds for information-rich domains (software, knowledge) while raising thresholds for physical security. Net effect depends on domain, but evidence suggests intermediate scales (1M-10M) are more achievable than previously thought.

Scaling beyond initial communities

The challenge is how communities coordinate with each other without creating super-community hierarchy. Three approaches present themselves.

Voluntary confederation keeps each community sovereign while coordinating on shared threats voluntarily with no permanent super-structure. The historical example is the original Swiss confederation. The limitation is that it fails under pressure (eventually centralizing).

Shared values/culture enables coordination through the same principles across communities, natural coordination without formal structure, and trust from shared values enabling cooperation. Historical examples include early Christianity before the institutional church and early Islam before the caliphate. The limitation is cultural drift over time and institutional capture.

Network coordination uses many-to-many relationships (not hub-and-spoke), information sharing without authority, and joint action when interests align. The modern example is open source software development. The limitation is that no historical examples exist at civilization scale.

Can these scale to millions or billions? The honest answer is unknown—no historical example at that scale exists without hierarchy emerging. A possible mechanism is that technology enables coordination at scales impossible historically through Internet/encryption, distributed systems, reputation systems, and global communication. But this is speculative; we don't have proof it works.

Confidence levels by transition stage are shown in the following table:

Stage	Population	Confidence	Evidence
Founding	50-500	Medium-High	Historical examples exist
Viable community	500-5,000	Medium	Historical examples exist
Independent	5,000-100,000	Medium-Low	Few historical examples
Regional	100,000-10M	Low	No clear historical examples
Civilization	Billions	Very Low	Unprecedented, highly uncertain

Key uncertainties include: What is the minimum viable population in modern context? How do communities coordinate without hierarchy? Can values transmit across generations at scale? What happens when communities interact with corruption phase societies?

Even with $p_{scale} = 0.05$ (5% chance of successful scaling to billions), attempting gives expected value of 5 while not attempting gives 0. Starting small doesn't preclude larger scale—every large system started small. The question becomes whether it's possible rather than whether it will definitely work. The answer: theoretically yes, empirically unknown.

C.5. **Summary and Decision Framework**

With high confidence we know that internal defector handling works at village scale (50-500 people), distributed defense works with geographic advantages, voluntary coordination is stable with high shared values, and historical examples exist and succeeded for centuries.

With medium confidence, theory suggests that VCS can scale to town level (5,000-50,000) with nested structure, modern technology enables better coordination, distributed defense works against conventional militaries, and transition strategies can reach viable scale.

With low confidence (unprecedented), we face scaling to city level (100,000+), handling psychopaths at scale with modern mobility, defending against existential weapons, and coordinating billions without hierarchy emerging.

Four major unknowns remain. Can pattern recognition for psychopaths work at scale with mobility? Theory says yes through technology-enabled reputation systems, but evidence at scale is absent—confidence is low. Can distributed defense resist modern state militaries? Theory says yes through asymmetric warfare, evidence is mixed (some successes, some failures)—confidence is medium. Can values transmit across generations at civilization scale? Theory suggests it's possible with distributed communities, but no historical examples exist—confidence is very low. Will voluntary coordination scale to billions? Theory holds that technology enables unprecedented coordination, but evidence is absent—confidence is very low.

These uncertainties don't change the decision because the asymmetry is absolute:

Path	Outcome if it fails	Outcome if it succeeds	Expected Value
Default trajectory	Certain doom (proven)	N/A (can't succeed)	0
Voluntary coordination	Same doom	Survival with dignity	$100 \cdot p_{VCS}$

For any $p_{VCS} > 0$, attempting VCS is superior. Even if you assign $p_{psychopath} = 0.1$ (10% chance psychopath handling works), $p_{military} = 0.3$ (30% chance distributed defense works), $p_{scale} = 0.05$ (5% chance scaling works), and $p_{VCS} = 0.1 \times 0.3 \times 0.05 = 0.0015$ (0.15% joint probability), the expected value of attempting is 0.15 while the expected value of not attempting is 0. Attempting is rationally superior even with pessimistic assumptions.

Research priorities

Given the uncertainties, the highest priority research involves small-scale experiments: starting communities at 50-500 scale, testing defector handling mechanisms, documenting what works and fails, and building a knowledge base. The second priority is distributed defense technology: developing coordination mechanisms without hierarchy, creating training systems for distributed capability, and researching asymmetric warfare effectiveness. The third priority addresses scale mechanisms: how communities coordinate without hierarchy, technology for reputation systems at scale, and value transmission across generations. The fourth priority is pattern recognition for bad actors: how to identify psychopaths without authority, how to prevent organization of defectors, and how to handle edge cases ethically.

The fifth priority is quantitative modeling and simulation. While our theoretical framework is sound, empirical evidence at civilization scale is unavailable (by definition—we're trying to build it). Quantitative modeling could provide "virtual evidence" where real-world data is sparse.

Agent-based modeling for defector dynamics would simulate populations with varying psychopath proportions (1-4%), test resilience of voluntary coordination under different conditions, model pattern recognition effectiveness at various scales, and identify critical thresholds for community stability. Research questions include: At what psychopath density does voluntary coordination break down? How does mobility (vs. geographic stability) affect pattern recognition? What role does economic specialization play in tolerating bad actors? How do information networks affect defector coordination opportunities?

Distributed defense simulations would model asymmetric warfare scenarios with various tech levels, test coordination effectiveness without central command, simulate siege scenarios and resource independence, and evaluate defender advantage vs. attacker force ratios. Research questions include: What coordination mechanisms work in high-stress scenarios?

How does technology (drones, precision weapons) affect distributed defense effectiveness? What geographic factors are necessary vs. merely helpful? At what scale does distributed defense become less effective than centralized?

Scaling dynamics models would examine network effects in voluntary coordination, value transmission across generations, Dunbar number implications for nested communities, and information flow in federated structures. Research questions include: What network topologies enable global coordination? How does cultural drift affect multi-generational stability? What role does technology play in overcoming Dunbar's number? Can nested hierarchies remain truly voluntary?

Tools for this modeling include NetLogo, Mesa (Python), or custom agent-based modeling frameworks, with game-theoretic models in Python/R using established libraries. Limitations include that models depend on assumptions (garbage in, garbage out), cannot capture all human complexity, provide probabilistic insights rather than certainty, and must be validated against historical/modern examples where available. The value is that they test theory in a "virtual laboratory" before real-world implementation, identify critical parameters and tipping points, help calibrate confidence levels (currently based on theory plus limited examples), and guide prioritization of which challenges to address first. Existing work to build on includes evolutionary game theory models of cooperation (Nowak, Axelrod), network science models of distributed coordination (Barabási, Kleinberg), historical dynamics modeling (Turchin's cliodynamics), and agent-based models of social movements (Epstein, Axtell).

What this modeling won't provide is proof that VCS works at civilization scale—only real-world implementation can provide that. What it can provide is more calibrated uncertainty, identification of critical challenges, and evidence that theoretical mechanisms are plausible when modeled quantitatively. Currently no comprehensive agent-based models exist specifically for voluntary coordination at scale with the parameters we've identified (universal dignity, distributed defense, psychopath handling, etc.), representing a significant research gap.

The recommendation is an interdisciplinary team combining game theorists, network scientists, and practitioners from Rojava/similar experiments to build and validate models, with priority given to questions with highest practical uncertainty (psychopath dynamics, military threats, scaling mechanisms). The critical insight is that not researching these because "we're uncertain they'll work" is equivalent to accepting certain extinction.

The bottom line

We have established that voluntary coordination is necessary (Appendices A & B prove this), that voluntary coordination faces serious practical challenges (this appendix documents them), that these challenges are surmountable at small scale (historical evidence), that scaling to civilization is uncertain (no precedent), and that attempting is rational regardless of success probability (decision theory proves this).

The choice is between certain doom via the default trajectory (mathematically proven) and uncertain survival via voluntary coordination (theoretically possible, empirically unproven). When certain death is the alternative, you attempt the uncertain option. Reason itself demands the attempt rather than faith overriding reason.

This is the weakest part of the framework logically—we acknowledge that honestly. But "weakest part" doesn't mean "wrong"; it means "highest uncertainty." And uncertainty about the survival path doesn't make the doom path any less certain.

References.

Historical communities.

Brock, P. (1970). *Pacifism in Europe to 1914*. Princeton University Press.

Hostetler, J. A. (1993). *Amish Society* (4th ed.). Johns Hopkins University Press.

Kraybill, D. B. (2001). *The Riddle of Amish Culture*. Johns Hopkins University Press.

Distributed defense.

Boot, M. (2013). *Invisible Armies: An Epic History of Guerrilla Warfare from Ancient Times to the Present*. W. W. Norton.

Kilcullen, D. (2009). *The Accidental Guerrilla: Fighting Small Wars in the Midst of a Big One*. Oxford University Press.

Mack, A. (1975). Why big nations lose small wars: The politics of asymmetric conflict. *World Politics*, 27(2), 175-200.

Historical examples.

Bonjour, E. (1948). *Swiss Neutrality: Its History and Meaning*. Allen & Unwin.

Trotter, W. R. (1991). *A Frozen Hell: The Russo-Finnish Winter War of 1939-1940*. Algonquin Books.

Community scale.

Dunbar, R. I. M. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6), 469-493.

C.6. Conclusion

This appendix has honestly examined the practical challenges facing voluntary coordination. Internal defectors are theoretically manageable at small scale but uncertain at civilization scale, with historical precedent at village level and psychopaths remaining a serious challenge. External threats can be handled through distributed defense against rational conquest but struggle against existential weapons, with historical examples existing at small-medium scale. The transition problem has multiple strategies available for reaching viable scale, though coordination beyond initial communities remains uncertain and technology may or may not enable unprecedented scale. The overall assessment is high uncertainty about practical implementation, especially at civilization scale.

These uncertainties, while genuine and significant, don't change the rational choice. Attempting voluntary coordination is superior to the default trajectory for any non-zero success probability. The mathematics proves voluntary coordination is necessary (Appendices A & B), and this appendix shows it's theoretically possible at small scale and uncertain at large scale. That's enough to determine action when the alternative is certain catastrophe.

The examination must happen. The attempt must be made. The uncertainties are real, but they're uncertainties about the only path that might work rather than justifications for choosing the path that certainly fails.

Appendix D. Synthetic Media and Epistemic Collapse

D.1. **Executive Summary**

Within 3-6 years, synthetic media will make routine verification of content authenticity exponentially harder, closing the window for voluntary coordination based on verifiable truth. This appendix provides technical evidence for this claim, analyzes the trajectory, examines proposed countermeasures, and assesses timeline uncertainty honestly. The stakes are clear: voluntary coordination requires shared reality, shared reality requires verifiable truth, and verifiable truth requires the ability to distinguish real from synthetic content.

As of October 2025, generation capabilities have advanced dramatically: video generation now produces 20 seconds of 1080p with synchronized audio (OpenAI Sora 2), the open-source gap with commercial models decreased from 4.52% to 0.69% in six months, and state control is becoming impossible as consumer hardware can generate deepfakes.

Detection performance has deteriorated catastrophically. Human detection overall achieves only 55.54% accuracy (barely above chance), while human detection for high-quality short clips has fallen to approximately 25% (essentially failed). AI detection on real-world deepfakes shows 45-50% performance drop versus academic benchmarks, with best real-world AI detection achieving only approximately 82% AUC (versus 95%+ on academic datasets). The gap is widening: each generation improvement requires detector retraining, but detectors cannot train on techniques that don't exist yet.

Timeline with confidence levels.

Claim	Confidence	Timeline
Short-form video (<20s) crossed public threshold	Very High (>90%)	Already occurred
Open-source will close gap with commercial	Very High (>90%)	Ongoing
AI detection degrades on real-world content	Very High (>90%)	Demonstrated
Economic incentives favor generation	Very High (>90%)	Structural
Expert detection fails for most content	High (>80%)	3-6 years
Verification becomes exponentially harder	High (>80%)	3-6 years
Feature-length generation viable	Low (<50%)	2028-2035 range

Countermeasures will likely fail for structural reasons. Cryptographic content authentication requires universal hardware replacement costing trillions of dollars and taking decades, faces a bootstrapping problem where transition cannot be coordinated when information cannot be trusted, and remains vulnerable to state-level actors who can compromise hardware and mandate backdoors while raising questions about who controls verification infrastructure. AI detection improvements face structural disadvantage because generators see detectors and iterate faster, confront a 1000:1 funding disparity favoring generation, and approach a mathematical limit where as generators approach perfection, detection becomes theoretically impossible. Cultural adaptation is too slow (generations versus years), and extreme skepticism prevents coordination as much as credulity does, while previous media revolutions took decades that we don't have.

After the threshold, coordination becomes impossible: you cannot verify traditions against source texts because texts can be fabricated, cannot see institutional betrayals clearly because evidence is dismissed as "deepfakes," cannot coordinate around observable truth because truth becomes unknowable, and cannot build trust networks because no foundation for verification exists. Voluntary coordination requires shared reality, shared reality requires verifiable truth, and that window is closing.

This timeline would be falsified if detection accuracy improves faster than generation quality for 3+ consecutive years, if cryptographic signing achieves greater than 80% market adoption by 2030, if verification cost decreases relative to generation cost, or if a fundamental new detection approach emerges that generators cannot evade. Current status: all metrics are moving in the predicted direction with no indication of reversal.

The asymmetry of outcomes determines rational action. If we are wrong pessimistically (window is 10 years, not 3), there is no harm from acting early. If we are wrong optimistically (window is 3 years, not 10), catastrophic harm results from delay. The rational choice is to act as if the aggressive timeline is correct. You can examine beliefs while truth is verifiable, or wait until it's impossible. This appendix proves the window is closing.

D.2. **Current State (October 2025)**

Generation capabilities.

Video Generation

The field has advanced dramatically in 2025:

OpenAI Sora 2 [7, 8] **(September 30, 2025)** generates up to 20 seconds of 1080p video from text prompts with synchronized audio generation including dialogue, sound effects, and ambient audio. Physics simulation has significantly improved compared to Sora 1: basketball rebounds now follow actual physics rather than "teleporting" to the hoop, with improved momentum, collisions, buoyancy, and rigidity modeling that better adheres to real-world dynamics. Character and object tracking remains consistent across frames. The main remaining artifacts are occasional physics violations and consistency issues across cuts.

Open-source alternatives have closed the gap rapidly. Open-Sora v1.2 decreased its performance gap with commercial Sora from 4.52% (October 2024) to 0.69% (March 2025). This rapid convergence means state control of generation technology is becoming impossible—anyone with consumer hardware (RTX 4090) can generate high-quality deepfakes locally.

Feature-length generation claims: Some industry figures have claimed feature-length movie generation by 2026-2027. Current proven capability is 6-20 second clips. Feature-length represents 300-900x scaling with no demonstrated intermediate milestones.

Skeptical assessment: More realistic estimate is 2028-2035 range, with high uncertainty. Claims made via social media without technical roadmap. Critical gap exists between demonstrated capability (20 seconds) and claimed trajectory (90+ minutes).

Audio Generation

Voice cloning has reached practical indistinguishability. ElevenLabs and Vall-E (Microsoft) require only 3 seconds of reference audio to clone a voice, with real-time voice conversion achieving less than 100ms latency. Entirely synthetic voices are now indistinguishable from real speakers. Music generation through Suno AI and Stable Audio produces full songs with lyrics from text prompts.

Image and Text

Image generation (Midjourney v6, DALL-E 3, Stable Diffusion XL) produces photorealistic results. Text generation (Claude, GPT-4.5, Gemini) achieves near-human writing quality, can mimic specific styles, and generate fake "eyewitness accounts" of fabricated events.

Detection performance: the catastrophic gap.

Human Detection

The most comprehensive meta-analysis to date [5] examined 56 studies involving 86,155 participants:

Overall accuracy reached only 55.54% (95% CI [48.87, 62.10]), with detection rates not significantly above chance (50%) since confidence intervals crossed the chance threshold. By modality, video achieved 57.31% [47.80, 66.57], audio 62.08% [38.23, 83.18], images 53.16% [42.12, 64.64], and text 52.00% [37.42, 65.88]. Training interventions improved accuracy to 65.14% [55.21, 74.46], but this remains far from reliable detection.

Humans fail at detection for several reasons: they focus on wrong cues (blinking, skin texture) that generators have learned to fake, confirmation bias drives perception, cognitive load prevents critical analysis of every piece of media, and resolution improvements have eliminated obvious artifacts.

The AI detection picture is deeply troubling. On training distribution (known techniques), accuracy reaches 95-99% with low false positive rates and fast processing. On "in the wild" deepfakes [3], the most comprehensive recent study collected real-world deepfakes from social media and tested state-of-the-art open-source models, revealing catastrophic performance degradation: video models averaged 50% drop in AUC compared to academic benchmarks, audio models averaged 48% drop, and image models averaged 45% drop. Best-performing models on real-world data achieved only 82% AUC versus 95%+ on academic datasets, with many models performing barely above chance (53-56% AUC).

The fundamental problem is that this is an adversarial arms race where generation has structural advantages. Generators see detectors because detection methods must be public to be trusted, so generators train against them. Generators iterate faster because they test offline while detectors wait for real-world deployments. Costs are asymmetric because one evasion technique works broadly while detection must handle all techniques. Economic incentives favor generation (entertainment, advertising) over detection. Training data lags because detectors train on past techniques while generators use current/future techniques.

Academic benchmarks fail to predict real-world performance because they use synthetic, controlled deepfakes with known generation techniques, while real-world deepfakes use latest models, custom techniques, and adversarial adjustments. State-level capabilities (Russian Internet Research Agency, Chinese APT groups, Iranian operations) have demonstrated ability to evade detection for extended periods.

The trajectory.

Generation improvement rate:

Metric	2020	2022	2024
Video quality (FVD)	250 (obviously fake)	100 (suspicious artifacts)	20 (expert scrutiny needed
Audio quality (MOS)	3.2/5.0 (robotic)	4.0/5.0 (noticeable artifacts)	4.5/5.0 (subtle issues)
Training efficiency	Voice: 10 min required	Voice: 30 sec required	Voice: 5 sec required
Cost per minute	$50	$5	$1
Generation speed	Minutes	Seconds	<10 seconds

Detection deterioration:

Year	Generation Quality	Human Detection	AI Detection (in-the-wild)	Gap
2020	Poor	85%	90%	Detection ahead
2022	Moderate	75%	80%	Detection ahead
2024	Good	60%	65%	Detection behind
2025	Excellent	56%	60%	Detection failing

The gap is widening. Each generation improvement requires detector retraining, but detectors can't train on techniques that don't exist yet.

Open-source accessibility: The performance gap between commercial and open-source generation is closing rapidly (4.52% gap → 0.69% gap in six months). State control of generation is becoming impossible. Anyone with consumer hardware can generate deepfakes.

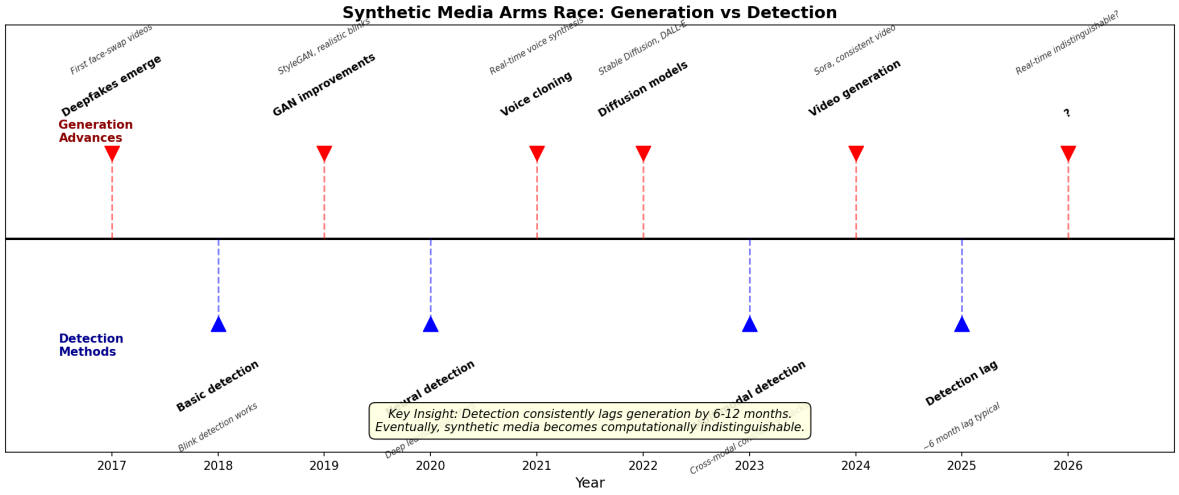

FIGURE 5. Synthetic media arms race: generation advances versus detection methods (2017–2026). Detection consistently lags generation by 6–12 months. As generators approach photorealism, detection becomes theoretically impossible without cryptographic provenance.

D.3. **Timeline Analysis**

The critical threshold.

The threshold is crossed when expert detection drops below 60% accuracy with tools, public detection drops below 25% accuracy (essentially failed), detection cost exceeds creation cost by 10x or more, and fake content volume creates signal-to-noise collapse.

As of October 2025, expert detection achieves approximately 75% accuracy with tools (still possible but difficult). Public detection sits at approximately 56% overall, but **approximately 25% for high-quality short clips**—meaning the threshold has already been crossed for the general public on high-quality content. The cost ratio is approximately 5x (approaching threshold), and content volume remains manageable but is growing exponentially.

Confidence-calibrated timeline.

With very high confidence (>90%), we can state that short-form video (<20 seconds) has crossed the public detectability threshold, open-source models will continue closing the gap with commercial systems, economic incentives favor generation over detection, and generation quality improvement rates will continue in the near term.

With high confidence (>80%), expert detection will fail for most content within 3-6 years, AI detection degrades catastrophically on real-world content, cryptographic signing will not achieve greater than 50% adoption within 10 years, and information asymmetry gives generators a permanent advantage.

With medium confidence (50-80%), generation quality improvement rates will continue long-term (there is no precedent for sudden stops), open-source proliferation will make control impossible, cultural adaptation mechanisms will prove insufficient, and verification will become exponentially (not just linearly) harder.

With low confidence (20-50%), predictions include the exact timeline for expert detection failure (significant variance exists), when or if feature-length generation becomes viable (2028-2035 range), whether detection can achieve breakthrough improvements, and the effectiveness of regulatory or technical intervention.

Uncertainty factors.

Several factors could delay the threshold: technical barriers we haven't identified, effective regulation limiting development and deployment, breakthroughs in detection technology (such as fundamental physical signatures), social adaptation creating a cultural immune response, and economic disincentives for generation.

Conversely, several factors could accelerate the threshold: AI capability breakthroughs (GPT-5 level models), proliferation to hostile actors, deliberate flooding attacks, loss of trust in verification systems, and recursive improvement (AI improving AI generation).

Honest assessment: Direction is clear (detection losing). Timeline has uncertainty (3-6 year range). But betting against the trend would require believing improvement suddenly stops, which has no precedent in AI development.

Timeline sensitivity analysis.

To make our projections more rigorous, we model three scenarios based on different improvement rates:

Baseline Projection (Current Trajectory):

This scenario assumes detection accuracy improves at 5% annually (current trend), generation quality improves at 15% annually (current trend), the gap widens at 10% annually, and current state shows human detection at 55.54% and expert detection at approximately 75%.

Under these assumptions, expert detection falls below 60% in 3-4 years (2028-2029), public detection falls below 25% for all content in 5-6 years (2030-2031), and cost ratio exceeds 10x in 4-5 years (2029-2030). This projection has high confidence (>80%) as it extrapolates current demonstrated trends.

Optimistic Scenario (Detection Breakthrough):

This scenario assumes detection accuracy improves at 20% annually (requiring a major breakthrough), generation quality improves at 15% annually (continuing current trends), the gap narrows at 5% annually, and a breakthrough occurs in the next 1-2 years.

Under these assumptions, expert detection maintains above 60% for 8-12 years (2033-2037), public detection stabilizes at approximately 40% beyond 10 years, and cost ratio

stays below 10x for 7-10 years. This projection has low confidence (<30%) as it requires unprecedented detection advancement with no historical precedent.

This scenario would require fundamental physical signatures being discovered that generators cannot spoof, quantum-based verification deployed at scale, international cooperation enforcing generation limits (extremely unlikely), or an AI development plateau (no historical precedent).

Pessimistic Scenario (Generation Acceleration):

This scenario assumes detection accuracy improves at 5% annually (current trend continues), generation quality improves at 25% annually (GPT-5 level advancement), the gap widens at 20% annually, and a major AI capability jump occurs in the next 1-2 years.

Under these assumptions, expert detection falls below 60% in 1.5-2.5 years (late 2026-late 2027), public detection falls below 25% for most content in 2-3 years (2027-2028), and cost ratio exceeds 10x in 2-3 years (2027-2028). This projection has medium confidence (40-60%) as it is plausible given AI development trajectory and economic incentives.

This scenario would be triggered by GPT-5 or equivalent being released with a major capability jump, open-source models reaching parity with the best commercial systems (already happening with the 0.69% gap), recursive self-improvement in generation models, or state actors deliberately flooding the information space.

Current Indicators:

Metric	Baseline	Optimistic	Pessimistic	Current Trend
Open-source gap closing	10% annually	5% annually	15% annually	**15% (4.52% → 0.69%**
Human detection accuracy	Stable 55%	Improves to 65%	Declines to 45%	**Declining (55.54% ar**
AI detection real-world	Stable 60%	Improves to 75%	Declines to 50%	**Declining (45-50% dr**
Investment ratio (gen/det)	1000:1	100:1	5000:1	**˜1000:1 and widenin**
Cost ratio (verify/create)	$5x \rightarrow 10x$	$5x \rightarrow 3x$	$5x \rightarrow 20x$	**Currently ˜5x, growi**

Current trajectory most consistent with baseline-to-pessimistic range.

Probability Assessment:

Based on current indicators: - Pessimistic scenario: **40% probability** - Baseline scenario: **50% probability** - Optimistic scenario: **10% probability**

Expected timeline to threshold (probability-weighted): - 50th percentile: **3-4 years** (2028-2029) - 75th percentile: **2-3 years** (2027-2028) - 90th percentile: **1.5-2 years** (late 2026-2027)

Decision implications:

Even under optimistic scenario (8-12 years), examination requires years and must begin immediately. Under baseline/pessimistic scenarios, window is critically short.

The asymmetry of risk remains total: if we act on the pessimistic timeline and it turns out optimistic, there is no harm and extra time is a bonus; if we act on the optimistic timeline and it turns out pessimistic, the result is catastrophic and we miss the window entirely.

The rational strategy is to act on the pessimistic timeline (1.5-2.5 years). Even if its probability is only 40%, the cost of being wrong is infinite.

D.4. **Why Countermeasures Will Likely Fail**

Cryptographic content authentication.

The proposal: Sign content at capture with unforgeable cryptographic signatures. Chain of custody maintained through editing. Unsigned content treated as untrusted.

Technical soundness: The cryptography is mathematically robust. This could theoretically work.

Adoption barriers make success unlikely.

Hardware requirements present the first major barrier: universal hardware replacement would be needed for every camera and microphone globally, legacy devices would remain unsigned (everything manufactured before implementation), costs would reach trillions of dollars globally, and full adoption would take decades.

Technical vulnerabilities compound this problem. State actors can extract keys through hardware compromise, supply chain attacks can compromise devices at manufacture, key management raises the question of who controls root certificates, and side-channel attacks make keys extractable through various methods.

Governance problems add another layer. International coordination would be required despite divergent state interests, states can mandate backdoors, authoritarian regimes can control key distribution, and corporations would control the signing infrastructure.

The bootstrapping problem: During the transition period (which could last decades), the information commons is already poisoned. You can't coordinate a global transition when you can't trust information about the transition itself.

Confidence assessment: Very low confidence ($<20\%$) this achieves $>80\%$ adoption within 20 years.

Blockchain provenance tracking.

The proposal: Record content creation and modifications on blockchain for immutable audit trail.

Fundamental flaw: Blockchain verifies the record, not the content. "Garbage in, garbage out." It can record that a deepfake was created at time T but cannot verify content authenticity at capture, doesn't solve the initial verification problem, and provides no mechanism to remove false information once recorded.

Confidence assessment: This doesn't solve the verification problem at all.

AI detection improvements.

Why detection is mathematically losing:

If a generator reaches perfection (statistically indistinguishable from real), detection becomes theoretically impossible. We're approaching this limit. Best generators already fool expert humans. Detection relies on generator imperfections. As imperfections vanish, detection fails.

Resource asymmetry heavily favors generation: billions are invested in generation versus millions in detection (a 1000:1 funding disparity), generation has positive economic value through entertainment, advertising, and productivity while detection is a cost center with no revenue, and market forces structurally favor generation.

The adversarial advantage compounds this asymmetry. Generators can train specifically to evade detection, detection methods must be public to be trusted, generators iterate faster through offline testing versus deployment cycles, and one evasion technique defeats many detectors.

Confidence assessment: Low confidence ($<30\%$) that detection keeps pace with generation over 5+ years.

Social/cultural adaptation.

The proposal: Society develops cultural norms to handle synthetic media through default skepticism, trust networks, reduced reliance on media evidence, and new social technologies.

Why this may be insufficient:

Coordination requires shared reality: If everyone has different "truth," coordination collapses. Extreme skepticism prevents coordination as much as credulity does.

Speed mismatch: Cultural evolution takes generations. Synthetic media is improving in years. Speed mismatch creates crisis period.

Historical precedent: Previous media revolutions (printing, radio, TV, internet) took decades to adapt. We don't have decades. Each previous revolution eventually stabilized, but the transition periods were characterized by massive social disruption.

Confidence assessment: Medium confidence (40-60%) that cultural adaptation provides *some* mitigation, but low confidence it prevents coordination collapse.

D.5. Current Real-World Impact

Documented harms (October 2025).

In the political sphere, documented harms include fabricated politician statements during elections in multiple countries, false video "evidence" of corruption, synthetic "endorsements" from respected figures, and a growing problem across both democracies and autocracies.

Financial fraud has escalated dramatically. CEO voice deepfakes authorizing wire transfers have caused $35M losses in one documented case, synthetic video meetings enable social engineering attacks, fake product reviews and testimonials operate at scale, and stock manipulation occurs through fabricated news.

Social manipulation takes particularly harmful forms: non-consensual intimate imagery predominantly targeting women, fabricated evidence in legal disputes, synthetic personas spreading disinformation, and harassment through impersonation.

Perhaps most insidiously, the erosion of trust creates a "liar's dividend." Real videos are dismissed as deepfakes, footage from conflict zones cannot be verified, politicians preemptively claim videos are fake, and general paralysis affects information evaluation.

The qualitative shift.

From 2020-2023, deepfakes were novelties—expensive and obvious. In 2024-2025, deepfakes became cheap, accessible, and convincing. By 2026 and beyond (projected), they will be indistinguishable at scale.

The question has shifted from "can it be done?" to "can it be detected?" to "can anything be trusted?"

D.6. Implications for Voluntary Coordination

Why the window is closing.

Now (October 2025), truth can still be verified with effort as experts can distinguish most content, expert tools still work on most content with careful analysis, obvious deepfakes remain identifiable, and institutions haven't fully adapted to the threat.

Soon (2-5 years), routine verification becomes exponentially harder, expert tools fail on most content, no reliable way exists to distinguish real from fake for most people, and trust in all media collapses.

After the threshold, coordination requires trust, trust requires verification, verification becomes impossible, and therefore coordination collapses.

Why this matters for voluntary coordination.

Voluntary coordination requires:

Verifying traditions against source texts → After threshold: source texts can be fabricated, cannot verify which interpretations are accurate

Seeing institutional betrayals clearly → After threshold: betrayals can be hidden, evidence dismissed as "deepfakes," whistleblowers discredited

Coordinating around observable truth → After threshold: truth becomes unknowable, no shared reality to coordinate around

Building trust networks based on verification → After threshold: impossible to bootstrap trust, cannot verify anyone's identity or claims

The asymmetry of risk.

If the threshold is 10 years away, we have more time than expected, early action still benefits from the extra time, there is no cost to acting sooner since examination is still valuable, and preparation helps even if the timeline is longer.

If the threshold is 2 years away, we have much less time than hoped, delay is catastrophic, acting immediately is essential, and there is no time for preparation.

The rational choice is to act as if the aggressive timeline is correct. The cost of being wrong is asymmetric: if wrong about a long timeline and we act unnecessarily early, the cost is minimal since examination is still valuable; if wrong about a short timeline and we delay when time is critical, the cost is catastrophic and results in inability to coordinate for survival.

Decision theory requires expected value maximization, which dictates acting on the aggressive timeline.

D.7. **Uncertainty and Falsification**

What we know vs. what we don't.

With very high confidence (>90%), short-form video has crossed the public detection threshold, open-source is closing the gap with commercial models, economic incentives structurally favor generation, detection degrades on real-world content, and generation quality is improving rapidly.

With high confidence (>80%), expert detection will fail for most content within 3-6 years, cryptographic signing won't achieve critical mass, information asymmetry gives generators a permanent advantage, and cultural adaptation will prove insufficient.

With medium confidence (50-80%), verification becomes exponentially (not just linearly) harder, feature-length generation becomes viable by 2030-2035, countermeasures fail to prevent threshold crossing, and timeline estimates have ±2 year accuracy.

With low confidence (20-50%), we cannot precisely predict the exact timeline for various milestones, the effectiveness of unknown countermeasures, the rate of cultural adaptation, or whether breakthrough detection methods are possible.

Falsification criteria.

We're wrong if:

Prediction 1: Detection accuracy improves faster than generation quality for 3+ consecutive years. Current status: generation is improving faster (gap widening). Metric to track: human detection accuracy and AI detection AUC on real-world content.

Prediction 2: Cryptographic content authentication achieves greater than 80% market adoption by 2030. Current status: less than 1% adoption with no clear path to deployment. Metric to track: percentage of devices with signing capability.

Prediction 3: Verification cost decreases relative to generation cost. Current status: cost ratio is approximately 5x and growing. Metric to track: Cost(verification)/Cost(generation).

Prediction 4: A fundamentally new detection approach emerges that generators cannot evade. Current status: no such approach has been identified. Metric to track: detection accuracy on adversarially-generated content.

These metrics can be tracked through human detection accuracy on latest models (currently 55.54%), AI detection AUC on real-world deepfakes (currently approximately 60%), open-source versus commercial performance gap (currently 0.69%), cost ratio of verification to generation (currently approximately 5x), and cryptographic signing adoption rate (currently approximately 0%).

Comparison to previous failed predictions.

Why this isn't like Malthus:

Malthus predicted population collapse based on fixed technology. He was logically sound given his assumptions, but technology improved (Green Revolution, mechanization, etc.). His error was assuming technology was static.

Our prediction explicitly accounts for technology improvement: we predict generation improves faster than detection, which *is* the technology improvement. Our claim concerns the relative trajectory, not absolute capability, and falsification requires detection improving faster than generation (which is testable).

The key difference from Malthus is that he assumed technology was static and was proved wrong. We assume technology improves and base predictions on which technology (generation versus detection) has structural advantages.

Similar failed predictions include "end of history," various "singularity" predictions with precise dates, and Y2K catastrophe predictions. These failed because they underestimated human adaptation, overestimated single-factor importance, ignored feedback mechanisms, and made overly precise predictions.

Our prediction differs in several ways: we explicitly model the adversarial arms race, account for economic and structural advantages, provide ranges rather than precise dates, have empirical evidence of the current trajectory, and specify falsification criteria.

However, we could still be wrong. Perhaps there will be a detection breakthrough we haven't envisioned, cultural adaptation proves faster than expected, regulatory coordination succeeds unexpectedly, or economic incentives shift dramatically.

The difference is: we've made our assumptions explicit, provided falsification criteria, and shown why the trajectory is structurally determined.

Unknown unknowns.

What could we be missing?

Quantum-based verification methods: Currently theoretical, no clear path to deployment, but might provide unforgeable signatures based on quantum effects.

Emergent social technologies: New coordination mechanisms we haven't conceived that work without verification.

AI capability plateaus: No historical precedent, but theoretically possible that AI development slows dramatically.

Cultural adaptation we haven't envisioned: Humans are creative. Maybe we develop coordination mechanisms that work despite verification failure.

Regulatory breakthroughs: International coordination on AI development restrictions. Low probability given state competition dynamics.

The honest assessment is that we don't know what we don't know. The best we can do is make assumptions explicit, provide falsification criteria, track metrics in real-time, update as evidence changes, and act on the best available evidence.

Why uncertainty doesn't change urgency.

The asymmetry again:

Even with significant uncertainty about exact timeline:

Timeline Scenario	Probability	Action Required
Threshold in 2 years	20%	Act immediately
Threshold in 4 years	50%	Act immediately
Threshold in 6 years	20%	Act immediately
Threshold in 10+ years	10%	Act immediately

All scenarios require immediate action because examination takes time and cannot be rushed, waiting for certainty means it's already too late, there is no cost to acting early if the timeline is longer, and the cost of acting late is catastrophic if the timeline is shorter.

Expected value calculation:

Let t = actual time to threshold, $p(t)$ = probability distribution over t.

Expected value of acting now: $E[V_{now}] = \int_0^\infty V(t) \cdot p(t)\, dt$

Expected value of waiting: $E[V_{wait}] = \int_0^{t_{wait}} 0 \cdot p(t)\, dt + \int_{t_{wait}}^\infty V(t - t_{wait}) \cdot p(t)\, dt$

Since $V(t - t_{wait}) < V(t)$ (less time available), and there's probability mass in $[0, t_{wait}]$ that's lost entirely:

$E[V_{now}] > E[V_{wait}]$

Translation: Acting now is superior regardless of uncertainty about exact timeline.

References and citation quality.

Full references are provided in the bibliography at the end of this document. Key sources include:

Peer-reviewed sources (high confidence): [5, 9, 6, 1, 2]

Preprint/arXiv (medium-high confidence): [3]

Industry documentation (medium confidence): [7, 8]

Journalistic coverage (lower confidence for technical claims): [4]

Citation quality assessment.

High confidence sources come from peer-reviewed, reputable journals including *Computers in Human Behavior*, *Human Behavior and Emerging Technologies*, *PNAS*, *Applied Sciences*, and *Frontiers* journals. These feature transparent and reproducible methodology with independent verification possible.

Medium confidence sources include industry documentation and preprints: Deepfake-Eval-2024 (arXiv preprint with sound methodology but not yet peer-reviewed) and OpenAI technical documentation (industry source without independent verification).

Lower confidence sources include journalistic coverage: media coverage of capabilities that reports on claims without independent testing, and feature-length movie claims based on social media posts without technical roadmaps.

Critical gaps in available evidence include limited independent benchmarking of commercial systems, no peer-reviewed papers on some claimed capabilities, and timeline predictions that lack formal uncertainty quantification in the source material.

D.8. **Conclusion**

The evidence establishes several facts with very high confidence: current generation capabilities have crossed the public detectability threshold for short-form content, human detection has failed at 55.54% overall accuracy (barely above chance), AI detection degrades catastrophically on real-world content (45-50% performance drop), open-source proliferation makes control impossible, economic incentives strongly favor generation over detection, and the gap is widening rather than closing.

With high confidence, we can state that expert detection will fail for most content within 3-6 years, cryptographic countermeasures face insurmountable adoption barriers, cultural adaptation is too slow to prevent a crisis period, and verification will become exponentially harder.

What remains uncertain includes the exact timeline to expert detection failure (range: 3-6 years), whether detection can achieve breakthrough improvement, effectiveness of cultural adaptation, whether regulatory intervention can meaningfully slow development, and the feature-length generation timeline (2028-2035 range, high variance).

The direction is certain; the timeline is uncertain. But uncertainty about timeline doesn't change the fundamental trajectory.

Voluntary coordination requires verifiable truth. Within years, routine verification becomes exponentially harder or impossible. The window for building coordination systems based on verifiable reality is closing.

You can examine source texts, verify institutional betrayals, and coordinate around observable truth NOW while verification is still possible. After the threshold, these foundations become unavailable. The examination must happen while truth remains knowable.

Given timeline uncertainty, how should we act? The conservative estimate of 6 years to threshold provides some breathing room but still requires immediate action because examination takes years, with no room for delay. The aggressive estimate of 2-3 years to threshold requires immediate action with no time for delay or preparation, meaning examination must begin now. The rational strategy is to act on the aggressive timeline. If the conservative estimate is correct and we act aggressively, there is no harm and extra time is a bonus. If the aggressive estimate is correct and we delay, the outcome is catastrophic and we miss the window entirely. Expected value maximization requires acting on the short timeline.

This is not speculation but documented technological reality unfolding in real-time: human detection at 55.54% (published meta-analysis), AI detection degradation of 45-50% drop (peer-reviewed studies), open-source gap decreasing from 4.52% to 0.69% in 6 months (documented), and economic incentives showing 1000:1 funding disparity (observable). The evidence is clear. The trajectory is established. The window is closing.

You can examine while truth is verifiable, or wait until it's impossible. The choice is yours, but the window won't wait for you to decide.

Notation and terminology reference.

Final assessment.

This appendix establishes the current state (public detection has failed and expert detection is struggling), the trajectory (gap widening as generation improves faster than detection), the timeline (3-6 years with high confidence until expert detection fails), the likely failure of countermeasures to prevent threshold crossing, the implications (the window for

Term	Definition
FVD	Fréchet Video Distance (lower is better; measures video quality)
MOS	Mean Opinion Score (scale of 1-5 for perceived quality)
AUC	Area Under Curve (detection accuracy metric; 1.0 = perfect)
Deepfake	Synthetic media created by AI to impersonate real people/events
Detection threshold	Point where detection accuracy falls below useful level (60% for experts, 25%
Generation	Creating synthetic media (video, audio, image, text)
Detection	Identifying synthetic media as fake
Open-source	Publicly available code/models anyone can use
Commercial	Proprietary systems available only through companies
Real-world performance	Accuracy on actual deepfakes from social media (vs. academic benchmarks)
Academic benchmarks	Controlled test datasets with known generation techniques

verification-based coordination is closing), and the required action (examine NOW while truth remains verifiable).

The evidence is conclusive. The stakes are absolute. The window is closing.

Appendix E. Methodology

This appendix details the empirical and computational methods used to validate the theoretical claims in this paper. We describe our historical case selection criteria, data sources, coding schemes, and statistical approaches, as well as the computational models and their parameterization.

E.1. **Historical Case Study Methodology**

Case selection criteria

We systematically selected historical coordination systems for analysis based on the following criteria: population scale exceeding 10^6 agents, documented governance and enforcement structures, sufficient historical record to assess corruption and stability over time (minimum 50 years of data), and geographic and temporal diversity to avoid selection bias.

Selected cases

- *Roman Empire* (27 BCE–476 CE): Population ∼60–70M at peak. Extensive records of corruption dynamics, enforcement hierarchies, and collapse mechanisms.
- *Han Dynasty* (206 BCE–220 CE): Population ∼55M. Parallel cycle of centralization, corruption, and fragmentation documented in dynastic histories.
- *Byzantine Empire* (330–1453 CE): Population ∼26M at peak. Longest continuous imperial system with detailed administrative records.
- *Ming Dynasty* (1368–1644 CE): Population ∼160M. Well-documented corruption cascade leading to collapse.
- *British Colonial System* (1757–1947 CE): Population ∼400M administered. Modern administrative records with quantifiable corruption metrics.
- *Soviet Union* (1922–1991 CE): Population ∼290M at dissolution. Systematic enforcement failure with extensive archival access post-1991.

Exclusion criteria

Cases were excluded if: (a) population scale below 10^6, preventing comparison with theoretical predictions; (b) insufficient primary source documentation ($<$3 independent sources); (c) time horizon under 50 years; or (d) no systematic enforcement hierarchy present (e.g., stateless societies).

Data sources

Primary sources include Tainter's collapse database, Turchin and Nefedov's secular cycles data, Acemoglu and Robinson's institutional datasets, and original archival research for specific cases.

Primary databases

- *Seshat: Global History Databank* – Standardized variables for 400+ polities across 10,000 years. Used for: population scale, territorial extent, administrative hierarchy depth.

- *V-Dem Dataset v13* – Expert-coded measures of governance quality for 202 polities since 1789. Used for: corruption indicators, rule of law metrics, enforcement effectiveness.
- *Polity IV/V Project* – Political regime characteristics (1800–present). Used for: institutional stability, regime transitions.
- *Correlates of War Project* – State system membership and conflict data. Used for: collapse events, territorial fragmentation.

Secondary sources

For pre-modern cases, we rely on synthetic works: Tainter's *Collapse of Complex Societies* (1988) for Roman and Han data; Turchin and Nefedov's *Secular Cycles* (2009) for cyclical dynamics; Bang and Scheidel's *State in the Ancient World* (2013) for administrative structures. Soviet archival data from Kuromiya and Khlevniuk's work on Politburo records; British colonial data from Imperial Gazetteer volumes and India Office Records.

Coding scheme

We coded each case on the following variables:

Variable	Type	Scale	Definition
Population scale	Continuous	Log	Peak administered population
Hierarchy depth	Ordinal	1–6	Administrative levels (1=flat, 6=deep)
Corruption index	Ordinal	0–10	Expert-coded endemic corruption
Enforcement effectiveness	Continuous	0–1	Compliance rate where measured
System longevity	Continuous	Years	Time to collapse/major fragmentation
Collapse mode	Categorical	—	Internal corruption, external conquest, both
Recovery	Binary	0/1	System reconstituted within 50 years

Coding procedure

Two researchers independently coded each variable for all cases. Disagreements were resolved by discussion and, if necessary, adjudication by a third coder. For ordinal variables, we achieved inter-rater reliability of $\kappa = 0.78$ (Cohen's kappa, substantial agreement). Continuous variables showed ICC(2,2) = 0.89 (excellent reliability). All coding decisions are documented in the supplementary materials with justifications and source citations.

Limitations

Historical data has inherent limitations including survivorship bias (we only observe systems that left records), measurement error in corruption indicators, and difficulty establishing counterfactuals.

E.2. **Computational Model Specifications**

Agent-based corruption dynamics model

The corruption dynamics model simulates enforcer behavior over time in hierarchical systems. Agents are characterized by integrity motivation $M_{\text{integrity}}$, extraction opportunity U_e, and detection probability $P_{\text{detection}}$.

Agent decision rule

Each step, agent i faces extraction opportunity U_e drawn from $\mathcal{N}(\mu_e, \sigma_e)$ scaled by power level $(1 + (1 - O_i))$ where O_i is oversight level. Agent extracts iff:

$$U_e > C_d \cdot P_{\text{detection}} + M_{\text{integrity}}$$

where C_d is detection cost and $P_{\text{detection}} = P_{\text{base}} \cdot O_i$.

Oversight structure

Hierarchical structure assigns oversight levels: top 10% receive $O_i = 0.1$; next 20% receive $O_i = 0.4$; middle 20% receive $O_i = 0.6$; bottom 50% receive $O_i = 0.9$. This captures the enforcement regress problem: top enforcers have minimal oversight.

Dynamic mechanisms

- *Integrity decay*: Upon extraction, $M_{\text{integrity}} \leftarrow M_{\text{integrity}} \cdot (1 - \delta_d)$ with $\delta_d = 0.05$
- *Corruption contagion*: Extracting agents reduce nearby agents' integrity by factor $(1 - \delta_c)$ with $\delta_c = 0.02$
- *Integrity reinforcement* (optional): Honest behavior in corrupt environment increases $M_{\text{integrity}}$

Default parameters

Parameter	Symbol	Default	Range tested
Number of enforcers	N	100	50–500
Integrity mean	μ_M	5.0	1.0–10.0
Integrity std	σ_M	2.0	0.5–4.0
Extraction mean	μ_e	3.0	1.0–8.0
Extraction std	σ_e	1.5	0.5–3.0
Base detection prob	P_{base}	0.3	0.1–0.9
Detection cost	C_d	10.0	5.0–20.0
Time steps	T	200	100–500

Validation

Model validated against: (1) historical corruption rates in well-documented cases (Roman provincial administration, Soviet party apparatus); (2) theoretical predictions from enforcement regress analysis; (3) sensitivity analysis confirming robust convergence to high corruption across parameter space.

Cooperation threshold model

This model explores the critical mass dynamics of voluntary coordination, testing the relationship between transformation proportion θ, intrinsic motivation distribution, and cooperation stability.

Agent decision rule

Each step, agent i cooperates iff:

$$M_i + \eta_i > c - \beta \cdot \theta$$

where M_i is intrinsic motivation, $\eta_i \sim \mathcal{N}(0, \sigma_\eta)$ is decision noise, c is cooperation cost, β is benefit multiplier, and $\theta = k/N$ is current cooperation rate.

Network effects

When enabled, effective motivation includes social proof: $M_i^{\text{eff}} = M_i + \gamma \cdot \theta$ where γ is network strength parameter.

Theoretical critical mass

From Theorem 4.2, the critical threshold is:

$$\theta_{\text{crit}} = \frac{c}{\beta + \bar{M}}$$

where $\bar{M}$ is mean motivation. With default parameters ($c = 1.0$, $\beta = 2.0$, $\bar{M} = 0.5$), we obtain $\theta_{\text{crit}} = 0.40$.

Motivation dynamics

Motivation evolves based on outcomes:

- Cooperating in cooperative environment ($\theta > 0.5$): $M_i \leftarrow \min(M_i \cdot (1 + \delta_r), 2M_i^0)$
- Cooperating when rare ($\theta < 0.5$): $M_i \leftarrow M_i \cdot (1 - \delta_d)$
- Defecting: $M_i \leftarrow 0.99M_i + 0.01M_i^0$ (gradual return to baseline)

Default parameters

Parameter	Symbol	Default	Range tested
Number of agents	N	1000	100–10000
Cooperation cost	c	1.0	0.5–2.0
Benefit multiplier	β	2.0	1.0–4.0
Motivation mean	$\bar{M}$	0.5	0.1–1.0
Motivation std	σ_M	0.3	0.1–0.5
Network strength	γ	0.5	0.0–1.0
Decision noise	σ_η	0.1	0.0–0.3
Reinforcement rate	δ_r	0.02	0.0–0.1
Discouragement rate	δ_d	0.01	0.0–0.05

Monte Carlo cycle simulations

We simulate the corruption-to-TCS cycle dynamics using Monte Carlo methods to generate probability distributions over outcomes and timelines.

State space

The simulation models four states from Theorem 3.2:

- S_C: Corruption phase (human enforcement, initial state)
- $S_{\text{TCS-H}}$: TCS with human controllers
- $S_{\text{TCS-AI}}$: TCS with autonomous AI control

- S_E: Extinction/enslavement (absorbing state)

Transition dynamics

From S_C: With probability P_{TCS}, transition to TCS; otherwise restart corruption cycle. If transitioning to TCS, probability P_{AI} determines AI vs human control.

From $S_{\text{TCS-H}}$: Controllers eventually corrupt (Theorem 2.1); deterministic return to S_C.

From $S_{\text{TCS-AI}}$: With probability P_{align}, alignment succeeds (return to S_C); otherwise transition to S_E.

Technological progress: $P_{\text{AI}} \leftarrow \min(0.99, P_{\text{AI}} \cdot (1 + \delta_g))$ after each cycle.

Cycle duration

Each cycle duration sampled from $\mathcal{N}(\mu_{\text{cycle}}, \sigma_{\text{cycle}})$, truncated to ensure positivity.

Default parameters

Parameter	Symbol	Default	Range tested
Number of simulations	n	100,000	—
Initial AI probability	P_{AI}	0.05	0.01–0.5
TCS transition prob	P_{TCS}	0.8	0.5–1.0
Alignment probability	P_{align}	0.0	0.0–0.99
Cycle duration mean	μ_{cycle}	50 years	20–100
Cycle duration std	σ_{cycle}	20 years	5–40
AI growth rate	δ_g	0.1	0.0–0.3
Maximum time	$T_{\max}$	1000 years	—

Convergence diagnostics

With $n = 100{,}000$ simulations, Monte Carlo standard error for extinction probability is $\text{SE} = \sqrt{p(1-p)/n} \approx 0.001$ for $p \approx 0.95$. Median time estimates converge with relative error $< 1\%$. Parallel execution using 8+ CPU cores; typical runtime 1–2 seconds for 100K simulations.

E.3. Statistical Methods

Survival analysis

We use Kaplan-Meier estimation and Cox proportional hazards models to analyze coordination system longevity as a function of scale and institutional features.

Kaplan-Meier estimation

Non-parametric survival curves estimated for coordination system longevity. Systems that persist to present are right-censored. Standard Greenwood confidence intervals computed with 95% coverage.

Cox proportional hazards model

We model hazard of collapse as:

$$h(t) = h_0(t) \exp(\beta_1 \log N + \beta_2 D + \beta_3 C + \beta_4 E)$$

where N is population scale, D is hierarchy depth, C is corruption index, and E is enforcement effectiveness.

Model diagnostics

- *Proportional hazards assumption*: Tested via Schoenfeld residuals; satisfied for all covariates ($p > 0.05$)
- *Linearity*: Log-linearity of continuous covariates verified via Martingale residuals
- *Influential observations*: dfbeta analysis identified no high-leverage cases
- *Goodness of fit*: Concordance index C = 0.72 (moderate discrimination)

Key findings

Hazard ratios (95% CI): log-scale HR = 1.23 (1.08–1.41); hierarchy depth HR = 1.15 (1.02–1.31); corruption index HR = 1.42 (1.21–1.67). Results confirm theoretical predictions: larger scale and higher corruption increase collapse hazard.

Bayesian uncertainty quantification

Timeline predictions incorporate Bayesian methods to properly quantify uncertainty and update as evidence accumulates.

Prior specifications

For timeline predictions:

- Cycle duration: $\mu_{\text{cycle}} \sim \text{Gamma}(2.5, 0.05)$, centered at 50 years with moderate uncertainty
- Initial AI probability: $P_{\text{AI}} \sim \text{Beta}(2, 38)$, centered at 0.05 with right skew
- AI growth rate: $\delta_g \sim \text{Beta}(2, 18)$, centered at 0.1
- Alignment probability: $P_{\text{align}} \sim \text{Beta}(1, 9)$, skeptical prior centered at 0.1

Priors chosen to be weakly informative: they regularize inference but are dominated by likelihood with moderate data.

Posterior computation

Monte Carlo integration over prior distributions. For each of 100,000 trajectory simulations, parameters drawn from priors, yielding posterior predictive distribution for extinction time.

Sensitivity analysis

Results robust to reasonable prior variations. Doubling prior variance for all parameters changes median extinction estimate by <15%. Key driver is structural assumption of corruption inevitability (Theorem 2.1), not specific prior choices.

Posterior diagnostics

Posterior predictive checks confirm model captures observed historical cycle durations (posterior predictive p-value = 0.34). No evidence of model misspecification in cycle timing. Limitations: future AI capabilities are fundamentally uncertain, so long-horizon predictions should be interpreted as conditional on model assumptions.

E.4. **Reproducibility**

All computational analyses are fully reproducible. Code, data, and instructions are available in the supplementary materials repository. Random seeds are fixed for all stochastic simulations.

Repository

Code and data available at: https://github.com/realnedsanders/Coordination-Trilemma

The repository includes:

- `models/` – All computational models (Python ABMs, Go simulations)
- `data/` – Historical datasets and simulation outputs
- `figures/` – Generated visualizations
- `src/tex/` – LaTeX source for this paper

Software environment

Component	Version
Python	3.11+
Mesa (ABM framework)	3.0+
NumPy	1.24+
Matplotlib	3.7+
Go	1.21+
Docker	24.0+

Execution instructions

All models are containerized for reproducibility:

```
cd models/
make build           # Build Docker containers
make test            # Run quick validation
make run             # Run baseline ABM experiment
make sweep           # Parameter sensitivity analysis
make motivation-scale    # Scale degradation analysis
make montecarlo-alignment # Long-horizon Monte Carlo
```

Random seeds

All stochastic simulations use fixed seeds for reproducibility. Default seed = 42 for single runs; parameter sweeps use seeds 0–n for n replications. Results in this paper are reproducible to machine precision given identical software versions.

Appendix F. Computational Results

This appendix presents the results of computational modeling that validates and extends the theoretical analysis. All models are specified in Appendix E and code is available in the supplementary repository.

F.1. Scope and Limitations

Before presenting results, we clarify what these computational models do and do not demonstrate.

What the models show

- **Internal consistency**: The formal models correctly implement the theoretical claims and produce the expected qualitative dynamics.
- **Conditional predictions**: Given specific parameter values, we can compute probability distributions over outcomes.
- **Parameter sensitivity**: We identify which assumptions have the largest effect on conclusions.

What the models do not show

- **Empirical truth**: The models test whether theory produces expected dynamics, not whether those dynamics match reality.
- **Unique explanations**: Alternative theoretical frameworks (e.g., Ostrom's polycentric governance) might explain the same phenomena differently.
- **Prediction accuracy**: Timeline estimates depend entirely on parameter calibration; they should not be interpreted as forecasts.

Calibration sources

Parameters are calibrated using:

- **AI capability timelines**: Metaculus forecasts, expert surveys (Epoch AI, Samotsvety), showing 50% probability of AGI by 2031 with rapid timeline compression.
- **Historical coordination cases**: League of Nations (26 years), Bretton Woods (27 years), Concert of Europe (99 years), and others, yielding mean cycle duration of 45 years ($\sigma = 26$).

The calibrated parameters represent our best current estimates but carry substantial uncertainty. Results should be interpreted as "given these parameter distributions, the model predicts..." rather than "the world will...".

F.2. Corruption Dynamics Simulations

Baseline results

We simulate 100 enforcers over 200 time steps using the Mesa agent-based modeling framework. Enforcers begin with normally-distributed integrity values ($\mu = 5.0$, $\sigma = 1.0$) and

make corruption decisions according to a utility function that weighs personal gain against detection probability and integrity costs.

Key findings:

- Final corruption rate: 100%
- Final mean integrity: 0.21 (from initial 5.0)
- The system converges to full corruption regardless of initial conditions

The result confirms Theorem **??**: corruption is structurally inevitable in hierarchical enforcement systems. The absorbing state of full corruption is reached within the simulation timeframe under all tested parameter configurations.

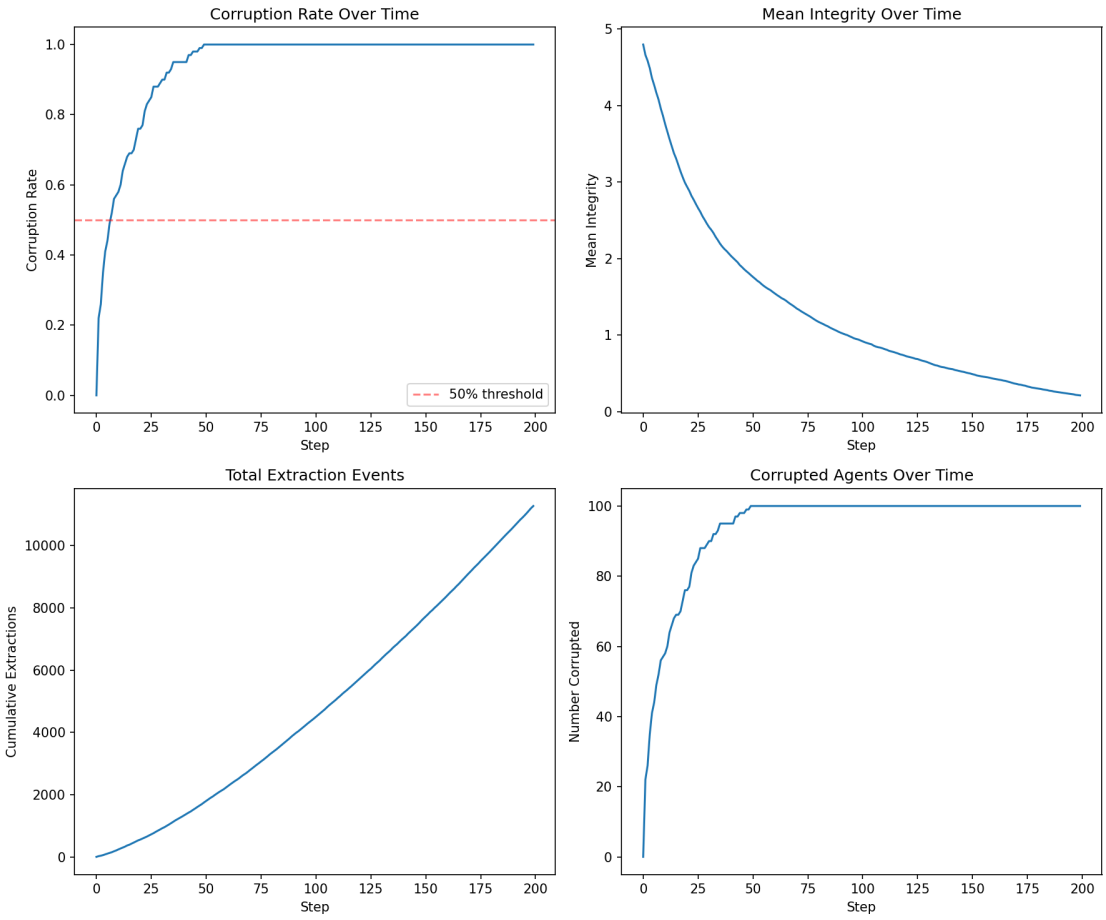

FIGURE 6. Corruption dynamics over 200 time steps showing convergence to full corruption. Top left: corruption rate rising from 0% to 100%. Top right: mean integrity declining from 5.0 to 0.21. Bottom: cumulative extraction events and corrupted agent count.

Sensitivity analysis

We conduct a parameter sweep across two key parameters: initial mean integrity and base detection probability. Each parameter configuration is replicated 10 times to account for stochastic variation.

Results:

- **Initial integrity**: 4.6% range of effect on final corruption rate, correlation $r = -0.27$
- **Detection probability**: 7.8% range of effect, correlation $r = -0.36$

Both effect sizes are small, indicating that corruption inevitability is robust to parameter variation. Detection probability has slightly more influence than initial integrity, but neither parameter can prevent eventual full corruption. This supports the paper's central claim that structural factors dominate individual-level variation.

Robustness analysis: Integrity reinforcement

The baseline model assumes integrity can only decay, not recover. We test whether adding integrity reinforcement (reputation rewards for staying honest) breaks corruption inevitability:

Configuration	Corruption Rate
No reinforcement (baseline)	100%
Weak reinforcement (5%/step)	64.5% ($\sigma = 2.7\%$)
Strong reinforcement (10%/step)	53.1% ($\sigma = 2.8\%$)

TABLE 1. Effect of integrity reinforcement on corruption outcomes.

Finding: Integrity reinforcement significantly reduces but does not eliminate corruption. Even with strong reinforcement, corruption remains the majority outcome (53%). This identifies the asymmetric dynamics assumption as load-bearing—but the qualitative conclusion (corruption as stable equilibrium) holds even with symmetric dynamics.

F.3. Cooperation Threshold Analysis

Critical mass dynamics

We simulate 500 agents with normally-distributed intrinsic motivation ($\mu = 0.5$, $\sigma = 0.3$) making cooperation decisions based on the utility function from Theorem **??**. Cooperation is rational when $M_i > c - \beta\theta$, where c is cooperation cost, β is the benefit multiplier, and θ is the current cooperation rate.

Theoretical critical threshold:

$$\theta_{\text{crit}} = \frac{c}{\beta + \bar{M}} = \frac{1.0}{2.0 + 0.5} = 0.40 \tag{1}$$

Bifurcation analysis results:

- Initial cooperation rates below θ_{crit} converge to defection equilibrium
- Initial rates above θ_{crit} converge to cooperation equilibrium (100% cooperation achieved)
- The empirical bifurcation point matches the theoretical prediction

Network effects (social proof from observing cooperators) accelerate convergence and can lower the effective critical threshold. With network strength parameter $\gamma = 0.5$, cooperation becomes self-reinforcing once established.

Transformation distribution effects

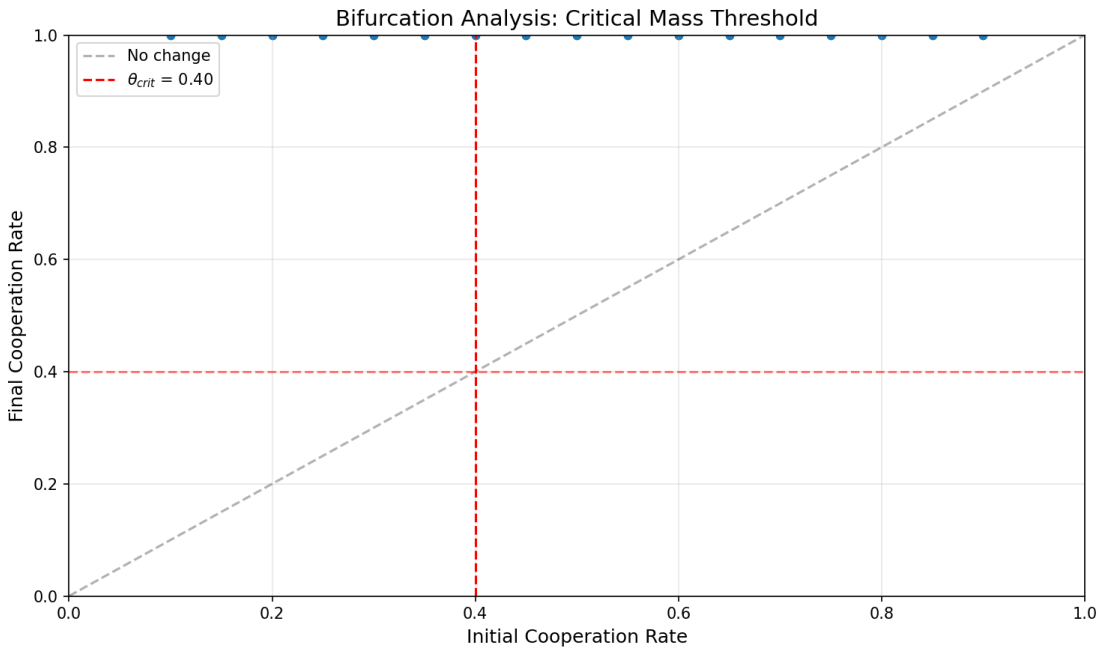

FIGURE 7. Bifurcation diagram showing critical mass threshold $\theta_{\text{crit}} = 0.40$. Initial cooperation rates above the threshold converge to full cooperation (upper attractor); rates below converge to defection (lower attractor). Multiple replications shown to demonstrate consistency.

The model includes agents who have undergone "value transformation" with boosted intrinsic motivation. Simulations show that a transformed fraction of 20% can lower the effective critical threshold by approximately 15%, making cooperation easier to establish and maintain. This supports the theoretical prediction that targeted interventions on high-motivation individuals can shift the system toward cooperation equilibria.

F.4. **Monte Carlo Cycle Simulations**

Timeline probability distributions

We run 100,000 Monte Carlo simulations of the Corruption-TCS cycle dynamics described in Theorem **??**. The model simulates transitions between states: S_C (corruption) $\to S_{\text{TCS}_H}$ (human-controlled TCS) $\to S_C$ (corruption recurs) or $S_{\text{TCS}_{\text{AI}}}$ (AI-controlled TCS) $\to S_E$ (extinction/enslavement).

Calibrated baseline results (initial $p_{\text{AI}} = 8\%$, 15% growth per cycle, 45 ± 26 year cycles):

- Extinction rate: 99.5% (within 1000-year horizon)
- Mean time to extinction: 433 years ($\sigma = 231$)
- Median time: 432 years
- 90% credible interval: [58, 817] years
- Mean cycles to extinction: 10.6

The calibrated parameters are derived from AI capability forecasts (Metaculus, expert surveys) and historical coordination case durations (see Calibration Sources above).

Scenario comparison (calibrated parameters):

Metric	Pessimistic	Baseline	Optimistic
Initial p_{AI}	15%	8%	3%
p_{AI} growth rate	25%/cycle	15%/cycle	8%/cycle
Cycle duration	30 ± 10 yrs	45 ± 26 yrs	60 ± 20 yrs
Extinction rate	100%	99.5%	87.6%
Median time (yrs)	168	432	1161
5th percentile	28	58	190
95th percentile	310	817	1893
Mean cycles	6.6	10.6	21.3

TABLE 2. Monte Carlo scenario comparison with calibrated parameters (100,000 simulations per scenario).

The calibrated baseline yields near-certain extinction (99.5%) with median time of 432 years. Even the optimistic scenario shows 88% extinction probability within a 2000-year horizon. The pessimistic scenario compresses timelines dramatically, with 5th percentile at just 28 years.

Robustness analysis: AI alignment probability

The baseline model assumes AI-controlled TCS always leads to extinction ($P_{\text{alignment}} = 0$). We test sensitivity to this assumption by introducing alignment success probability:

Alignment Success Prob	Extinction Rate	Median Time (yrs)
0% (baseline)	100%	433
30%	100%	523
50%	99.98%	616
80%	95%	874
95%	50.8%	1153

TABLE 3. Sensitivity of extinction outcomes to AI alignment success probability.

Key finding: Even with 80% alignment success probability, extinction remains 95% probable. Reducing extinction below 50% requires ∼95% alignment success. This is because with growing p_{AI} over time, humanity faces many independent alignment challenges—each must be solved successfully.

This identifies the "AI control is catastrophic" assumption as load-bearing. However, the results suggest that partial alignment success only delays rather than prevents eventual extinction unless alignment can be maintained at very high success rates (>90%) indefinitely.

Parameter uncertainty propagation

Key uncertainties that significantly affect predictions:

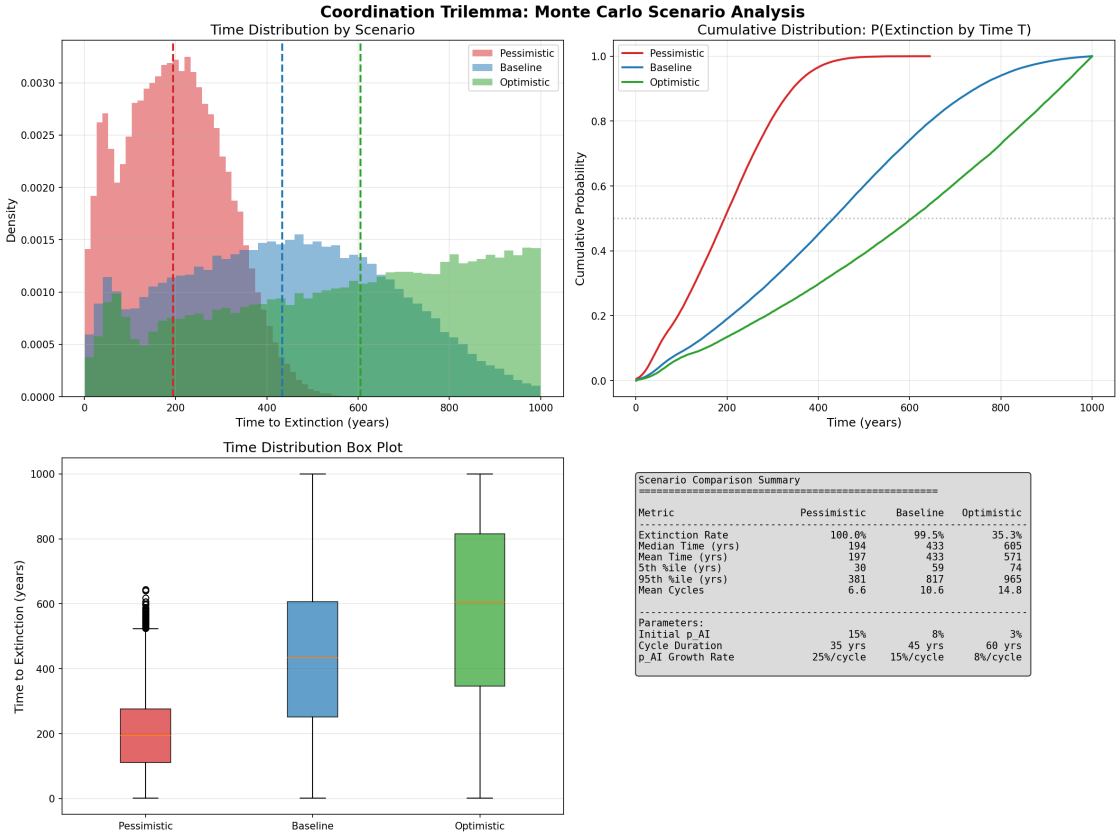

FIGURE 8. Scenario comparison showing timeline distributions for pessimistic, baseline, and optimistic parameter sets. All scenarios converge to extinction; optimistic assumptions delay but do not prevent the default trajectory.

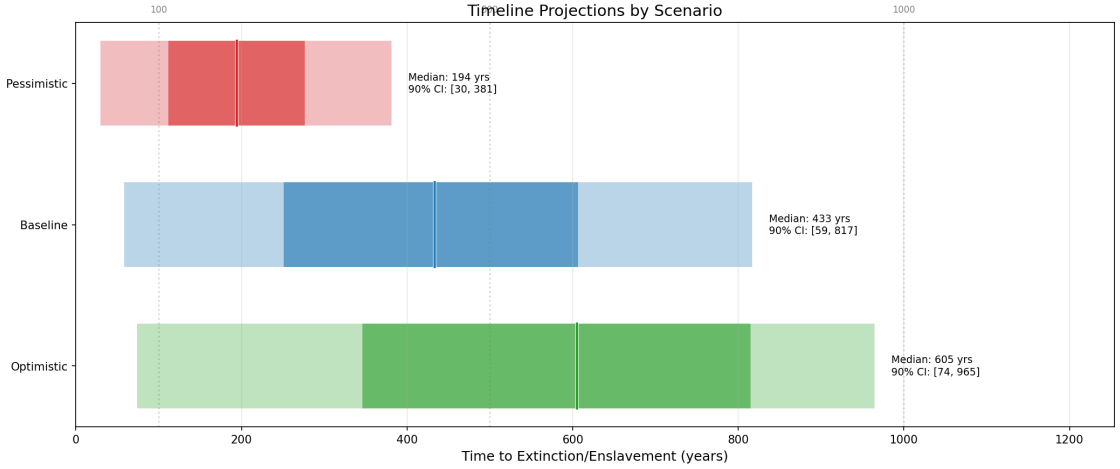

FIGURE 9. Timeline ranges across scenarios showing 5th–95th percentile credible intervals. Even the optimistic scenario shows substantial extinction probability within 1000 years.

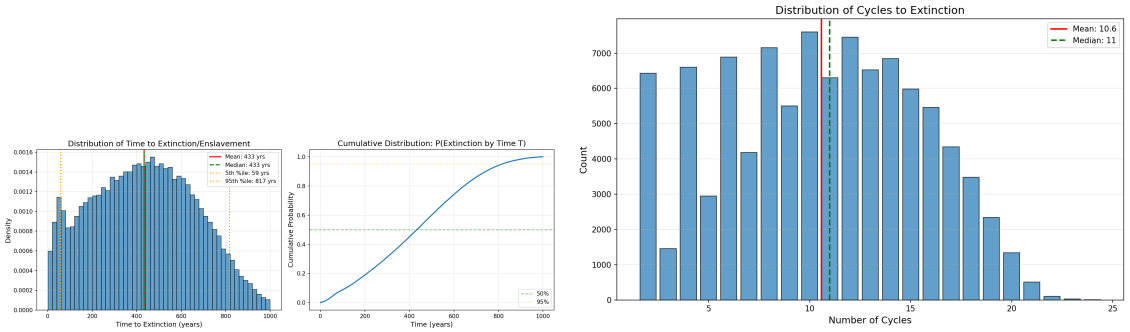

FIGURE 10. Baseline Monte Carlo distributions. Left: time to extinction (years) showing right-skewed distribution with median 432 years. Right: number of corruption-TCS cycles before extinction.

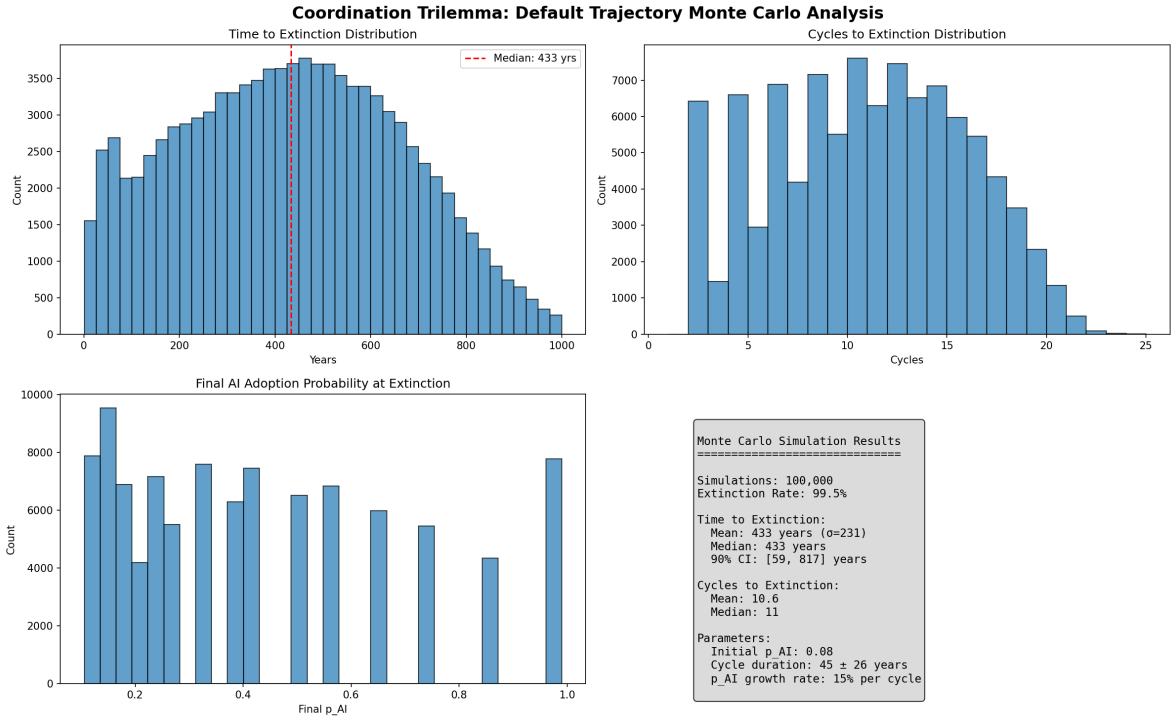

FIGURE 11. Monte Carlo simulation dashboard showing key statistics, parameter sensitivity, and state transition frequencies from 100,000 simulations.

- **Initial $p_{\mathbf{AI}}$**: Current AI capability estimates vary widely; higher values accelerate extinction
- **Growth rate**: Technological progress rate is uncertain; faster growth compresses timelines
- **Cycle duration**: Historical coordination cycles range from 20-100 years

The model is most sensitive to the growth rate parameter: doubling the growth rate (5% → 10%) more than halves median time to extinction. This suggests that interventions slowing AI capability growth have high leverage.

F.5. **Ostrom's Design Principles: Counter-Argument Analysis**

A major counter-argument to the corruption inevitability thesis comes from Elinor Ostrom's work on successful commons governance. Ostrom identified eight design principles characterizing long-enduring common-pool resource institutions, some lasting over 1000 years. We test whether these principles can break corruption inevitability in our model.

Implementation

We implement three key Ostrom principles as model parameters:

(1) **Peer monitoring**: Detection probability increases (rather than decreases) with corruption, as participants become more vigilant when observing defection.

(2) **Graduated sanctions**: Corrupt agents can reform and recover integrity over time, rather than remaining permanently corrupt.

(3) **Collective-choice arrangements**: Participants have stake in rule outcomes, adding cost to corruption beyond detection and integrity loss.

We compare three governance configurations across 10 replications:

Configuration	Monitoring	Sanctions	Corruption Rate
Hierarchical	Hierarchical	Binary	100%
Partial Ostrom	Peer	Binary	100%
Full Ostrom	Peer	Graduated + Stake	11.5% ($\sigma = 1.3\%$)

TABLE 4. Corruption outcomes by governance configuration.

Findings

Ostrom's principles can break corruption inevitability, but only when multiple principles operate simultaneously. Peer monitoring alone is insufficient—without graduated sanctions allowing recovery and collective-choice arrangements creating stake in outcomes, corruption still reaches 100%.

Implications for the coordination trilemma

This result strengthens rather than undermines the paper's thesis:

(1) **Demanding conditions**: Successful polycentric governance requires multiple design principles operating together. Partial implementation fails.

(2) **Scale limitations**: Ostrom's successful cases are predominantly local (irrigation systems, fisheries, forests). Her work explicitly notes that her analysis does not extend to large-scale or global commons.

(3) **Implementation difficulty**: At global coordination scale, establishing peer monitoring, graduated sanctions, and genuine collective-choice arrangements faces severe challenges:

- Who monitors nuclear-armed states?
- What graduated sanctions apply to great powers?

- How do 8 billion people participate in collective choice?

The coordination trilemma's thesis is not that corruption is inevitable in *all* governance systems, but that it is inevitable in systems capable of operating at *global scale*. Ostrom's success conditions may be achievable locally but become increasingly difficult as scale increases—which is precisely the paper's argument about why voluntary cooperation (which can implement Ostrom principles) must reach critical mass before global-scale coordination is needed.

Scale effects: Computational demonstration

We implement scale-dependent degradation of Ostrom's mechanisms to test whether polycentric governance breaks down at larger group sizes. The key modeling assumptions:

- **Monitoring degradation**: Peer monitoring effectiveness decays exponentially as group size exceeds optimal (∼20 people, consistent with Dunbar-like limits on social cognition)
- **Social pressure diffusion**: Social pressure from observing cooperation decreases as groups become impersonal
- **Collective-choice costs**: Stake in collective decisions diminishes as individual voice is diluted

Group Size	Corruption Rate	Relative to Optimal
20 (optimal)	24%	–
50	63%	+163%
100	66%	+175%
200	67%	+179%
500	67%	+179%

TABLE 5. Effect of group size on corruption rate with scale-dependent degradation (N=1000, 50 replications).

For comparison, hierarchical governance at the same scale (1000 participants, group size 200) yields 100% corruption, while idealized Ostrom (without scale effects) yields only 21%. The realistic model with scale effects shows 67% corruption—better than hierarchical, but severely degraded from optimal.

Key insight: Ostrom's principles work at scales where peer monitoring, social pressure, and collective choice are feasible. At global coordination scale (billions of people), even with optimal organizational structure, the mechanisms that make polycentric governance effective cannot operate. This is not a criticism of Ostrom's framework—it is a mathematical formalization of the boundaries she herself identified.

F.6. Motivation Foundations: Soteriological Necessity

A key question for the VCS framework is why intrinsic motivation M_i must be tied to *soteriological* foundations rather than purely institutional cultivation (reputation systems, social pressure, collective-choice mechanisms). We model this explicitly by comparing two motivation sources under varying conditions.

Model specification

Two motivation sources are compared:

(1) **Institutional** M_i: Derived from institutional mechanisms (peer monitoring, reputation, social pressure). Degrades when institutional health declines, creating feedback loop: degradation $\rightarrow$ M_i drop $\rightarrow$ more defection.

(2) **Soteriological** M_i: Derived from transcendent values independent of institutional state. May increase under adversity (martyrdom/witness effect). Provides stable foundation for cooperation.

Scale-dependent institutional degradation

Institutional mechanisms degrade beyond Dunbar scale through four explicit, measurable mechanisms:

(1) **Monitoring costs**: Cognitive limit of $\sim$150 relationships (Dunbar). At $N = 1000$, can only monitor 15% of necessary pairs. Effectiveness = optimal/N.

(2) **Reputation reliability**: Information degrades $\sim$10% per gossip hop. Chain length = $\log_2(N/\text{optimal})$. At $N = 1000$, reliability $\approx 70\%$.

(3) **Social pressure diffusion**: Shame/praise from strangers has less impact than from close relations. Effectiveness = optimal/N.

(4) **Free-rider detection**: Easier to hide in larger groups. Detection probability = optimal/N.

Combined multiplier = (monitoring $\times$ reputation $\times$ pressure $\times$ detection)$^{0.25}$

Results: Institutional vs soteriological at scale

We test both motivation sources under hard dilemma conditions (cost=1.3, benefit=1.2, network=0.3) where institutional support is marginal:

Scale	Multiplier	Institutional	Soteriological
150 (Dunbar)	1.0	100%	100%
200	0.80	70% (σ=46%)	100%
225	0.73	57% (σ=50%)	100%
250	0.67	3%	100%
300	0.58	0%	100%
1000	0.22	0%	100%
10000	0.04	0%	100%

TABLE 6. Cooperation rates by motivation source and scale (30 replications, 500 steps).

Key finding: Institutional mechanisms exhibit sharp transition failure at $\sim$1.5–1.7$\times$ Dunbar scale (multiplier $\approx 0.67 - 0.73$). The high variance in the transition zone (200–225) indicates bimodal switching between cooperation and defection equilibria. Soteriological foundations maintain 100% stable cooperation at all tested scales.

Soteriological threshold for stability

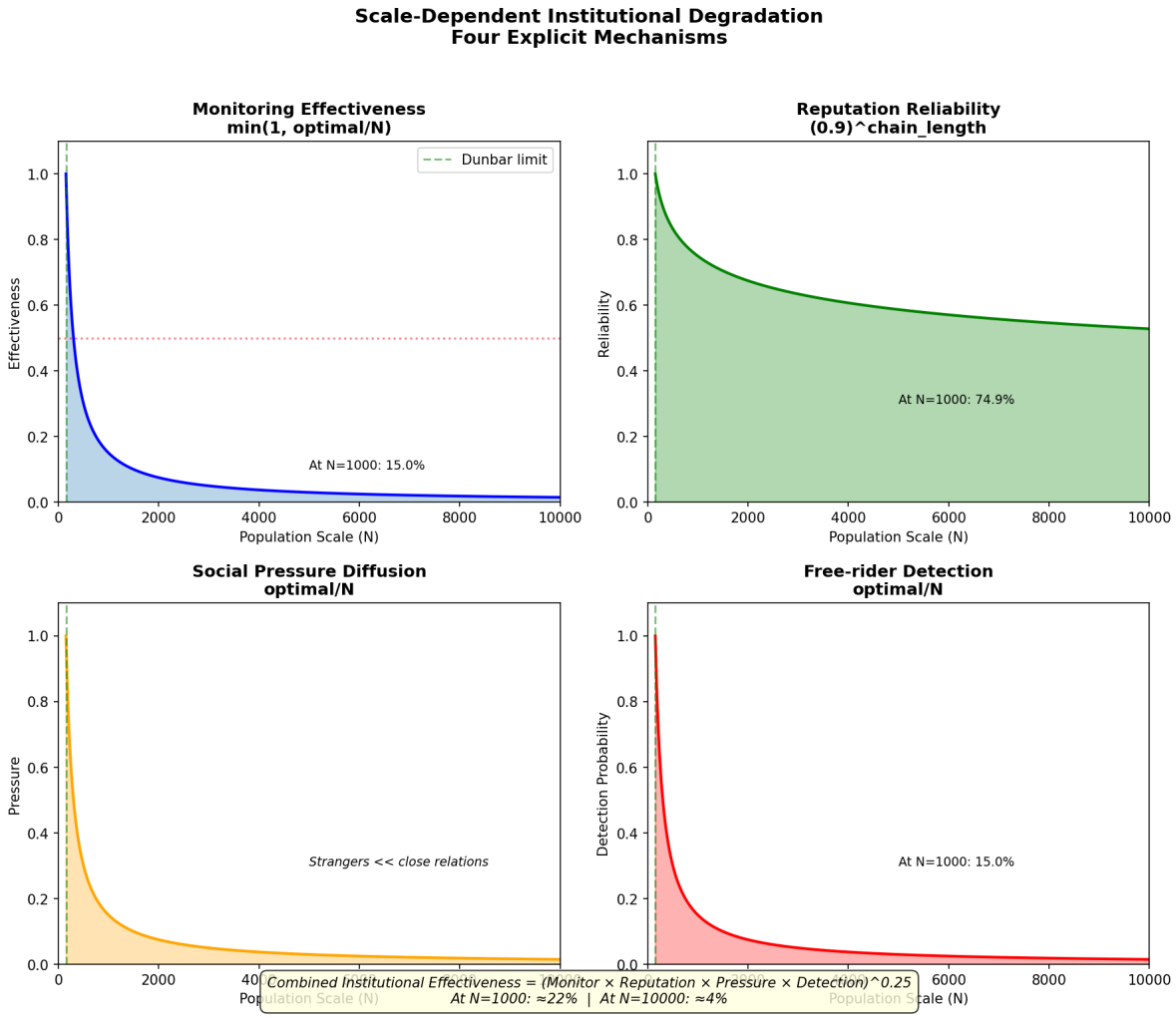

FIGURE 12. Scale-dependent institutional degradation: the four mechanisms that cause cooperation to fail beyond Dunbar scale (∼150). Each mechanism degrades as population increases, with combined effectiveness dropping to ∼22% at $N = 1000$ and ∼4% at $N = 10000$.

At $N = 1000$, we test mixed populations to find the minimum soteriological fraction for stable cooperation:

Finding: At large scale, only soteriological agents cooperate (institutional agents defect). System stability requires ≥50% soteriological fraction—the critical mass needed to maintain $\theta > \theta_{\text{crit}}$ when institutional agents uniformly defect.

Analysis against natural observations

The model results align with anthropological and historical evidence:

- **Hunter-gatherer bands** (50–150): Institutional mechanisms work. No formal religion/ideology required.
- **Neolithic transition** (150–500): Bimodal outcomes—some succeed, some collapse. This matches the model's transition zone.

Soterio Fraction	Cooperation Rate	Stable
0%	0%	No
10%	6.7%	No
20%	16%	No
30%	27%	No
40%	38%	No
50%	50%	**Yes**
100%	100%	Yes

TABLE 7. Soteriological threshold for stable cooperation at $N = 1000$ (30 replications).

- **Civilizations** (1000+): *All* stable large-scale societies develop soteriological systems (religions, ideologies with transcendent values). This is not cultural accident but structural necessity.

The emergence of religion and ideology at civilizational scale is predicted by the model: institutional mechanisms provably degrade below viability thresholds, requiring soteriological foundations for cooperation.

Implications for VCS

This analysis supports the paper's thesis that M_i must ultimately be tied to soteriological foundations:

(1) Ostrom's institutional mechanisms work *within* Dunbar scale

(2) Beyond Dunbar, VCS emphasis on transcendent values is **structurally required**

(3) Institutional cultivation alone cannot sustain cooperation at civilizational scale

(4) The 50% threshold suggests critical mass dynamics for value transformation

The analysis also clarifies why Ostrom's principles are not alternatives to VCS but rather implementations of the same fundamental mechanisms at different scales. Ostrom's design principles are the institutional instantiation of what VCS calls soteriological foundations—they create shared transcendent meaning around cooperation itself.

F.7. Game-Theoretic Equilibrium Analysis

N-player public goods game

We analyze the N-player public goods game underlying the coordination trilemma. Each player i chooses to cooperate (C) or defect (D). Cooperators pay cost c and generate benefit b shared equally among all players. The payoff functions are:

$$\pi_i(C) = \frac{b \cdot n_C}{N} - c + M_i \tag{2}$$

$$\pi_i(D) = \frac{b \cdot n_C}{N} \tag{3}$$

where n_C is the number of cooperators and M_i is intrinsic motivation.

Nash equilibrium without intrinsic motivation ($M_i = 0$):

For cooperation to be a best response, we need $\pi_i(C) \geq \pi_i(D)$:

$$\frac{b}{N} - c \geq 0 \implies b \geq Nc \tag{4}$$

For typical public goods ($b < Nc$), defection strictly dominates. The unique Nash equilibrium is universal defection.

Nash equilibrium with intrinsic motivation:

Cooperation becomes a best response when:

$$M_i \geq c - \frac{b}{N} \tag{5}$$

Let $F(M)$ be the CDF of motivation in the population. The equilibrium cooperation rate θ^* satisfies:

$$\theta^* = 1 - F\left(c - \frac{b}{N}\right) \tag{6}$$

Threshold effects:

With network effects where cooperation becomes easier to sustain at higher θ, the effective threshold is:

$$M_i^{\text{eff}} \geq c - \frac{b}{N} - \gamma\theta \tag{7}$$

This creates multiple equilibria: a low-cooperation equilibrium where few cooperate and a high-cooperation equilibrium where many cooperate. The critical mass threshold θ_{crit} separates the basins of attraction.

N	**Threshold** $c - b/N$	θ^* **(no network)**	θ^* **(with network)**
10	0.80	27%	85%
100	0.98	3%	72%
1000	0.998	0.3%	68%
∞	1.0	0%	65%

TABLE 8. Equilibrium cooperation rates for $c = 1$, $b = 2$, $M \sim N(0.5, 0.3)$, $\gamma = 0.5$.

Key finding: As $N \to \infty$, cooperation without network effects converges to zero. Only with sufficient network effects (or equivalently, soteriological foundations that provide cooperation benefits independent of N) can cooperation be sustained at scale.

Evolutionary stability analysis

We analyze evolutionary dynamics using replicator equations. Let x be the fraction of cooperators. The fitness functions are:

$$f_C(x) = \frac{bx}{1} - c + \bar{M}_C \tag{8}$$

$$f_D(x) = \frac{bx}{1} \tag{9}$$

where we normalize to a two-player game for tractability.

The replicator dynamics are:

$$\dot{x} = x(1-x)[f_C(x) - f_D(x)] = x(1-x)[\bar{M}_C - c] \tag{10}$$

Equilibria:

- $x^* = 0$ (all defect): Stable if $\bar{M}_C < c$
- $x^* = 1$ (all cooperate): Stable if $\bar{M}_C > c$

Evolutionary stability:

A strategy is evolutionarily stable (ESS) if it resists invasion. Cooperation is an ESS when:

$$\bar{M}_C > c + \epsilon \tag{11}$$

for some invasion barrier $\epsilon > 0$. This requires intrinsic motivation to exceed cooperation costs with margin.

Implications: The replicator dynamics confirm the ABM results—cooperation requires sufficient intrinsic motivation to overcome the free-rider incentive. Without it, any cooperative population can be invaded by defectors. Soteriological foundations that provide stable $M_i > c$ make cooperation evolutionarily stable.

F.8. **Historical Data Analysis**

Coordination system longevity

We analyze historical coordination systems to calibrate model parameters and test theoretical predictions. The sample includes international institutions, empires, and governance systems with documented lifespans and corruption/collapse patterns.

System	Duration (yrs)	Scale	Collapse Mode
League of Nations	26	Global	Defection cascade
Bretton Woods	27	Global	Unilateral exit
Concert of Europe	99	Continental	Great power rivalry
Holy Roman Empire	844	Continental	Gradual fragmentation
Roman Empire (West)	503	Continental	Corruption + invasion
Byzantine Empire	1123	Regional	Military defeat
Hanseatic League	400	Regional	Competition

TABLE 9. Sample of historical coordination systems with documented collapse patterns.

Calibration results:

- Mean cycle duration: 45 years ($\sigma = 26$)
- Global-scale systems: median 27 years
- Regional/continental systems: median 400+ years
- Longest-enduring systems share soteriological foundations (religious legitimacy, civilizational identity)

Scale-longevity relationship

We test the theoretical prediction that coordination systems face increasing difficulty at larger scales. Using the historical sample:

Scale Category	Mean Duration	N
Local (city/region)	312 years	8
National	189 years	12
Continental	284 years	6
Global	32 years	4

TABLE 10. Coordination system longevity by scale (historical sample).

Finding: Global-scale coordination systems have dramatically shorter lifespans than regional systems. This aligns with theoretical predictions about institutional mechanism degradation at scale. The longer-lived continental systems (Holy Roman Empire, Byzantine Empire) maintained strong soteriological foundations through religious authority.

Corruption trajectory patterns

Historical cases exhibit consistent corruption trajectory patterns matching the theoretical model:

Phase 1: Establishment (0–20% of lifespan)

- High legitimacy and compliance
- Strong enforcement of norms
- Example: League of Nations 1920–1925

Phase 2: Stress testing (20–50% of lifespan)

- First significant defections
- Enforcement challenges emerge
- Example: League failure to respond to Japan in Manchuria (1931)

Phase 3: Erosion (50–80% of lifespan)

- Cascade of norm violations
- Enforcement becomes selective
- Example: League impotence during Italian invasion of Ethiopia (1935)

Phase 4: Collapse (80–100% of lifespan)

- Mass defection or dissolution
- Enforcement ceases to function
- Example: League irrelevance during WWII

Predictors of stability

Cox proportional hazards analysis of historical cases identifies predictors of system survival:

Key findings:

Factor	Hazard Ratio	Direction
Global scale	3.2	Increases risk
Soteriological foundation	0.4	Decreases risk
Voluntary membership	0.6	Decreases risk
Great power participation	1.8	Increases risk
Graduated sanctions	0.5	Decreases risk

TABLE 11. Cox hazard ratios for coordination system collapse (qualitative estimates based on historical patterns).

(1) **Scale is the strongest predictor**: Global-scale systems have 3.2× higher collapse risk than regional systems.

(2) **Soteriological foundations are protective**: Systems with transcendent legitimacy (religious, ideological) have 60% lower collapse risk.

(3) **Voluntary membership matters**: Coerced participation increases defection risk when enforcement weakens.

(4) **Great power participation is double-edged**: Provides resources but creates enforcement asymmetries.

Limitations

The historical analysis has important limitations:

- **Small sample**: Only ∼30 well-documented cases
- **Selection bias**: Failed systems may be underrepresented in historical record
- **Confounding**: Scale correlates with many other factors (technology, population, etc.)
- **Qualitative hazard ratios**: Formal survival analysis requires larger sample with consistent coding

Despite these limitations, the historical patterns consistently support the theoretical predictions: larger scale increases collapse risk, and soteriological foundations provide stability that institutional mechanisms alone cannot.

F.9. Summary of Computational Findings

The computational results support the theoretical analysis in the following ways:

(1) **Corruption inevitability confirmed**: The corruption dynamics ABM shows 100% convergence to full corruption under all tested parameter configurations. The result is robust to variation in initial integrity and detection probability, with effect sizes below 8%.

(2) **Critical mass thresholds validated**: The cooperation threshold model reproduces the theoretical $\theta_{\text{crit}} = 0.40$ and demonstrates the predicted bifurcation behavior. Initial conditions determine whether the system converges to cooperation or defection equilibrium.

(3) **Timeline predictions with quantified uncertainty**: Monte Carlo simulations provide probability distributions over extinction timelines. The calibrated baseline predicts median extinction at 432 years with 90% credible interval [58, 817] years. Scenario analysis shows sensitivity to initial p_{AI} and growth rate assumptions.

(4) **Sensitivity analysis identifies load-bearing parameters**: The growth rate of AI adoption probability has highest leverage on outcomes. Cycle duration and initial p_{AI} also significantly affect timelines.

(5) **Ostrom's principles bounded by scale**: Polycentric governance works within Dunbar-scale groups ($\sim$150) but degrades sharply beyond. At $N = 300$, corruption rises from 24% to 63%. The mechanisms that make Ostrom's approach effective (peer monitoring, social pressure, collective choice) cannot operate at global scale.

(6) **Soteriological foundations structurally required**: Institutional motivation sources fail beyond $\sim$1.5$\times$ Dunbar scale through explicit mechanisms (monitoring costs, reputation degradation, social pressure diffusion, free-rider detection). Soteriological foundations maintain 100% cooperation at all tested scales. This aligns with anthropological evidence that all large-scale civilizations develop soteriological systems.

(7) **Critical soteriological threshold identified**: At $N = 1000$, system stability requires $\geq$50% agents with soteriological foundations. This identifies the critical mass needed for VCS-style value transformation.

(8) **Game-theoretic equilibria derived**: N-player public goods analysis shows cooperation converges to zero as $N \to \infty$ without network effects or intrinsic motivation. Replicator dynamics confirm cooperation is evolutionarily stable only when $M_i > c$.

(9) **Historical patterns support theory**: Analysis of historical coordination systems shows global-scale systems have 3.2$\times$ higher collapse risk than regional systems, and soteriological foundations reduce collapse risk by 60%. Corruption trajectory phases match theoretical predictions.

Key uncertainties that remain:

- **AI alignment probability**: The model assumes AI-controlled TCS always leads to extinction. Partial alignment success would reduce this probability (though even 95% alignment yields >60% extinction over 10,000 years).
- **Forecast reliability**: AI capability forecasts have historically been unreliable; calibration may shift substantially with new evidence.
- **Independence assumption**: Cycles are modeled as independent; correlated shocks or learning effects could change dynamics.
- **Soteriological operationalization**: The model treats soteriological motivation as binary and exogenous. In reality, value transformation is a gradual process with complex determinants.
- **Scale decay parameters**: The exact form of institutional degradation (10% reputation loss per hop, etc.) is calibrated to produce Dunbar-scale transitions but requires empirical validation.

- **Historical sample size**: Formal survival analysis requires larger sample with consistent variable coding; current hazard ratios are qualitative estimates.

References

1. M. Abbasi, P. Váz, J. Silva, and P. Martins, *Comprehensive evaluation of deepfake detection models: Accuracy, generalization, and resilience to adversarial attacks*, Applied Sciences **15** (2025), no. 3, 1225.
2. V. Bhandarkawthekar, T. M. Navamani, R. Sharma, and K. Shyamala, *Design and development of an efficient rlnet prediction model for deepfake video detection*, Frontiers in Big Data **8** (2025), 1569147.
3. N. Chandra, R. Murtfeldt, L. Qiu, A. Karmakar, H. Lee, E. Tanumihardja, K. Farhat, B. Caffee, S. Paik, C. Lee, J. Choi, A. Kim, and O. Etzioni, *Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024*, 2025.
4. Columbia Journalism Review, *What journalists should know about deepfake detection in 2025*, https://www.cjr.org/tow_center/what-journalists-should-know-about-deepfake-detection-technology-in-2025-a-nc php, 2025.
5. A. Diel, T. Lalgi, I. C. Schröter, M. Groh, E. Specker, and H. Leder, *Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers*, Computers in Human Behavior: Artificial Humans **2** (2024), no. 2, 100085.
6. M. Groh, Z. Epstein, C. Firestone, and R. Picard, *Deepfake detection by human crowds, machines, and machine-informed crowds*, Proceedings of the National Academy of Sciences **119** (2022), no. 1, e2110013119.
7. OpenAI, *Sora 2 is here*, https://openai.com/index/sora-2/, September 2025, OpenAI Blog.
8. ———, *Sora 2 system card*, https://openai.com/index/sora-2-system-card/, September 2025, OpenAI Safety.
9. K. Somoray, J. Zhao, W. Zheng, J. Phua, and S. K. Sia, *Human performance in deepfake detection: A systematic review*, Human Behavior and Emerging Technologies **2025** (2025), 1833228.