Dissensus AI Working Paper Series

DAI-2601

The Axiom of Consent:

Friction Dynamics in Multi-Agent Coordination

Murad Farzulla[1,2,*]

[1]Dissensus AI, London, UK [2]King's College London, London, UK

[*]Correspondence: murad@dissensus.ai ORCID: 0009-0002-7164-8704

February 2026

Abstract

Multi-agent systems face a fundamental coordination problem: agents must coordinate despite heterogeneous preferences, asymmetric stakes, and imperfect information. When coordination fails, friction emerges—measurable resistance manifesting as deadlock, thrashing, communication overhead, or outright conflict. This paper derives a formal framework for analyzing coordination friction from a single axiom: actions affecting agents require authorization from those agents in proportion to stakes.

From this axiom of consent, we establish the *kernel triple* $(\alpha, \sigma, \varepsilon)$—alignment, stake, and entropy—characterizing any resource allocation configuration. The friction equation $F = \sigma \cdot (1+\varepsilon)/(1+\alpha)$ predicts coordination difficulty as a function of preference alignment α, stake magnitude σ, and communication entropy ε. The Replicator-Optimization Mechanism (ROM) governs evolutionary selection over coordination strategies: configurations generating less friction persist longer, establishing consent-respecting arrangements as dynamical attractors rather than normative ideals.

We develop formal definitions for resource consent, coordination legitimacy, and friction-aware allocation in multi-agent systems. The framework yields testable predictions: MARL systems with higher reward alignment exhibit faster convergence; distributed allocations accounting for stake asymmetry generate lower coordination failure; AI systems with interpretability deficits produce friction proportional to the human-AI alignment gap. Applications to cryptocurrency governance and political systems demonstrate that the same equations govern friction dynamics across domains—a complexity science perspective on coordination under preference heterogeneity.

Keywords: multi-agent systems, coordination, friction, evolutionary dynamics, AI alignment, complexity science

arXiv Categories: cs.MA (primary), cs.GT, cs.SI

DOI: 10.48550/arXiv.2601.06692 | **Version:** v2.0.0 | **License:** CC-BY-4.0

Contents

1 Introduction

1.1 The Coordination Problem

Multi-agent systems face a fundamental tension: agents must coordinate despite heterogeneous preferences, asymmetric stakes, and imperfect information (Olson, 1965). When coordination succeeds, resources flow efficiently and system-level objectives emerge from local interactions. When coordination fails, the result is friction—measurable resistance manifesting as wasted computation, deadlock, thrashing, or outright conflict.

A striking empirical puzzle motivates this paper: why do structurally similar interventions produce radically different resistance across systems? Consider two classes of coordination events. Protocol upgrades proposed through established governance channels—Bitcoin Improvement Proposals, Ethereum's EIP process, community-ratified hard forks—generate remarkably low friction despite their technical complexity. The same systems respond to externally imposed changes with volatility amplification factors exceeding 5.7× baseline levels (Farzulla, 2025e). Both intervention types alter system conditions; both require participant adaptation; both carry economic stakes. Yet one class integrates smoothly while the other generates persistent turbulence.

This asymmetry is not unique to cryptocurrency. Distributed systems with participatory resource allocation exhibit lower coordination failure than systems with centralized allocation imposed without agent consent. Multi-agent reinforcement learning with aligned reward functions converges more reliably than systems with misaligned incentives. The pattern repeats: interventions aligned with stakeholder preferences encounter less friction than interventions misaligned with those preferences, even when the two produce identical immediate outcomes.

This paper argues that these patterns reflect a single underlying phenomenon: *friction*—the measurable resistance generated when decision authority diverges from consequence-bearing. Sornette et al. (2026) independently derive a structural friction framework from statistical physics, identifying learned human interaction structures as sources of alignment failure whose disruption by AGI constitutes an endogenous evolutionary shock. Their formalization converges with the axiomatization presented here from a completely different disciplinary foundation, providing strong convergent evidence for the friction-based approach. Further convergent evidence comes from opinion dynamics: Stokes et al. (2024) develop an agent-based model with pairwise affinity, memory capacity, and heterogeneous interaction thresholds, proving that consensus obtains under full interaction but that *limited interaction* drives collective extremisation and oscillatory dynamics. Their "limited interaction" is formally analogous to friction in our framework—structural constraints on who interacts with whom—and their finding that restricting interaction produces extremisation rather than moderation provides independent mathematical confirmation from social psychology of our prediction that friction does not merely slow coordination but qualitatively reshapes outcomes. We derive a complete formal framework from a single axiom: actions affecting agents require authorization from those agents in proportion to stakes. From this "axiom of consent," we establish the *kernel triple* $(\alpha, \sigma, \varepsilon)$—alignment, stake, and entropy—characterizing any consent-holding configuration.

The canonical friction equation predicts system behavior:

$$F = \sigma \cdot \frac{1+\varepsilon}{1+\alpha} \tag{1}$$

Friction F increases with stake magnitude σ, increases with information loss ε, and decreases with alignment α. When consent-holders are perfectly aligned with stake-holders ($\alpha = 1$) and information transmission is perfect ($\varepsilon = 0$), friction reduces to an irreducible baseline $\sigma/2$—the minimal coordination cost of delegation itself. When consent-holders are perfectly misaligned ($\alpha \to -1$), friction diverges.

1.2 Connection to Multi-Agent Systems

The framework provides natural primitives for multi-agent coordination:

- **Consent-holding** maps to resource allocation authority. An agent "holds consent" over a resource if it controls allocation decisions for that resource.
- **Stakes** map to consequence exposure. An agent's stake in a decision is the magnitude of utility change that decision can induce.
- **Alignment** maps to preference correlation. Two agents are aligned if their utility functions correlate positively over the relevant decision space.
- **Entropy** maps to communication overhead. Information loss between decision-makers and affected agents creates coordination uncertainty.
- **Friction** maps to coordination failure. Wasted computation, deadlock, thrashing, and explicit conflict are manifestations of friction.
- **Legitimacy** maps to sustainable coordination. Arrangements where voice tracks stakes persist; arrangements where voice and stakes diverge generate accumulating friction until reconfiguration.

The Replicator-Optimization Mechanism (ROM) (Farzulla, 2025g) governs evolutionary selection on coordination configurations:

$$\frac{dp_t(\tau)}{dt} = \sum_{\tau'} p_t(\tau') \cdot \sigma(\tau') \cdot \frac{L(\tau')}{1 + F(\tau')} \cdot M_S(\tau' \to \tau) - p_t(\tau) \cdot \bar{\phi}_t \tag{2}$$

Configurations with high legitimacy L and low friction F persist; those generating unsustainable friction are selected against. This provides a formal basis for understanding why certain coordination patterns emerge and persist while others fail.

1.3 Existing Approaches and Their Limitations

Three theoretical traditions have addressed multi-agent coordination. Each captures important features; none provides the unified apparatus we seek.

Mechanism Design. The Hurwicz-Myerson-Maskin tradition (Hurwicz, 1960; Myerson, 1981; Maskin, 1999) provides powerful tools for incentive-compatible allocation. Recent work on commitment-enhanced communication (Avoyan and Ramos, 2023) demonstrates that credible coordination mechanisms significantly improve efficiency over cheap talk. Yet mechanism design assumes preference revelation: agents report preferences, and the mechanism aggregates them. This assumption breaks down when agents *cannot* reveal preferences (bandwidth constraints), *will not* reveal preferences (strategic concealment), or when revealed preferences are systematically distorted by power asymmetries. Mechanism design also lacks a natural treatment of stakes asymmetry: the agent most affected by a decision may have least voice in its governance.

Evolutionary Game Theory. Replicator dynamics (Taylor and Jonker, 1978), the Price equation (Price, 1970), and evolutionary stability concepts (Maynard Smith and Price, 1973) provide powerful tools for analyzing strategic interaction under selection. Vanderschraaf's work on inductive deliberation (Vanderschraaf, 2001) shows how correlated equilibria emerge from rational learning, while Golman and

Page (Golman and Page, 2009) demonstrate inherent speed-accuracy tradeoffs in decentralized decision-making. Yet evolutionary game theory as standardly formulated ignores normative structure: replicator dynamics describe what *persists*, not what coordination patterns *should* emerge. The gap between fitness and legitimacy—between what survives and what generates sustainable coordination—remains unbridged.

Multi-Agent Reinforcement Learning. MARL provides computational tools for learning coordination (Buşoniu et al., 2008; Zhang et al., 2021). Organizational research reveals the dual challenge of search and coordination: hierarchical influence structures outperform flat teams in environments requiring rapid convergence (Koçak and Puranam, 2023), while network centralization improves collective adaptability when two-way information flow is preserved (Bernstein et al., 2002). Yet MARL typically assumes reward functions are given, not derived. The framework asks how agents learn to coordinate given objectives; it does not ask how objectives themselves should be structured to minimize coordination failure. The consent-friction framework addresses this gap: alignment α measures reward function correlation, and the friction equation predicts coordination difficulty given alignment structure.

1.4 Our Contribution

This paper provides a unified formal framework for analyzing friction dynamics in multi-agent systems. Our contributions are threefold.

Contribution 1: Single Axiom to Complete Apparatus. We derive a complete formal framework from a single axiom: actions affecting agents require authorization from those agents in proportion to stakes. From this axiom, we derive the kernel triple $(\alpha, \sigma, \varepsilon)$. These three quantities characterize any consent-holding configuration and determine its friction profile. The kernel triple is not an analogy across domains but the *same mathematical structure* applied at different scales.

Contribution 2: Scale-Relative Framework via Coarse-Graining. The kernel triple operates at multiple scales simultaneously. An individual agent holds consent over local decisions. A coalition holds consent over collective resources. A system holds consent over global parameters. We develop a coarse-graining apparatus connecting micro-level dynamics to macro-level observables. Under lumpability conditions we specify, friction dynamics at scale S approximately preserve structure when projected to scale S'.

Contribution 3: Multi-Agent Coordination as Primary Application. We demonstrate the framework's power through detailed application to multi-agent systems:

(i) Resource allocation as consent-holding, with friction as coordination failure
(ii) Alignment verification via kernel triple measurement
(iii) ROM dynamics governing agent population evolution
(iv) Predictions for coordination stability and phase transitions

We also show how the framework applies to other domains—cryptocurrency governance, political legitimacy—as instances of the same underlying structure, demonstrating the generality of consent-friction dynamics.

1.5 Methodological Notes

Several methodological points require clarification.

Descriptive, Not Prescriptive. We do not claim that consent *ought* to be respected; we claim that configurations where consent is respected *exhibit* lower friction. This is an empirical claim, testable through the operationalizations we provide. The framework is descriptive: it predicts friction levels given consent configurations. Whether lower friction is desirable is a separate question.

However, we offer a bridge principle. *If* agents prefer lower coordination failure (lower friction), *then* configurations with higher consent alignment are instrumentally preferred. This conditional structure avoids the is-ought fallacy while grounding normative discourse in empirically tractable dynamics.

Falsifiability. The framework is falsifiable at the level of measurement apparatus. If the operationalizations we provide (Section 6) fail to predict friction across domains, the framework fails. We provide specific empirical predictions: multi-agent systems with higher reward alignment should exhibit lower coordination failure; systems with higher legitimacy should exhibit greater stability. Failure of these predictions would falsify the framework.

Scope Conditions. The framework applies where Lewontin's minimal conditions for selection hold (Lewontin, 1970): variation among configurations, differential persistence, and heritable transmission. Multi-agent systems with learning and adaptation satisfy these conditions naturally.

1.6 Roadmap

The paper proceeds as follows.

Section 2 develops the Axiom of Consent from first principles, deriving the kernel triple and establishing primitive definitions.

Section 3 presents the kernel triple formalism, connecting the axiom to the ROM evolutionary dynamics.

Section 4 develops the dynamical treatment: friction evolution, legitimacy dynamics, and key theorems including Lyapunov stability conditions.

Section 5 applies the framework to multi-agent coordination as the primary domain, with supplementary applications to cryptocurrency governance and political systems demonstrating cross-domain generality.

Section 6 develops the measurement apparatus, specifying operationalizations for alignment, stake, and entropy with focus on multi-agent system applicability.

Section 7 addresses objections, limitations, and pathological cases.

Section 8 concludes with implications for AI alignment, distributed systems design, and directions for future work.

2 The Axiom of Consent

2.1 Informal Statement

We begin with a claim that appears normative but is, upon examination, purely structural:

> **The Axiom of Consent (Informal)**
>
> *No entity may be bound by commitments it did not consent to, weighted by its stake in the outcome.*

The standard reading treats this as a moral principle—an *ought* claim about political legitimacy or individual rights. We propose a radically different interpretation. The axiom describes an unavoidable *structural* feature of any system where multiple agents interact in shared decision domains. It is not a prescription for how governance *should* operate but a description of a constraint that *all* governance arrangements face.

The insight is this: wherever decisions affecting multiple parties occur, some locus of control determines outcomes. This locus may be concentrated (a sovereign, an algorithm, a parent) or distributed (a vote, a market, a consensus mechanism). It may be explicit (constitutional authority) or implicit (first-mover advantage, social convention). But it cannot be absent. Even "leaving things to chance" discloses a prior decision to permit randomness; even "letting the market decide" reveals the meta-choice to instantiate market mechanisms.

We call this locus **consent-holding**—the custody of decision authority in a shared domain. The axiom's force lies not in normative assertion but in structural necessity: consent-holding is *unavoidable* wherever outcomes occur.

2.1.1 Why This Is Not Normative Ethics

Traditional normative ethics asks: *Who should hold consent?* The utilitarian answers: whoever maximizes aggregate welfare. The deontologist answers: whoever respects categorical duties. The contractualist answers: whoever would be chosen under idealized conditions. Each tradition offers criteria for adjudicating disputes about legitimate authority.

We ask a different question entirely: *What happens when consent-holding configurations misalign with stakes?* This is a descriptive question with empirical answers. When those with high stakes have low voice, friction accumulates. When those with low stakes have high voice, different frictions emerge. The relationship between consent-holding and consequence-bearing determines system dynamics regardless of normative evaluations.

This reframing transforms the axiom from a contested moral principle into a falsifiable structural hypothesis: configurations where consent-holding diverges systematically from stake-bearing exhibit predictable friction patterns. The axiom does not tell us who *should* hold consent; it predicts what *will* happen under various consent-holding configurations.

2.1.2 Consent-Holding as Structural Fact

Consider the minimal assumptions required for the axiom to apply:

1. Agents act in shared domains
2. Actions produce outcomes
3. Preferences and stakes differ across agents
4. Attention and capacity are finite

From these four premises—none of them normative—a structural conclusion follows: in any domain where a non-null outcome occurs, some procedure selected an action, and therefore some locus of control held the right to decide.

This locus may be:

- **Concentrated**: a monarch, CEO, parent, algorithm owner
- **Distributed**: a vote, board, consensus mechanism
- **Delegated**: a randomization rule, market mechanism, coin flip
- **Encoded**: an algorithm, smart contract, institutional procedure

But it cannot be absent. This is the axiom's structural content: *consent-holding is unavoidable in multi-agent coordination*. Even radical disagreement about who *should* hold consent presupposes that *someone* does.

2.2 Formal Axiom

We now state the axiom formally, integrating notation from the unified framework.

Axiom 2.1 (The Consent Principle). For any decision $d \in \mathscr{D}$ affecting agent $i \in \mathscr{A}$ with stake $s_i(d) > 0$, the **legitimacy** of the decision is determined by:

$$\text{Legitimate}(d) \iff \sum_{i \in S_d} s_i(d) \cdot \nVdash[\text{Consent}_i(d)] \geq \theta \cdot \sum_{i \in S_d} s_i(d) \tag{3}$$

where $S_d = \{i \in \mathscr{A} : s_i(d) > 0\}$ is the affected set, $\nVdash[\text{Consent}_i(d)]$ is the indicator of agent i's consent to decision d, and $\theta \in (0,1]$ is a threshold parameter.

This formal statement encodes several substantive commitments:

Stakes-weighting. Consent is weighted by stakes. An agent with stake $s_i(d) = 10$ who consents contributes more to legitimacy than an agent with stake $s_i(d) = 1$ who consents. This reflects the proportionality principle: those who bear greater consequences should have proportionally greater voice (Brighouse and Fleurbaey, 2010; Mackay, 2020; Afsahi et al., 2021).

Threshold structure. The threshold θ parameterizes how much stake-weighted consent is required for legitimacy. At $\theta = 1$, only unanimity among affected parties suffices. At lower values, supermajority or simple majority thresholds apply. This connects to the classical analysis of optimal majority rules by Buchanan and Tullock (1962), who showed that different decision types warrant different threshold structures. The framework does not specify θ; different domains may require different thresholds.

Affected set restriction. Only those with positive stakes are included in the legitimacy calculation. This addresses the boundary problem in democratic theory (Arrhenius, 2005; Goodin, 2007): who counts in aggregation? Our answer: those with nonzero stakes in the domain.

The binary consent indicator $\nVdash[\text{Consent}_i(d)]$ is a simplification. In practice, consent admits degrees—enthusiastic endorsement, reluctant acquiescence, resigned acceptance, passive non-resistance. We address this complexity through the alignment function below.

2.3 Derived Concepts

From the axiom, we derive three central concepts: alignment, aggregate alignment, and friction. These definitions operationalize the intuitive notions that ground the framework.

Definition 2.2 (Alignment)**.** For agent i in domain d at time t, the **alignment** $\alpha_i(d,t) \in [-1,1]$ measures the correlation between agent i's target function and the consent-holder's target function:

$$\alpha_i(d,t) = \text{corr}(T_i(S), T_{H(d,t)}(S)) = \frac{\text{Cov}(T_i, T_{H(d,t)})}{\sqrt{\text{Var}(T_i) \cdot \text{Var}(T_{H(d,t)})}} \tag{4}$$

where $T_i : \mathscr{O} \to \mathbb{R}$ is agent i's target function, $H(d,t)$ denotes the consent-holder in domain d at time t, and the correlation is computed under a probability measure over the state space S.

Alignment captures whether the consent-holder optimizes for outcomes that agent i also values. When $\alpha_i = 1$, the consent-holder's target function perfectly correlates with i's—they want the same things. When $\alpha_i = -1$, perfect misalignment: what the consent-holder pursues is precisely what i seeks to avoid. When $\alpha_i = 0$, the targets are orthogonal—the consent-holder's optimization is irrelevant to i's interests.

Definition 2.3 (Aggregate Alignment)**.** The **aggregate alignment** in domain d at time t is the stakes-weighted average of individual alignments:

$$\alpha(d,t) = \frac{\sum_{i \in S_d} s_i(d) \cdot \alpha_i(d,t)}{\sum_{i \in S_d} s_i(d)} \tag{5}$$

Aggregate alignment measures how well the consent-holder's optimization serves the affected population as a whole, weighted by stakes. A consent-holder who perfectly aligns with high-stakes parties but misaligns with low-stakes parties achieves higher aggregate alignment than one who does the reverse.

Definition 2.4 (Friction)**.** The **friction** in domain d at time t is:

$$F(d,t) = \sigma(d) \cdot \frac{1+\varepsilon(d,t)}{1+\alpha(d,t)} \tag{6}$$

where:

- $\sigma(d) = \sum_{i \in S_d} s_i(d)$ is the total stake magnitude
- $\varepsilon(d,t) \in [0,1]$ is the information entropy
- $\alpha(d,t) \in [-1,1]$ is the aggregate alignment

Remark 2.5 (Empirically-Refined Friction Form)**.** The MARL factorial experiment (Appendix C) reveals that the alignment–friction relationship is U-shaped rather than monotonic: neutral alignment ($\alpha = 0$) produces the worst coordination outcomes, while both cooperative ($\alpha > 0$) and adversarial ($\alpha < 0$) alignment reduce friction. This motivates an empirically-refined quadratic form:

$$F^{(2)}(d,t) = \sigma(d) \cdot \frac{1+\varepsilon(d,t)}{1+\alpha(d,t)^2} \tag{7}$$

The quadratic form agrees with the canonical form at $\alpha = 0$ (both yield $\sigma(1+\varepsilon)$) and at $\alpha = 1$ (both yield $\sigma(1+\varepsilon)/2$), but diverges in the adversarial regime ($\alpha < 0$): where the canonical form diverges as $\alpha \to -1$, the quadratic form remains bounded, with maximum friction at $\alpha = 0$ and symmetric attenuation toward both alignment extremes. The quadratic specification achieves $R^2 = 0.34$–0.43 in MARL validation compared to $R^2 = 0.05$–0.13 for the canonical form (Section 6). The canonical form is retained as the theoretical baseline satisfying the uniqueness conditions of Appendix B; the quadratic form is presented as an empirical refinement that relaxes the divergence desideratum D6 to a bounded non-monotonicity condition D6′ (see Appendix B for the formal derivation under D6′).

The friction function captures the *structural tension* in a consent-holding configuration. Three components interact:

Stake magnitude (σ) amplifies friction proportionally. High-stakes domains generate more friction than low-stakes domains, all else equal. A consent-holder making decisions about life and death faces more friction than one choosing office supplies.

Entropy (ε) captures information loss in the consent-holding relationship—the proportion of affected parties' preferences that the consent-holder does not know or cannot implement. Even perfectly aligned agents generate friction when entropy is high, because the consent-holder optimizes for a *misspecified* version of what affected parties actually want.

Alignment (α) appears in the denominator: higher alignment reduces friction, lower alignment amplifies it. As $\alpha \to -1$ (perfect misalignment), friction approaches infinity—the consent-holder actively pursues outcomes that affected parties seek to avoid.

This functional form is not arbitrary. It satisfies desirable properties:

- $F \geq 0$ for all valid inputs (friction is non-negative)
- As $\alpha \to 1$: $F \to \sigma(1+\varepsilon)/2$ (minimal friction proportional to stakes and entropy)
- As $\alpha \to -1$: $F \to \infty$ (unbounded friction under misalignment)
- With $\varepsilon = 0$ and $\alpha = 1$: $F = \sigma/2$ (irreducible baseline)

The irreducible baseline $\sigma/2$ reflects a deep insight: *delegation has friction even in paradise*. Even ideally aligned agents with perfect information incur transaction costs—the cognitive and coordinative overhead of consent-holding itself.

The quadratic form $F^{(2)} = \sigma(1+\varepsilon)/(1+\alpha^2)$ from Remark 2.5 satisfies modified properties:

- $F^{(2)} \geq 0$ for all valid inputs (friction is non-negative)
- Maximum at $\alpha = 0$: $F^{(2)} = \sigma(1+\varepsilon)$ (neutral alignment is worst)
- Minimum at $\alpha = \pm 1$: $F^{(2)} = \sigma(1+\varepsilon)/2$ (symmetric attenuation)
- No singularity: $F^{(2)}$ is bounded for all $\alpha \in [-1,1]$
- With $\varepsilon = 0$ and $\alpha = 1$: $F^{(2)} = \sigma/2$ (same irreducible baseline)

2.4 Properties

We establish basic properties of the friction function through formal propositions.

Proposition 2.1 (Zero Friction Condition)**.** *$F(d,t) = 0$ if and only if $\sigma(d) = 0$.*

Proof. If $\sigma(d) = 0$, then by Definition 2.4, $F(d,t) = 0 \cdot \frac{1+\varepsilon}{1+\alpha} = 0$.

Conversely, suppose $F(d,t) = 0$ with $\sigma(d) > 0$. Then:

$$0 = \sigma(d) \cdot \frac{1+\varepsilon(d,t)}{1+\alpha(d,t)}$$

Since $\sigma(d) > 0$, we require $(1+\varepsilon)/(1+\alpha) = 0$. But $\varepsilon \geq 0$ implies $1+\varepsilon \geq 1 > 0$, and $\alpha \leq 1$ implies $1+\alpha \leq 2 < \infty$. Thus the fraction is positive, contradiction. Therefore $\sigma(d) = 0$. ■ ■

This proposition captures an important insight: friction cannot be eliminated in any domain with positive stakes. *Zero friction requires zero stakes*—only domains where no one is affected can be friction-free.

Proposition 2.2 (Alignment Effect)**.** *For fixed $\sigma > 0$ and $\varepsilon \geq 0$:*

$$\frac{\partial F}{\partial \alpha} < 0$$

Friction decreases as alignment increases.

Proof.

$$\frac{\partial F}{\partial \alpha} = \sigma \cdot (1+\varepsilon) \cdot \frac{\partial}{\partial \alpha}\left(\frac{1}{1+\alpha}\right) = -\frac{\sigma(1+\varepsilon)}{(1+\alpha)^2} < 0$$

since $\sigma > 0$, $1+\varepsilon > 0$, and $(1+\alpha)^2 > 0$ for $\alpha > -1$. ■ ■

Proposition 2.3 (Stake Effect). *For fixed $\alpha < 1$ and $\varepsilon \geq 0$:*

$$\frac{\partial F}{\partial \sigma} > 0$$

Friction increases with stake magnitude when alignment is imperfect.

Proof.

$$\frac{\partial F}{\partial \sigma} = \frac{1+\varepsilon}{1+\alpha} > 0$$

since $\varepsilon \geq 0$ implies $1+\varepsilon \geq 1$, and $\alpha \leq 1$ implies $1+\alpha \leq 2$, hence the ratio is positive. ■ ■

Proposition 2.4 (Entropy Effect). *For fixed $\sigma > 0$ and $\alpha \in (-1, 1]$:*

$$\frac{\partial F}{\partial \varepsilon} > 0$$

Friction increases with information entropy.

Proof.

$$\frac{\partial F}{\partial \varepsilon} = \frac{\sigma}{1+\alpha} > 0$$

since $\sigma > 0$ and $1+\alpha > 0$ for $\alpha > -1$. ■ ■

These four propositions establish the basic comparative statics of the friction function. They formalize intuitions that will prove central to empirical applications: higher stakes amplify friction; better alignment reduces it; information loss increases it; and zero friction requires zero stakes.

2.5 Information-Theoretic Foundations

The friction function $F = \sigma(1+\varepsilon)/(1+\alpha)$ is not merely an intuitive construction but admits rigorous derivation from information-theoretic principles. We sketch three independent derivations that converge on this functional form.

2.5.1 Derivation from Constrained Optimization

Consider the optimization problem facing a consent-holder who must balance stakeholder preferences under information constraints. The consent-holder aims to maximize stakeholder welfare but faces limited channel capacity for learning preferences:

$$\max_{a \in \mathscr{A}} \sum_{i \in S_d} s_i(d) \cdot U_i(a) \quad \text{subject to} \quad I(a; P_i) \leq C_i \; \forall i \tag{8}$$

where $I(a; P_i)$ is mutual information between action a and stakeholder i's preferences P_i, and C_i is channel capacity. The Lagrangian dual yields friction as the *shadow price* of information constraints (Cover and Thomas, 2006):

$$F = \sum_{i \in S_d} s_i(d) \cdot \lambda_i \cdot D_{KL}(P_i \| \hat{P}_i) \tag{9}$$

where λ_i are dual variables and D_{KL} is Kullback-Leibler divergence between true and estimated preferences. Under appropriate regularity conditions, this reduces to the friction equation with $\varepsilon_i = D_{KL}(P_i \| \hat{P}_i)$ and α_i entering through the covariance structure of utilities.

2.5.2 Derivation from Information Decomposition

This derivation connects to Partial Information Decomposition (PID) (Mediano et al., 2022). Consider a consent-holder H receiving preference signals from stakeholders. The total information available for decision-making decomposes into:

- **Redundant information**: Shared structure across stakeholder utilities that the consent-holder can exploit (related to alignment α)
- **Unique information**: Stakeholder-specific preferences requiring individual attention
- **Synergistic information**: Information requiring the *full ensemble* of stakeholder signals to decode (captured by entropy ε)

Friction arises from the mismatch between information *required* for optimal decision-making and information *available* to the consent-holder. The multiplicative form $(1+\varepsilon)/(1+\alpha)$ reflects that alignment and entropy operate on different information dimensions: alignment captures redundant structure that simplifies coordination, while entropy captures synergistic complexity that resists compression.

2.5.3 Derivation from Diversity Decomposition

Recent work on diversity and institutional resilience (Beinhocker and Bednar, 2025) suggests a complementary derivation. Bednar's diversity decomposition identifies three dimensions of systemic variety:

- **Variety**: Magnitude of heterogeneity in stakeholder preferences $\to \sigma$
- **Modularity**: Degree of clustering in preference space $\to 1/(1+\alpha)$
- **Redundancy**: Overlap in stakeholder objectives $\to 1/(1+\varepsilon)$

A multiplicative combination of these dimensions—appropriate when they interact rather than aggregate additively—yields the friction form. High variety (stakes) amplified by low modularity (misalignment) and low redundancy (high entropy) produces maximal friction.

2.5.4 Convergent Derivations

The convergence of three independent derivations—information-theoretic optimization, PID decomposition, and diversity analysis—provides theoretical support for the friction function's form. This is not a proof of *uniqueness*; alternative functional forms satisfying the same constraints may exist. However, the convergence suggests that $F = \sigma(1+\varepsilon)/(1+\alpha)$ captures genuine structural features of consent-holding configurations rather than arbitrary modeling choices.

For formal uniqueness results under specific axiomatic constraints, see Appendix B.

2.5.5 Functional Form Status: Phenomenological Ansatz

We state explicitly what this derivation *does not* establish. The friction function $F = \sigma(1+\varepsilon)/(1+\alpha)$ is a **phenomenological ansatz**—a functional form chosen for its desirable mathematical properties and convergent motivations, not uniquely derived from first principles.

What we have shown: Three independent theoretical perspectives—constrained optimization, information decomposition, and diversity analysis—suggest this general form. The function satisfies intuitive boundary conditions (non-negativity, monotonicity in alignment and entropy, stake-proportionality). Multiple derivation routes converging on similar structures is *suggestive* but not *conclusive*.

What we have not shown: That this is the *unique* function satisfying these constraints. Alternative forms—e.g., $F = \sigma \cdot e^{-\alpha}(1+\varepsilon)$, or weighted geometric means, or entropy-based divergence measures—might satisfy similar boundary conditions while yielding different quantitative predictions.

The honest claim: We propose F as a *working hypothesis*—a tractable functional form with plausible theoretical motivation that enables empirical testing. The framework's value lies not in the specific functional form but in the *conceptual architecture*: friction as primitive, consent as derived, the kernel triple $(\alpha, \sigma, \varepsilon)$ as sufficient statistics for governance dynamics.

If empirical observation suggests alternative functional forms fit data better, the framework can accommodate such revisions without abandoning its core conceptual structure. The phenomenological status of F is a feature, not a bug: it signals that the framework is falsifiable and revisable in light of evidence (see Section 6 for sensitivity analysis).

2.6 Philosophical Foundations

The axiom of consent engages with several traditions in moral and political philosophy. We position the framework relative to these traditions, identifying points of connection and departure.

2.6.1 Relationship to Scanlon's Contractualism

T.M. Scanlon's contractualism (Scanlon, 1998) represents the most sophisticated contemporary attempt to ground morality in consent-like reasoning. For Scanlon, an act is wrong if it would be disallowed by any principle that no one could reasonably reject.

The axiom of consent shares Scanlon's emphasis on what affected parties would accept, but diverges in three critical respects:

First, Scanlon's framework operates in *hypothetical* mode—it asks what principles *would* be accepted under idealized conditions of full information and mutual recognition. Our framework is *empirical*: we observe what configurations *actually* produce friction and derive conclusions from structural analysis.

Second, Scanlon's "reasonable rejection" criterion requires specifying what counts as reasonable—a move that reintroduces normative commitments at the foundation. Our stakes-weighting criterion is purely structural: we measure actual stakes and actual voice, not idealized reasonableness.

Third, Scanlon seeks principles that *no one* could reasonably reject—a unanimity condition. Our framework acknowledges that unanimity is typically unattainable and parameterizes the threshold θ to accommodate majority, supermajority, or other decision rules.

Despite these differences, the frameworks are complementary. Scanlonian contractualism provides normative guidance about what *should* count as reasonable; our framework provides empirical predictions about what configurations *will* generate friction regardless of normative evaluation.

2.6.2 The Proportionality Principle

Why weight consent by stakes? The answer lies in a principle with deep roots in political philosophy: *those who bear consequences should have proportional voice in decisions that produce them.*

This proportionality principle appears across traditions:

- In democratic theory, it underlies the "all-affected interests" principle (Goodin, 2007): all those affected by a decision should have a say in making it.
- In corporate governance, it motivates stakeholder theory: those affected by corporate decisions—employees, communities, suppliers—should have voice proportional to their stakes.

- In international relations, it grounds debates about who should participate in climate negotiations, trade agreements, and global governance.

Our stakes-weighting operationalizes this principle formally. An agent with stake $s_i = 100$ who is excluded from decision-making contributes ten times as much to legitimacy deficits as an agent with stake $s_i = 10$ who is excluded. This is not an arbitrary weighting but a reflection of the proportionality principle's substantive content.

2.6.3 The Affected Interests Principle

Robert Goodin's "all-affected interests" principle (Goodin, 2007) holds that all those affected by a decision should have standing to participate in making it. Our framework extends this principle in two directions:

First, we distinguish *standing* from *voice*. Having standing to participate does not guarantee effective voice. The affected set S_d includes all agents with positive stakes; the legitimacy function $L(d,t)$ measures how much effective voice they actually possess.

Second, we operationalize "affected" through the stakes function $s_i(d)$. Rather than treating affectedness as binary (affected or not), we measure its magnitude. An agent whose entire livelihood depends on a decision is more affected than one for whom the decision is peripheral.

This operationalization addresses a persistent challenge in democratic theory: how to identify the relevant constituency for a given decision. Our answer is empirical: measure stakes, include those with positive stakes, weight by magnitude.

2.6.4 Distinction from Unanimity Requirements

The axiom of consent might appear to require unanimity—after all, it invokes the consent of affected parties. This interpretation would render the axiom practically useless, since unanimity is rarely achievable.

We reject the unanimity reading. The axiom does not claim that *every* affected party must consent for a decision to be legitimate. It claims that legitimacy is a *function* of stakes-weighted consent, parameterized by threshold θ.

Different domains appropriately employ different thresholds:

- Constitutional amendments might require supermajority consent ($\theta = 0.67$ or higher)
- Legislative decisions might require simple majority consent ($\theta = 0.5$)
- Emergency decisions might permit lower thresholds when deliberation is infeasible
- Decisions affecting fundamental rights might require higher thresholds

The framework does not prescribe θ; it analyzes consequences given various θ values. This is another respect in which the framework is descriptive rather than normative: we predict friction patterns under different threshold choices without insisting that one threshold is uniquely correct.

2.7 The Friction-First Methodology

We now articulate the methodological innovation at the heart of the framework: the inversion of the traditional relationship between consent and friction.

2.7.1 Traditional Approach: Consent Primitive, Friction Derivative

Traditional moral and political philosophy treats consent as the primitive concept. Consent is something agents possess, grant, or withhold. Friction—conflict, resistance, instability—is a *derivative* phenomenon that arises when consent is violated.

On this traditional view:

1. Consent is structurally prior: agents have consent to give or refuse

2. Legitimacy is defined by consent: an arrangement is legitimate if agents consent to it

3. Friction is pathological: it indicates consent violation and calls for remedy

4. The goal is consent: we should seek arrangements to which all would consent

This approach faces severe difficulties. Consent is notoriously difficult to observe directly. How do we know if consent was genuine? Informed? Uncoerced? Free from manipulation? These questions admit no clean answers, and disputes about consent's presence or absence prove interminable.

2.7.2 Our Approach: Friction Observable, Consent Derived

We invert the traditional hierarchy:

Methodological Inversion

Traditional: Consent → Friction (consent is primitive, friction is derivative)
Friction-First: Friction → Consent (friction is observable, consent is derived)

On the friction-first approach:

1. **Friction is observable**: We can detect friction directly through behavioral indicators—protest, exit, litigation, noncompliance, sabotage, violence

2. **Friction is measurable**: We can quantify friction through proxies—turnover rates, litigation frequency, shadow economy size, emigration, regulatory reversals

3. **Consent is pattern-description**: "Consent" describes certain low-friction configurations; it is not a metaphysical property but an empirical pattern

4. **The goal is friction-minimization**: Rather than seeking unattainable perfect consent, we seek configurations that minimize destructive friction

This inversion has several methodological advantages:

Observability. We can observe friction directly through its behavioral manifestations. We cannot observe consent directly; we can only infer it from behavior and testimony, both of which are unreliable.

Measurability. Friction admits quantification through multiple proxies. Different friction types—corresponding to Hirschman's exit, voice, and loyalty categories (Hirschman, 1970)—can be tracked empirically. Consent, by contrast, resists quantification—what would it mean to measure "0.7 units of consent"?

Neutrality. Friction measurement requires no normative judgments. We observe that friction occurs without evaluating whether it is "justified." This descriptive stance enables empirical research without begging normative questions.

Universality. Friction appears in all shared domains regardless of cultural context. Disagreements about abortion, taxation, territorial boundaries, algorithm design, and family relationships all generate friction. The content differs; the dynamic is constant.

Substrate-agnosticism. The friction-first approach applies to any system with multiple agents and shared outcomes—humans, institutions, algorithms, hybrid systems. We need not resolve debates about who possesses the metaphysical capacity for consent; we simply observe friction patterns wherever they occur.

2.7.3 Empirical Tractability

The friction-first methodology enables empirical research that consent-first approaches cannot support. Consider the testable predictions that emerge:

1. **Legitimacy-friction correlation**: Domains with lower stakes-weighted voice exhibit higher friction indicators (protest, litigation, exit).

2. **Reconfiguration effects**: Reforms that increase alignment between consent-holding and stakes reduce friction.

3. **Duration effects**: Longer consent-holding durations predict greater friction upon reconfiguration (via the belief-transfer mechanism developed in subsequent sections).

4. **Threshold effects**: Friction increases discontinuously when legitimacy falls below critical thresholds.

These predictions are falsifiable. We can measure stakes distributions, voice allocations, and friction indicators across domains and jurisdictions. We can track changes over time following institutional reforms. We can compare predictions with observed outcomes.

This empirical tractability distinguishes the axiom of consent framework from purely philosophical approaches. We offer not only conceptual analysis but a research program with testable implications.

2.7.4 The Asymptotic Horizon

The friction-first approach reconfigures what "consent" means. Rather than a binary property that arrangements possess or lack, consent becomes an *asymptotic horizon*—a limit approached but never reached.

Perfect consent would require:

- Complete information (all affected parties know all relevant facts)

- Perfect communication (preferences are fully expressed and understood)

- Zero coercion (no party faces undue pressure)

- Full participation (all affected parties are included)

- Dynamic updating (consent tracks changing preferences)

No actual arrangement satisfies these conditions. Every consent is partial, imperfect, provisional. The question is never "was consent achieved?" but "how close did this configuration come?"

This asymptotic framing avoids the false binary that plagues consent discourse. We need not determine whether consent was "really" given; we can measure how much friction the configuration generates and compare it to alternatives. The direction matters—movement toward the consent horizon—even when arrival is impossible.

Table 1: The Kernel Triple Across Scales

Component	Symbol	Interpretation	Empirical Proxies
Alignment	α	Correlation of target functions	Survey congruence, revealed preference
Stake	σ	Magnitude of optimization at risk	Tax burden, employment dependence
Entropy	ε	Information loss in consent transfer	Transparency indices, misperception scores
Friction	F	System tension	Protest frequency, litigation rates
Legitimacy	L	Stakes-weighted voice	Franchise breadth, stakeholder inclusion

The framework's parsimony lies in this kernel triple: alignment, stake, and entropy. From these three measurable quantities, we derive friction and legitimacy. The same structure appears at every scale—from interpersonal relationships to international institutions—with scale-specific interpretations but invariant mathematical form.

This concludes the formal statement of the axiom of consent. In subsequent sections, we develop its dynamic implications and evolutionary mechanics (Section 4), and demonstrate applications across domains (Section 5).

3 The Kernel Triple Formalism

The axiom of consent provides a static characterization of legitimacy: voice should track stakes, mediated by alignment and entropy. But governance is dynamic—configurations evolve, institutions transform, norms shift. How does friction change over time? Under what conditions do consent-respecting arrangements emerge, persist, or dissolve? To answer these questions, we require a dynamical treatment that maps the axiom's core concepts onto evolutionary mechanics.

This section introduces the *kernel triple formalism*—the mathematical apparatus connecting the axiom's $(\alpha, \sigma, \varepsilon)$ structure to replicator-mutator dynamics via a scale-relative parameterization (ρ_S, w_S, M_S) (Hofbauer and Sigmund, 1998; Nowak, 2006). The key insight is structural: the axiom's alignment-stake-entropy triple instantiates directly into the survival-weight-mutation kernel governing type dynamics. Consent-respecting configurations are not imposed by fiat but emerge as attractors under selection pressure.

3.1 Motivation: From Statics to Dynamics

The friction function (Eq. 6) characterizes tension at a point in time. But political reality is evolutionary: revolutionary movements gain adherents or dissipate; institutional reforms propagate or stall; governance norms spread or contract. A purely static framework cannot address the central question of political theory: *why do some arrangements persist while others collapse?*

Three considerations motivate the dynamical extension:

Temporal evolution. Friction is not constant. As stakes shift, information flows change, and alignments drift, the friction landscape transforms. A configuration that minimizes friction today may become untenable tomorrow. We require equations governing $\partial F/\partial t$.

Types, not individuals. Political dynamics operate at multiple scales simultaneously. Individual preference changes matter, but so do institutional reforms, cultural shifts, and paradigm transitions. A statistical-mechanics approach—tracking distributions over *types* rather than individual trajectories—provides the appropriate level of abstraction. This parallels the move from Newtonian mechanics (individual trajectories) to thermodynamics (ensemble distributions) in physics.

Scale-relativity. Different observables become relevant at different resolutions. The voter and the polity occupy different scales; describing both requires scale-relative primitives. The atomic unit itself must become a parameter, not a fixed assumption.

3.2 The Scale-Relative Kernel

We formalize the notion of scale and the parameters defined relative to it.

Definition 3.1 (Scale)**.** A **scale** S specifies:

(i) A **type space** T_S—the set of distinguishable configurations at scale S

(ii) An **observable algebra** $\mathscr{O}_S$—the measurable quantities at scale S

(iii) A **resolution parameter** $r_S > 0$—the characteristic spatiotemporal granularity

Scales are observer-relative measurement choices, not objective features of reality. The same system admits description at multiple scales, with different scales revealing different dynamics. At agent scale, the atomic unit is an intentional agent; at institutional scale, the atomic unit is an institution; at cultural scale, the atomic unit is a belief system or practice.

Definition 3.2 (Atomic Agent)**.** Given scale S, the **atomic agent** Atom_S is the minimal unit of analysis—the entity treated as indivisible for purposes of description at that scale. Atomicity is not ontological fundamentality but resolution-relative non-decomposition.

This scale-relativity principle distinguishes the formalism from domain-specific applications. We are not claiming that "everything is selection" as metaphysics, but that many domains *admit* a selection-transmission description once one chooses appropriate scale-relative parameters. Recent work on evolutionary stability (Porter and Wikman, 2026) provides formal conditions under which such consent-holding configurations resist invasion by alternative arrangements, connecting the kernel triple's dynamical predictions to the classical stability theory of evolutionary game theory.

3.3 The Triple (ρ_S, w_S, M_S)

At each scale S, dynamics are governed by three functions constituting the *kernel triple*.

Definition 3.3 (Survival Function)**.** The **survival function** $\rho_S : T_S \times \mathscr{G}_S \times \Delta(T_S) \to [0,1]$ maps a type τ, an interaction network $G \in \mathscr{G}_S$, and a population state $p \in \Delta(T_S)$ to a survival probability:

$$\rho_S(\tau, G, p) \in [0,1] \tag{10}$$

The survival function captures frequency-dependent and density-dependent selection: a type's persistence probability depends on what other types exist and how they interact.

Definition 3.4 (Weight Function)**.** The **weight function** $w_S : T_S \to \mathbb{R}_{\geq 0}$ assigns to each type its **intrinsic weight**—baseline replication capacity or resource access:

$$w_S(\tau) \geq 0 \tag{11}$$

Weight captures advantages that persist regardless of competitive context: structural resource access, incumbency effects, network centrality advantages.

Definition 3.5 (Mutation Kernel)**.** The **mutation kernel** $M_S : T_S \times T_S \to [0,1]$ specifies transmission probabilities:

$$M_S(\tau' \to \tau) = \Pr(\text{type } \tau \text{ produced from } \tau') \tag{12}$$

Row-stochasticity requires $\sum_{\tau \in T_S} M_S(\tau' \to \tau) = 1$ for all τ'. The mutation kernel captures imperfect transmission: copies diverge from originals, reforms produce unintended variants, imitation introduces errors.

The kernel triple (ρ_S, w_S, M_S) provides a complete parameterization of selection-transmission dynamics at scale S. Different domains instantiate different kernel functions:

Table 2: Scale-Specific Instantiations of the Kernel Triple

Scale	**Atoms**$_S$	ρ_S **(Survival)**	M_S **(Mutation)**
Cellular	Cell	Replication rate	Point mutation, horizontal transfer
Organism	Individual	Darwinian fitness	Genetic recombination
Agent	Intentional agent	Strategy payoff	Learning, imitation
Institutional	Institution	Legitimacy$/(1+$Friction$)$	Reform, evolution
Cultural	Belief/practice	Transmission × retention	Copying error, reinterpretation

3.4 The ROM Update Equation

The kernel triple governs temporal evolution through the *replicator-optimization mechanism* (ROM) equation (Farzulla, 2025g).

Theorem 3.1 (Type Dynamics)**.** *Given scale S with type space T_S, kernel triple (ρ_S, w_S, M_S), and population distribution $p_t \in \Delta(T_S)$, the temporal evolution of type frequencies is:*

$$\boxed{\frac{dp_t(\tau)}{dt} = \sum_{\tau' \in T_S} p_t(\tau') \cdot w_S(\tau') \cdot \rho_S(\tau', G_{S,t}, p_t) \cdot M_S(\tau' \to \tau) - p_t(\tau) \cdot \bar{\phi}_t} \tag{13}$$

where the mean fitness $\bar{\phi}_t = \sum_{\tau' \in T_S} p_t(\tau') \cdot w_S(\tau') \cdot \rho_S(\tau', G_{S,t}, p_t)$ normalizes the dynamics, ensuring $\sum_\tau dp_t(\tau)/dt = 0$ and preserving the probability simplex.

The ROM equation is not novel—it is the weighted replicator-mutator equation, well-established in evolutionary game theory (Hadeler, 1981; Page and Nowak, 2002). What is novel is the explicit kernel parameterization enabling systematic cross-domain instantiation, and the specific consent-friction instantiation we develop below.

Interpretation of components:

- $p_t(\tau')$: Prevalence of type τ' at time t (population share)
- $w_S(\tau')$: Intrinsic weight of τ' (resource advantage)
- $\rho_S(\tau', G, p)$: Survival probability of τ' given network and population state
- $M_S(\tau' \to \tau)$: Probability that τ' produces τ in transmission
- $\bar{\phi}_t$: Mean fitness, ensuring normalization

The first term sums over all types that could produce τ via transmission, weighted by prevalence, intrinsic weight, survival, and transmission probability. The second term removes the current proportion of τ at rate equal to mean fitness. Types with above-average effective fitness (product of weight, survival, and transmission to self) increase in frequency; those below average decrease.

3.5 Consent Instantiation: Mapping $(\alpha, \sigma, \varepsilon)$ to (ρ, w, M)

The axiom of consent defines friction in terms of alignment (α), stake (σ), and entropy (ε). The kernel triple provides the dynamical scaffolding. The central theoretical contribution is establishing the *structural correspondence* between these formalisms.

Definition 3.6 (Consent-Weighted Survival)**.** In the consent domain, the survival function takes the **legitimacy-friction form**:

$$\rho_S^{\text{consent}}(\tau, G, p) = \frac{L(\tau)}{1 + F(\tau)} \tag{14}$$

where $L(\tau)$ is the legitimacy of configuration τ (distributional match between stakes and voice) and $F(\tau)$ is the friction generated by τ.

This functional form ensures that survival increases with legitimacy and decreases with friction. Configurations with perfect legitimacy ($L = 1$) and zero friction ($F = 0$) achieve $\rho = 1$. Configurations with zero legitimacy achieve $\rho = 0$ regardless of friction. The denominator $(1 + F)$ bounds survival probability appropriately.

Definition 3.7 (Consent-Friction Mapping)**.** The axiom's terms map to kernel components as follows:

Axiom Term	Kernel Term	Interpretation
α (alignment)	Encoded in ρ_S	Higher alignment → higher survival; alignment enters friction function inversely, thus survival positively
σ (stake)	w_S (weight)	Stakes determine influence on dynamics; higher stakes → larger weight in evolutionary pressure
ε (entropy)	M_S (mutation)	Information loss as type transition probability; high entropy → increased transmission noise

The correspondence is not metaphorical but structural. Alignment enters the survival function through friction: from Eq. 6, $F \propto (1+\varepsilon)/(1+\alpha)$, so high alignment reduces friction and increases $\rho_S = L/(1+F)$. Stakes operate as weights: configurations affecting high-stakes agents exert proportionally larger selection pressure. Entropy operates as mutation: information loss in consent transfers manifests as noisy transmission of institutional templates.

3.5.1 Stake as Weight

The mapping $\sigma \mapsto w_S$ requires elaboration. In the axiom, stakes $s_i(d)$ quantify an agent's sensitivity to outcomes in domain d. In the kernel formalism, weights $w_S(\tau)$ quantify a type's baseline replication capacity.

The connection operates through aggregation. For a configuration (type) τ representing a governance arrangement:

$$w_S(\tau) = \sum_{i \in \text{Stakeholders}(\tau)} s_i \cdot \mathbf{1}[i \text{ supports persistence of } \tau] \tag{15}$$

High-stakes stakeholders supporting an arrangement contribute more to its evolutionary weight. This captures the empirical observation that institutions with powerful supporters persist longer, ceteris paribus. The weight function thus aggregates individual stakes into type-level evolutionary advantage.

3.5.2 Entropy as Mutation

The mapping $\varepsilon \mapsto M_S$ captures how information loss affects institutional transmission. When consent-holders imperfectly understand delegator preferences (high ε), the policies they implement diverge from intended outcomes. This divergence is structurally equivalent to mutation in transmission.

Formally, the baseline mutation kernel $M_0(\tau' \to \tau)$ is modulated by entropy:

$$M_S(\tau' \to \tau) = M_0(\tau' \to \tau) \cdot \big(1 + \lambda \cdot \bar{\varepsilon}(\tau')\big) \tag{16}$$

where $\bar{\varepsilon}(\tau')$ is the average entropy across agents in configuration τ' and $\lambda > 0$ scales the effect. High entropy increases off-diagonal elements (transitions away from τ'), reflecting that information-impoverished configurations produce more transmission errors.

After modulation, row-stochasticity must be restored:

$$M_S(\tau' \to \tau) \leftarrow \frac{M_S(\tau' \to \tau)}{\sum_{\tau''} M_S(\tau' \to \tau'')} \tag{17}$$

3.5.3 Alignment as Survival Modulator

Alignment enters survival through the friction function. Recall the friction function (Definition 2.4):

$$F(\tau) = \sum_{i \in A} s_i \cdot \frac{1 + \varepsilon_i(\tau)}{1 + \alpha_i(\tau)} \tag{18}$$

The inverse dependence on alignment ensures that configurations with high consent-holder/stakeholder alignment generate lower friction and thus higher survival probability. We can express this modulation explicitly:

$$\rho_S^{\text{consent}}(\tau) = \rho_S^{\text{base}}(\tau) \cdot \exp\left(-\lambda \cdot F(\tau)\right) \tag{19}$$

where ρ_S^{base} is a baseline survival function and $\lambda > 0$ scales friction's effect. This exponential suppression form ensures:

- $\rho_S > 0$ for all finite friction
- $\rho_S \to 0$ as friction diverges
- Smooth interpolation between low and high friction regimes

3.6 Network Structure and Alignment Operationalization

The alignment function α captures correlation between consent-holder and stakeholder utilities. In practice, operationalizing α requires decomposing the utility correlation structure into interpretable components. The Local-Global (LoGo) decomposition from network science provides this operationalization.

3.6.1 The LoGo Decomposition

Consider the correlation matrix $\mathbf{C}$ of stakeholder utilities, where $C_{ij} = \text{corr}(U_i, U_j)$. Rather than treating this as an unstructured matrix, we decompose it into hierarchical components reflecting network topology.

Definition 3.8 (Local-Global Alignment Decomposition)**.** The aggregate alignment α decomposes as:

$$\alpha = \omega_L \cdot \alpha_{\text{local}} + \omega_G \cdot \alpha_{\text{global}} \tag{20}$$

where:

- α_{local}: Within-cluster utility correlation (agents with shared local interests)
- α_{global}: Cross-cluster backbone correlation (system-wide alignment on common objectives)
- ω_L, ω_G: Weights determined by network modularity, with $\omega_L + \omega_G = 1$

High modularity networks (many isolated clusters, weak cross-cluster links) imply $\omega_L \gg \omega_G$, so alignment is dominated by local coordination. Low modularity networks (hierarchical backbone, strong cross-cluster integration) imply $\omega_G \gg \omega_L$, so alignment depends on system-wide consensus.

3.6.2 Interpretation for Consent Dynamics

The LoGo decomposition reveals why some governance arrangements generate persistent friction despite apparently high local consent. When α_{local} is high but α_{global} is low, consent-holders may achieve legitimacy within their immediate constituency while generating friction with disconnected stakeholder groups.

Proposition 3.2 (Modular Governance Fragility)**.** *In high-modularity networks ($\omega_L \approx 1$), governance arrangements stable under local metrics exhibit fragility to cross-cluster shocks:*

$$\frac{\partial F}{\partial \alpha_{global}} \approx \frac{-\sigma(1+\varepsilon)}{(1+\alpha)^2} \cdot \omega_G \tag{21}$$

Small modularity reduction (increasing ω_G) can precipitate large friction increases if $\alpha_{global} \ll \alpha_{local}$.

This explains governance crises following integration: when previously isolated stakeholder groups become connected (through technology, migration, or institutional reform), hidden global misalignment becomes salient. The LoGo framework thus operationalizes α while providing predictive structure about friction dynamics under network evolution.

3.7 The Unification: Isomorphism of Structure

The central claim of this section is now statable precisely.

Theorem 3.3 (Axiom-Kernel Correspondence)**.** *The axiom of consent's kernel triple $(\alpha, \sigma, \varepsilon)$ and the ROM kernel triple (ρ_S, w_S, M_S) are structurally isomorphic under the consent-friction instantiation. Specifically:*

*(i) **Alignment-Survival:** α enters ρ_S through the friction function; higher alignment yields higher survival probability via $\rho_S = L/(1+F)$ where $F \propto (1+\varepsilon)/(1+\alpha)$.*

*(ii) **Stake-Weight:** σ determines w_S; aggregate stakes of supporting stakeholders constitute the type's evolutionary weight.*

*(iii) **Entropy-Mutation:** ε modulates M_S; information loss increases transmission noise, widening the mutation kernel's dispersion.*

Proof sketch. The correspondence follows from the definitions. For (i): substituting $F = \sigma(1+\varepsilon)/(1+\alpha)$ into $\rho_S = L/(1+F)$ yields survival as a function of alignment. For (ii): weighting type dynamics by stakeholder stakes is precisely the ROM weight function's role. For (iii): entropy's information-loss interpretation matches mutation's imperfect-transmission interpretation; both increase variance in type production. The structural isomorphism holds because both formalisms decompose the same phenomenon—consent-weighted persistence—into the same three components, viewed from static (axiom) and dynamic (kernel) perspectives. ■ ■

This unification is the key theoretical contribution. It shows that the axiom of consent is not merely a normative principle but has dynamical implications: consent-aligned configurations are evolutionarily favored. "Ought" connects to "is" not through logical derivation but through selection: what persists is constrained by what generates less friction.

3.8 Properties of the Consent-Kernel Dynamics

We now establish formal properties of the consent-friction instantiation.

Proposition 3.4 (Consent-Aligned Survival Advantage)**.** *Under the consent-friction instantiation, types with higher consent alignment exhibit higher survival probability:*

$$\alpha(\tau_1) > \alpha(\tau_2) \implies \rho_S(\tau_1) > \rho_S(\tau_2) \tag{22}$$

holding stakes, entropy, and legitimacy constant.

Proof. From Eq. 6, friction $F \propto (1+\varepsilon)/(1+\alpha)$. For fixed ε, increasing α decreases F. From Eq. 14, $\rho_S = L/(1+F)$. Decreasing F (with L held constant) increases ρ_S. ■ ■

Proposition 3.5 (Friction as Selection Pressure)**.** *Friction acts as negative selection pressure on type prevalence. Types generating high friction decrease in frequency, ceteris paribus:*

$$\frac{\partial}{\partial F(\tau)}\left[\frac{dp_t(\tau)}{dt}\right] < 0 \tag{23}$$

Proof. The ROM equation yields $dp_t(\tau)/dt$ proportional to $\rho_S(\tau) - \bar{\rho}$. Since $\rho_S = L/(1+F)$, $\partial \rho_S/\partial F = -L/(1+F)^2 < 0$. Higher friction reduces survival, reducing the growth rate differential. ■ ■

Proposition 3.6 (Consent Equilibrium Convergence)**.** *Under ergodicity conditions (irreducible, aperiodic mutation kernel; bounded survival function), the system converges to a stationary distribution p^* satisfying detailed balance. In the consent domain, p^* assigns higher mass to consent-respecting configurations.*

Proof sketch. Standard results on replicator-mutator dynamics (Page and Nowak, 2002) establish existence of stationary distributions under ergodicity. The mutation kernel's irreducibility ensures all types are accessible; aperiodicity prevents cycles. The stationary distribution p^* satisfies $\sum_{\tau'} p^*(\tau')w(\tau')\rho(\tau')M(\tau' \to \tau) = p^*(\tau)\bar{\phi}^*$ for all τ. Since consent-respecting types have higher ρ_S, they receive higher mass at equilibrium. Full proof requires specifying regularity conditions on L, F, and M; see Farzulla (2025g) for technical details. ■ ■

3.9 The Belief-Transfer Extension

A distinctive feature of the consent-friction instantiation is the *belief-transfer mechanism*: consent-holding duration affects subsequent dynamics. When an agent holds consent over a domain for extended periods, their subjective perception shifts from "holding consent for d" toward "owning authority over d." This psychological ownership accumulates over time and affects the mutation kernel.

Definition 3.9 (Ownership Accumulation)**.** The ownership-perception $O_A(d,t) \in [0,1]$ of agent A over domain d evolves as:

$$\frac{dO_A}{dt} = \beta \cdot (1 - O_A) \cdot \mathbf{1}[A \text{ holds consent for } d] \tag{24}$$

where $\beta > 0$ is the transfer rate.

This logistic-type equation ensures ownership saturates at 1 for long-tenure consent-holders. The accumulation rate β may vary with domain sensitivity, institutional design, or cultural context.

Definition 3.10 (Ownership-Modulated Mutation)**.** Ownership perception modulates the mutation kernel:

$$M_S(\tau' \to \tau) = M_0(\tau' \to \tau) \cdot \exp\left(-\gamma\left(\bar{O}(\tau') - \bar{O}(\tau)\right)\right) \tag{25}$$

where $\bar{O}(\tau)$ is average ownership-perception in configuration τ and $\gamma > 0$ is the entrenchment parameter.

This modulation creates two effects:

1. **Entrenchment:** Transitions *away* from high-ownership configurations are suppressed (the exponential is negative when $\bar{O}(\tau') > \bar{O}(\tau)$).
2. **Reform resistance:** Incumbents with accumulated ownership resist transitions that would reduce their authority.

The Arrhenius-like exponential form ensures transitions remain possible but increasingly difficult with ownership accumulation. This generates the prediction that regime transition probability decreases exponentially with incumbent tenure—a testable distinction from generic "institutional stickiness" explanations.

3.10 Coarse-Graining and Scale Transitions

The scale-relativity of the kernel triple raises the question of how dynamics at one scale relate to dynamics at another. The *coarse-graining operator* formalizes this relationship.

Definition 3.11 (Coarse-Graining Operator)**.** For scales S (fine) and S' (coarse), the coarse-graining operator $\pi_{S\to S'} : \Delta(T_S) \to \Delta(T_{S'})$ maps fine-grained type distributions to coarse-grained distributions.

Properties:

1. π is surjective but not injective (information loss)
2. Transitivity: $\pi_{S\to S''} = \pi_{S'\to S''} \circ \pi_{S\to S'}$

The central question is whether ROM structure is preserved under coarse-graining. This is not generally guaranteed—projecting dynamics onto coarser state spaces can introduce memory effects (Mori-Zwanzig structure) that break the Markovian replicator-mutator form.

Theorem 3.7 (Lumpability Conditions)**.** *ROM structure is preserved under coarse-graining $\pi : T_S \to T_{S'}$ if and only if:*

(i) ***Transition uniformity:*** *For all $\tau_i, \tau_k \in T_S$ with $\pi(\tau_i) = \pi(\tau_k)$, and all macro-types $T' \in T_{S'}$:*

$$\sum_{\tau_j : \pi(\tau_j) = T'} M_S(\tau_i \to \tau_j) = \sum_{\tau_l : \pi(\tau_l) = T'} M_S(\tau_k \to \tau_l) \tag{26}$$

(ii) ***Survival homogeneity:*** *$\rho_S(\tau_i) = \rho_S(\tau_k)$ whenever $\pi(\tau_i) = \pi(\tau_k)$.*

Under these conditions, the coarse-grained dynamics satisfy ROM with kernel $(\rho_{S'}, w_{S'}, M_{S'})$ where:

- $\rho_{S'}(T) = \rho_S(\tau)$ for any $\tau \in T$ (well-defined by condition ii)
- $w_{S'}(T) = \sum_{\tau\in T} w_S(\tau) p(\tau|T)$ (weighted by conditional distribution)
- $M_{S'}$ inherits transition rates from M_S

When lumpability fails, the coarse observer sees dynamics that appear non-Markovian—history-dependence emerges from integrating out fine-grained degrees of freedom. This connects to the Mori-Zwanzig formalism in statistical mechanics (Zwanzig, 1960) and explains why "emergence" appears mysterious: it is what coarse-graining looks like when lumpability conditions fail.

3.11 Summary: The Kernel Triple as Dynamical Backbone

This section established the kernel triple formalism connecting the axiom of consent to evolutionary dynamics. The key results are:

1. **Scale-relative parameterization:** The kernel triple (ρ_S, w_S, M_S) provides domain-general apparatus for selection-transmission dynamics at any scale.

2. **Consent instantiation:** The axiom's $(\alpha, \sigma, \varepsilon)$ maps structurally onto (ρ_S, w_S, M_S) via:

   - Alignment entering survival through the friction function
   - Stakes determining evolutionary weight
   - Entropy modulating the mutation kernel

3. **Selection for consent:** Consent-respecting configurations exhibit higher survival probability; friction acts as negative selection pressure.

4. **Belief-transfer dynamics:** Ownership accumulation modulates the mutation kernel, generating entrenchment effects and testable predictions about tenure-transition relationships.

5. **Scale coherence:** Under lumpability conditions, ROM structure is preserved across scales; when conditions fail, apparent emergence arises from coarse-graining.

The kernel triple provides the dynamical backbone for the axiom of consent. What remains is to examine its empirical implications (Section 6) and domain-specific applications (Section 5).

Full technical details, proofs, and computational validation appear in the companion paper (Farzulla, 2025g).

4 Core Dynamics

The kernel triple formalism established in Section 3 maps the axiom's static characterization onto evolutionary mechanics. We now develop the full dynamical treatment: how friction evolves, how legitimacy changes, and what equilibrium properties emerge. The central result is that consent-respecting configurations are not normative ideals imposed from without but *attractors* under selection pressure—what persists is constrained by what generates less friction.

4.1 The ROM Equation: Generalized Replicator-Mutator Dynamics

We begin by situating the ROM (Replicator-Optimization Mechanism) equation within the broader landscape of evolutionary dynamics, then establish its consent-specific instantiation.

4.1.1 Connection to Standard Replicator Dynamics

The classical replicator equation (Taylor and Jonker, 1978) governs frequency dynamics in populations under selection:

$$\frac{dp_t(\tau)}{dt} = p_t(\tau)\left[\pi(\tau, p_t) - \bar{\pi}(p_t)\right] \tag{27}$$

where $\pi(\tau, p_t)$ is the payoff to type τ given population state p_t, and $\bar{\pi}(p_t) = \sum_{\tau'} p_t(\tau')\pi(\tau', p_t)$ is mean population payoff. Types with above-average payoff increase in frequency; those below average decrease.

The replicator-mutator extension (Hadeler, 1981; Page and Nowak, 2002; Traulsen and Hauert, 2009; Metz et al., 1996) incorporates imperfect transmission, and recent work by Shen et al. (2026) extends these dynamics to reputation-based voluntary participation games where agents can choose whether to engage—a mechanism formally analogous to consent withdrawal in the present framework:

$$\frac{dp_t(\tau)}{dt} = \sum_{\tau'} p_t(\tau') \cdot \pi(\tau', p_t) \cdot M_S(\tau' \to \tau) - p_t(\tau) \cdot \bar{\pi}(p_t) \tag{28}$$

The first term sums over all types that could produce τ via transmission with probability $M(\tau' \to \tau)$, weighted by payoff. The second term maintains normalization.

The ROM equation (Farzulla, 2025g) generalizes this framework through three innovations:

1. **Decomposed fitness:** Rather than a monolithic payoff π, fitness decomposes into weight $w_S(\tau)$ (intrinsic resource access) and survival $\rho_S(\tau, G, p)$ (frequency- and density-dependent persistence).
2. **Network dependence:** Survival depends on the interaction network $G_{S,t}$, not merely population frequencies.
3. **Scale parameterization:** All components are explicitly indexed by scale S, enabling systematic cross-scale analysis.

Definition 4.1 (ROM Equation)**.** The **Replicator-Optimization Mechanism** governing type dynamics at scale S is:

$$\boxed{\frac{dp_t(\tau)}{dt} = \sum_{\tau' \in T_S} p_t(\tau') \cdot w_S(\tau') \cdot \rho_S(\tau', G_{S,t}, p_t) \cdot M_S(\tau' \to \tau) - p_t(\tau) \cdot \bar{\phi}_t} \tag{29}$$

where $\bar{\phi}_t = \sum_{\tau'} p_t(\tau') \cdot w_S(\tau') \cdot \rho_S(\tau', G_{S,t}, p_t)$ is mean effective fitness.

The product $w_S(\tau') \cdot \rho_S(\tau', G, p)$ replaces the payoff function π of classical replicator dynamics. This decomposition is not merely notational—it enables consent-specific instantiation where weight encodes stakes and survival encodes alignment-modulated legitimacy.

4.1.2 Consent-Specific ROM Instantiation

Under the consent-friction instantiation developed in Section 3.5, the ROM equation becomes:

Theorem 4.1 (Consent-Friction ROM)**.** *With kernel components instantiated as:*

$$w_S(\tau) = \sum_{i \in Stakeholders(\tau)} s_i \cdot \mathbf{1}[i \text{ } supports \text{ } \tau] \tag{30}$$

$$\rho_S(\tau, G, p) = \frac{L(\tau)}{1 + F(\tau)} \tag{31}$$

$$M_S(\tau' \to \tau) \propto M_0(\tau' \to \tau) \cdot (1 + \lambda \bar{\varepsilon}(\tau')) \tag{32}$$

the ROM equation becomes:

$$\frac{dp_t(\tau)}{dt} = \sum_{\tau'} p_t(\tau') \cdot \sigma(\tau') \cdot \frac{L(\tau')}{1 + F(\tau')} \cdot M_S(\tau' \to \tau) - p_t(\tau) \cdot \bar{\phi}_t \tag{33}$$

where $\sigma(\tau') = \sum_{i \in Stakeholders(\tau')} s_i$ is total stake and $\bar{\phi}_t$ normalizes.

This equation governs the evolutionary dynamics of consent-holding configurations. Configurations with high legitimacy L and low friction F achieve higher survival probability; those with high stakes σ from supporting stakeholders achieve higher weight. Together, these determine which configurations persist.

4.1.3 Connection to Learning Dynamics Literature

The ROM equation is a weighted replicator-mutator system (Hadeler, 1981; Page and Nowak, 2002). This connection is substantive, not merely notational. Recent work on learning dynamics in games (Pangallo et al., 2022; Galla and Farmer, 2013) demonstrates that convergence to equilibria is *non-generic*: most games produce cycles, chaos, or limit sets of positive measure rather than stable fixed points.

Galla and Farmer (2013) establish that even in simple two-player games, reinforcement learning dynamics exhibit deterministic chaos when the game matrix satisfies mild heterogeneity conditions. Pangallo et al. (2022) extend this to generic n-player games, showing that the best-response structure typically admits no globally attracting equilibria. These results suggest that evolutionary stability is the exception rather than the rule.

The consent-friction instantiation achieves convergence where generic games fail through two mechanisms:

1. **Potential structure:** The legitimacy-weighted survival function $\rho_S = L/(1+F)$ creates a quasi-potential landscape. Rather than arbitrary payoff matrices, the consent-friction fitness is derived from the scalar friction function, inducing approximate gradient flow dynamics. Configurations descend the friction surface rather than cycling through payoff-indifferent regions.
2. **Mutation regularization:** The entropy-modulated mutation kernel M_S provides regularization, smoothing the fitness landscape and preventing the discontinuous best-response dynamics that generate chaos in standard learning. The kernel ensures that type transitions are probabilistic rather than deterministic, damping the oscillatory modes that Galla and Farmer identify as chaos sources.

This explains why consent-respecting configurations emerge as attractors: they occupy basins in a friction-derived potential landscape, while misaligned configurations occupy saddles or repellers. The ROM framework inherits the mathematical structure of evolutionary game theory while avoiding its generic convergence failures through the specific functional form of consent-friction fitness.

4.2 Friction Dynamics

The friction function (Eq. 6) characterizes tension at a point in time. Political reality is dynamic: stakes shift, alignments drift, information channels open and close. We develop the temporal dynamics of friction.

4.2.1 The Friction Equation

From Definition 2.4, friction is $F = \sigma(1+\varepsilon)/(1+\alpha)$ (Eq. 6). Differentiating with respect to time:

Proposition 4.2 (Friction Evolution)**.** *The temporal evolution of friction is governed by:*

$$\frac{dF}{dt} = \frac{\partial F}{\partial \sigma}\frac{d\sigma}{dt} + \frac{\partial F}{\partial \alpha}\frac{d\alpha}{dt} + \frac{\partial F}{\partial \varepsilon}\frac{d\varepsilon}{dt} \qquad (34)$$

Substituting partial derivatives from Propositions 2.2–2.4:

$$\boxed{\frac{dF}{dt} = \frac{1+\varepsilon}{1+\alpha}\frac{d\sigma}{dt} - \frac{\sigma(1+\varepsilon)}{(1+\alpha)^2}\frac{d\alpha}{dt} + \frac{\sigma}{1+\alpha}\frac{d\varepsilon}{dt}} \tag{35}$$

Proof. Direct application of the chain rule to Eq. 6:

$$\frac{\partial F}{\partial \sigma} = \frac{1+\varepsilon}{1+\alpha}$$
$$\frac{\partial F}{\partial \alpha} = -\frac{\sigma(1+\varepsilon)}{(1+\alpha)^2}$$
$$\frac{\partial F}{\partial \varepsilon} = \frac{\sigma}{1+\alpha}$$

Substitution into the total derivative yields the result. ■ ■

This equation reveals the levers of friction change:

- **Stake dynamics** ($d\sigma/dt$): Entry of new stakeholders increases friction; exit decreases it. The coefficient $(1+\varepsilon)/(1+\alpha)$ implies that stake changes matter more when entropy is high or alignment is low.
- **Alignment dynamics** ($d\alpha/dt$): Improved alignment reduces friction; the negative coefficient reflects the friction-reducing effect of consent. The quadratic denominator implies that alignment improvements near the $\alpha \to -1$ pole have disproportionate effect.
- **Entropy dynamics** ($d\varepsilon/dt$): Improved information transmission (decreasing ε) reduces friction. Transparency initiatives, deliberation, and communication infrastructure all operate through this channel.

4.2.2 Stability Analysis

We analyze the stability of friction equilibria.

Definition 4.2 (Friction Equilibrium)**.** A **friction equilibrium** is a configuration $(\sigma^*, \alpha^*, \varepsilon^*)$ such that $dF/dt = 0$.

Proposition 4.3 (Equilibrium Conditions)**.** *A friction equilibrium obtains when:*

$$\frac{1+\varepsilon^*}{1+\alpha^*}\frac{d\sigma}{dt} = \frac{\sigma^*(1+\varepsilon^*)}{(1+\alpha^*)^2}\frac{d\alpha}{dt} - \frac{\sigma^*}{1+\alpha^*}\frac{d\varepsilon}{dt} \tag{36}$$

That is, friction stabilizes when stake growth is exactly offset by alignment improvements and entropy reduction.

Theorem 4.4 (Friction Stability)**.** *Let $(\alpha(t), \sigma(t), \varepsilon(t))$ evolve continuously with $\alpha \in (-1, 1]$, $\sigma \geq 0$, $\varepsilon \in [0, 1]$. Friction $F = \sigma(1+\varepsilon)/(1+\alpha)$ satisfies $dF/dt \leq 0$ if and only if:*

$$\frac{d\sigma}{dt} + \frac{\sigma}{1+\varepsilon}\frac{d\varepsilon}{dt} \leq \frac{\sigma}{1+\alpha}\frac{d\alpha}{dt} \tag{37}$$

Equivalently: friction decreases when the growth rate of the numerator $\sigma(1+\varepsilon)$ does not exceed the growth rate of the denominator $(1+\alpha)$.

Corollary 4.5 (Sufficient Conditions for Friction Decrease)**.** *Any of the following implies $dF/dt \leq 0$:*

(i) *Stakes non-increasing, entropy non-increasing, alignment non-decreasing:* $d\sigma/dt \leq 0$, $d\varepsilon/dt \leq 0$, $d\alpha/dt \geq 0$

(ii) *Logarithmic growth bound:* $\frac{d}{dt}\ln[\sigma(1+\varepsilon)] \leq \frac{d}{dt}\ln(1+\alpha)$

Proof of Theorem 4.4. From the chain rule, $dF/dt = \frac{1+\varepsilon}{1+\alpha}\frac{d\sigma}{dt} - \frac{\sigma(1+\varepsilon)}{(1+\alpha)^2}\frac{d\alpha}{dt} + \frac{\sigma}{1+\alpha}\frac{d\varepsilon}{dt}$. Multiplying by $(1+\alpha)/(1+\varepsilon) > 0$ and rearranging yields (37) as necessary and sufficient. ■ ■

Remark 4.3 (Why "Bounded Stakes" Is Insufficient). An earlier formulation claimed bounded stakes suffice for stability. This is false: stakes can approach a bound while still growing ($d\sigma/dt > 0$), contributing positively to dF/dt and potentially overwhelming alignment improvements. The corrected condition (37) makes explicit that governance adaptation rate must scale with stakeholder growth rate—a design principle explaining why rapid scaling often degrades governance quality even when intentions remain good.

This theorem provides precise conditions under which consent-respecting configurations emerge as attractors: when alignment improvement outpaces the combined growth of stakes and entropy, friction decreases monotonically.

4.3 Legitimacy Evolution

Friction characterizes tension; legitimacy characterizes acceptance. We develop the connection between these concepts and their co-evolution.

4.3.1 The Legitimacy Function

From the Doctrine of Consensual Sovereignty (Farzulla, 2025b), legitimacy quantifies the degree to which effective voice tracks stakes:

Definition 4.4 (Legitimacy)**.** The **legitimacy** of a consent-holding configuration in domain d at time t is:

$$L(d,t) = \frac{\sum_{i \in S_d} s_i(d) \cdot \text{eff\_voice}_i(d,t)}{\sum_{i \in S_d} s_i(d)} \tag{38}$$

where $S_d = \{i : s_i(d) > 0\}$ is the affected set and $\text{eff\_voice}_i \in [0,1]$ is agent i's effective voice over domain d.

Legitimacy is the stake-weighted mean of effective voice. When all stakeholders have voice proportional to stakes, $L = 1$. When voice is concentrated among low-stake agents while high-stake agents are excluded, $L < 1$.

4.3.2 Legitimacy-Friction Relationship

The relationship between legitimacy and friction is inverse but not symmetric.

Proposition 4.6 (Legitimacy-Friction Coupling)**.** *Legitimacy and friction are related by:*

$$\rho_S(\tau) = \frac{L(\tau)}{1+F(\tau)} \tag{39}$$

where $\rho_S(\tau)$ is survival probability for configuration τ. High legitimacy increases survival; high friction decreases it.

This coupling ensures that configurations cannot achieve high survival through legitimacy alone if they generate substantial friction, nor can they persist through low friction alone if they lack legitimacy. Both conditions are necessary for evolutionary success.

4.3.3 Legitimacy Dynamics

Theorem 4.7 (Legitimacy Evolution)**.** *Under the consent-friction dynamics, legitimacy evolves according to:*

$$\frac{dL}{dt} = \underbrace{\sum_{i \in S_d} \frac{s_i}{\sum_j s_j} \frac{d(\mathit{eff\_voice}_i)}{dt}}_{\text{voice dynamics}} + \underbrace{\sum_{i \in S_d} \frac{\mathit{eff\_voice}_i - L}{\sum_j s_j} \frac{ds_i}{dt}}_{\text{stake reweighting}} \tag{40}$$

Proof. Apply the quotient rule to Eq. 38. Let $N = \sum_i s_i \cdot \text{eff\_voice}_i$ and $D = \sum_i s_i$. Then:

$$\begin{aligned}
\frac{dL}{dt} &= \frac{1}{D}\frac{dN}{dt} - \frac{N}{D^2}\frac{dD}{dt} \\
&= \frac{1}{D}\sum_i \left(s_i \frac{d(\text{eff\_voice}_i)}{dt} + \text{eff\_voice}_i \frac{ds_i}{dt} \right) - \frac{L}{D}\sum_j \frac{ds_j}{dt} \\
&= \sum_i \frac{s_i}{D}\frac{d(\text{eff\_voice}_i)}{dt} + \sum_i \frac{\text{eff\_voice}_i - L}{D}\frac{ds_i}{dt}
\end{aligned}$$

which is the stated result. ■ ■

The two terms have distinct interpretations:

- **Voice dynamics:** Legitimacy increases when agents gain effective voice, weighted by their stakes. Democratic reforms, franchise expansion, and stakeholder empowerment operate through this channel.
- **Stake reweighting:** When agents with above-average voice ($\text{eff\_voice}_i > L$) gain stakes, legitimacy increases. When agents with below-average voice gain stakes, legitimacy decreases. This captures the legitimacy crisis that emerges when new stakeholders enter without corresponding voice.

4.4 Key Theorems

We now establish the central theoretical results of the consent-friction framework.

4.4.1 Theorem: Consent-Holding Necessity

Theorem 4.8 (Consent-Holding Necessity)**.** *In any domain d where a non-null outcome obtains, there exists at least one agent A such that A holds consent over d. That is, consent-holding is a structural invariant of multi-agent coordination, not a contingent feature of particular arrangements.*

Proof. Suppose domain d produces outcome $o \neq \varnothing$. By the definition of domain, d is a locus of decision where actions affect outcomes. Since o obtains, some action a was executed (even "doing nothing" is an action with consequences).

For action a to occur, some selection procedure π determined a from the space of possible actions $\mathscr{A}_d$. This procedure π may be:

- Concentrated: a single agent decides
- Distributed: multiple agents jointly decide
- Encoded: an algorithm or rule determines the outcome
- Random: a stochastic mechanism selects

In each case, some locus of control $H(d)$ determines which procedure applies. Even if π is "let randomness decide," the meta-decision to permit randomization was made by some agent. Define this locus as the consent-holder.

Therefore, $\exists A : A = H(d)$, the consent-holder for domain d. ■ ■

This theorem establishes that disputes about consent are never about *whether* consent-holding exists, but about *who* holds it and whether that holding is legitimate. The axiom of consent is not a normative ideal but a structural reality.

4.4.2 Theorem: Inevitable Friction

The next theorem is our "impossibility" result, connecting to Arrow's tradition in social choice.

Theorem 4.9 (Inevitable Friction). *For any domain d with $|S_d| \geq 2$ (at least two stakeholders) and heterogeneous preferences, there exists no consent-holding configuration achieving $F(d) = 0$ with $\sigma(d) >$ 0.*

Proof. From Proposition 2.1, $F = 0$ requires $\sigma = 0$. But $\sigma(d) = \sum_{i \in S_d} s_i(d)$, and $S_d = \{i : s_i(d) > 0\}$ by definition. Therefore $|S_d| \geq 2$ with positive stakes implies $\sigma(d) > 0$.

Contradiction. Therefore no configuration achieves $F = 0$ when $|S_d| \geq 2$ and $\sigma > 0$. ■ ■

Corollary 4.10 (Irreducible Friction). *The minimal achievable friction in domain d with positive stakes is:*

$$F_{\min}(d) = \frac{\sigma(d)}{2} \tag{41}$$

achieved when $\alpha = 1$ (perfect alignment) and $\varepsilon = 0$ (zero entropy).

Proof. Substitute $\alpha = 1, \varepsilon = 0$ into Eq. 6:

$$F_{\min} = \sigma \cdot \frac{1+0}{1+1} = \frac{\sigma}{2}$$

This is the global minimum since $\partial F / \partial \alpha < 0$ and $\partial F / \partial \varepsilon > 0$. ■ ■

The irreducible baseline $\sigma/2$ represents the *coordination cost of delegation*—the minimal friction inherent in having any consent-holder at all. Even perfectly aligned agents with perfect information incur transaction costs. This is not a design flaw but a structural feature of collective decision-making.

Remark 4.5 (Connection to Arrow's Impossibility). Arrow's theorem (Arrow, 1951) demonstrates that no social welfare function satisfies minimal fairness axioms without dictatorship. Our Theorem 4.9 is analogous: no consent-holding configuration achieves zero friction with positive stakes. The difference is interpretive: Arrow treats impossibility as a *problem* requiring domain restrictions or relaxed axioms. We treat inevitable friction as *proof*—confirmation that the axiom captures a structural truth about collective decision-making.

4.4.3 Theorem: Convergence to Consent-Respecting Equilibria

Theorem 4.11 (ROM Convergence). *Under the ROM dynamics (Eq. 29) with:*

(i) ***Irreducibility:*** *The mutation kernel M_S is irreducible (all types accessible from all other types)*

(ii) ***Aperiodicity:*** *The Markov chain induced by M_S is aperiodic*

(iii) ***Bounded survival:*** $0 < \rho_{\min} \leq \rho_S(\tau, G, p) \leq \rho_{\max} < \infty$ *for all τ, G, p*

the system converges to a unique stationary distribution $p^* \in \Delta(T_S)$ satisfying:

$$p^*(\tau) \propto \sum_{\tau'} p^*(\tau') \cdot w_S(\tau') \cdot \rho_S(\tau') \cdot M_S(\tau' \to \tau) \tag{42}$$

Moreover, types with higher survival probability $\rho_S = L/(1+F)$ receive higher mass in p^*.

Proof sketch. The ROM equation defines a continuous-time Markov process on the probability simplex $\Delta(T_S)$. Under irreducibility and aperiodicity, the embedded discrete chain is ergodic. Bounded survival ensures the flow is well-defined.

By the Perron-Frobenius theorem applied to the weighted transition matrix $Q(\tau', \tau) = w_S(\tau') \cdot \rho_S(\tau') \cdot M_S(\tau' \to \tau)$, there exists a unique positive eigenvector p^* with eigenvalue $\bar{\phi}^*$. This eigenvector is the stationary distribution.

For the moreover clause: consider two types τ_1, τ_2 with $\rho_S(\tau_1) > \rho_S(\tau_2)$ and identical weight and transmission. The ratio of stationary masses is:

$$\frac{p^*(\tau_1)}{p^*(\tau_2)} = \frac{\rho_S(\tau_1)}{\rho_S(\tau_2)} > 1$$

Thus higher survival yields higher stationary mass. ■ ■

Corollary 4.12 (Selection for Consent)**.** *In the consent-friction instantiation, the stationary distribution p^* assigns higher mass to configurations with:*

1. *Higher legitimacy L*
2. *Lower friction F*
3. *Higher stakeholder support (entering through w_S)*

This corollary is the central result connecting "ought" to "is": consent-respecting configurations are evolutionarily favored. What *should* happen (from normative perspectives emphasizing consent) is what *will* happen (under selection dynamics), not because of normative force but because consent-respecting arrangements survive.

4.5 Lumpability and Cross-Scale Dynamics

We established in Section 3.10 that the kernel triple operates at multiple scales. The question of how dynamics at one scale relate to dynamics at another is formalized through lumpability conditions.

Definition 4.6 (Exact Lumpability)**.** A partition $\mathscr{P} = \{T_1, \ldots, T_k\}$ of type space T_S is **exactly lumpable** for the ROM dynamics if for all $T_i, T_j \in \mathscr{P}$ and all $\tau, \tau' \in T_i$:

$$\sum_{\tau'' \in T_j} w_S(\tau) \cdot \rho_S(\tau) \cdot M_S(\tau \to \tau'') = \sum_{\tau'' \in T_j} w_S(\tau') \cdot \rho_S(\tau') \cdot M_S(\tau' \to \tau'') \tag{43}$$

Theorem 4.13 (Preservation of ROM Structure)**.** *If partition $\mathscr{P}$ is exactly lumpable for the ROM dynamics at scale S, then the coarse-grained dynamics at scale S' (induced by $\mathscr{P}$) also satisfy ROM with kernel triple $(\rho_{S'}, w_{S'}, M_{S'})$ where:*

$$\rho_{S'}(T_i) = \rho_S(\tau) \textit{ for any } \tau \in T_i \textit{ (well-defined by lumpability)} \tag{44}$$

$$w_{S'}(T_i) = \sum_{\tau \in T_i} w_S(\tau) \cdot p(\tau | T_i) \tag{45}$$

$$M_{S'}(T_i \to T_j) = \sum_{\tau \in T_i} \sum_{\tau' \in T_j} p(\tau | T_i) \cdot M_S(\tau \to \tau') \tag{46}$$

Proof. The coarse-grained state is $P_t(T_i) = \sum_{\tau \in T_i} p_t(\tau)$. Differentiating:

$$\begin{aligned}\frac{dP_t(T_i)}{dt} &= \sum_{\tau \in T_i} \frac{dp_t(\tau)}{dt} \\ &= \sum_{\tau \in T_i} \left[\sum_{\tau' \in T_S} p_t(\tau') \cdot w_S(\tau') \cdot \rho_S(\tau') \cdot M_S(\tau' \to \tau) - p_t(\tau) \cdot \bar{\phi}_t \right]\end{aligned}$$

Under exact lumpability, the transition rates from any $\tau' \in T_j$ to aggregate T_i are identical. Grouping by source partition:

$$\sum_{\tau \in T_i} \sum_{\tau' \in T_S} (\cdots) = \sum_{T_j \in \mathscr{P}} \sum_{\tau' \in T_j} p_t(\tau') \cdot w_S(\tau') \cdot \rho_S(\tau') \cdot \sum_{\tau \in T_i} M_S(\tau' \to \tau)$$

Define $M_{S'}(T_j \to T_i) = \sum_{\tau \in T_i} M_S(\tau' \to \tau)$ (well-defined by lumpability). The coarse dynamics become:

$$\frac{dP_t(T_i)}{dt} = \sum_{T_j} P_t(T_j) \cdot w_{S'}(T_j) \cdot \rho_{S'}(T_j) \cdot M_{S'}(T_j \to T_i) - P_t(T_i) \cdot \bar{\Phi}_t$$

which is ROM at scale S'. ■ ■

Remark 4.7 (Emergence and Lumpability Failure). When lumpability conditions fail, the coarse-grained dynamics are *not* Markovian—they exhibit history-dependence that arises from integrating out fine-grained degrees of freedom. This is the Mori-Zwanzig phenomenon (Zwanzig, 1960): apparent memory effects at coarse scales emerge from memoryless dynamics at fine scales when the projection operator is not lumpable.

"Emergence" is what lumpability failure looks like to a coarse observer. The framework thus provides a precise account of when macro-level dynamics are autonomous (lumpability holds) versus when they require micro-level specification (lumpability fails).

4.6 The Belief-Transfer Extension

A distinctive feature of consent dynamics is *temporal accumulation*: the longer an agent holds consent, the more they perceive ownership. This creates path-dependence in the mutation kernel.

Definition 4.8 (Ownership Dynamics)**.** The ownership-perception $O_A(d,t)$ of agent A over domain d evolves as:

$$\frac{dO_A}{dt} = \beta \cdot (1 - O_A) \cdot \mathbf{1}[A = H(d,t)] - \gamma \cdot O_A \cdot \mathbf{1}[A \neq H(d,t)] \tag{47}$$

where $\beta > 0$ is the accumulation rate and $\gamma > 0$ is the decay rate.

Ownership accumulates logistically while holding consent and decays exponentially when consent is lost. The asymmetry (β, γ may differ) captures the empirical observation that ownership perceptions develop slowly but erode slowly as well—the "sticky" nature of perceived authority.

Theorem 4.14 (Tenure-Transition Relationship)**.** *Let τ_H denote the tenure of current consent-holder H. The probability of regime transition decreases exponentially with tenure:*

$$\Pr(transition|\tau_H) \propto \exp(-\gamma \cdot O_H(\tau_H)) \tag{48}$$

where $O_H(\tau_H) = 1 - \exp(-\beta \tau_H)$ is the ownership level after tenure τ_H.

Proof. Regime transitions occur when the mutation kernel permits type change. From Eq. 25, transitions away from high-ownership configurations are suppressed by factor $\exp(-\gamma(\bar{O}(\tau') - \bar{O}(\tau)))$.

For the current holder's configuration τ', $\bar{O}(\tau') = O_H(\tau_H)$. Transition to a new configuration τ involves $\bar{O}(\tau) \approx 0$ (new holder has no accumulated ownership).

Therefore:

$$M_S(\tau' \to \tau) \propto \exp(-\gamma \cdot O_H(\tau_H))$$

Transition probability is proportional to this kernel element, yielding the stated result. ■ ■

This theorem generates the empirically testable prediction that regime longevity exhibits Weibull-like survival curves with increasing hazard as $\tau_H \to 0$ and decreasing hazard as $\tau_H \to \infty$. Long-tenured regimes become progressively harder to unseat—not because of inherent superiority but because of accumulated ownership perception.

4.7 Comparison with Standard Evolutionary Game Theory

We conclude this section by situating the ROM framework within the broader evolutionary game theory literature.

Table 3: ROM Framework vs. Standard Evolutionary Game Theory

Feature	**Standard Replicator Dynamics**	**ROM Framework**
Fitness	Monolithic payoff $\pi(\tau, p)$	Decomposed: $w_S(\tau) \cdot \rho_S(\tau, G, p)$
Network structure	Typically ignored or implicit	Explicit network $G_{S,t}$ in survival function
Scale	Fixed, typically agent-level	Parameterized by scale S; explicit coarse-graining
Mutation	Optional extension	Integral: entropy-modulated M_S
Normative content	Descriptive only	Bridge principle connecting fitness to legitimacy
Cross-domain	Domain-specific instantiation	Isomorphic structure across domains

The ROM framework's distinctive contribution is the consent-friction instantiation that maps alignment, stakes, and entropy onto the kernel components. This enables:

1. **Prediction:** Friction and legitimacy become observable quantities with testable dynamics.
2. **Design:** Institutional interventions can target specific kernel components (improve α, reduce ε, redistribute σ).
3. **Unification:** The same mathematics applies across political, economic, and computational domains.

4.7.1 Positioning Within Evolutionary Game Theory

Several features distinguish the consent-friction instantiation from standard EGT approaches and warrant explicit positioning relative to the literature.

Replicator dynamics and equilibrium selection. The classical replicator equation (Eq. 27) selects among strategies based on relative payoff. A standard result (Hofbauer and Sigmund, 1998; Weibull,

1995) is that replicator dynamics converge to Nash equilibria in potential games but may cycle or exhibit chaos in generic games (Galla and Farmer, 2013). The ROM equation inherits this structure but imposes additional constraint through the friction-derived fitness landscape. In the consent-friction instantiation, the payoff function is not arbitrary but derives from a scalar friction potential: $\rho_S = L/(1+F)$, where F is itself a function of alignment, stakes, and entropy. This structure makes the consent-friction game closer to a potential game than a generic game, explaining why convergence results (Theorem 4.11) hold under conditions that would produce chaos in unconstrained replicator dynamics.

Connection to tenable strategy blocks. In standard EGT, a *tenable strategy block* (Maynard Smith and Price, 1973) is a set of strategies that resists invasion by any strategy outside the block. The consent-friction framework generates an analogous structure: consent-respecting configurations form a "tenable consent block" in which no misaligned configuration can invade, because misalignment generates friction that reduces survival probability below the block's mean fitness. The friction function thus provides a *mechanism* for evolutionary stability that standard ESS analysis takes as primitive—it explains *why* certain strategy sets are invasion-resistant rather than merely identifying which ones are.

Extending beyond Nash and correlated equilibrium. Standard equilibrium concepts—Nash, correlated, evolutionary stable—describe configurations where no agent benefits from unilateral deviation. The consent-friction framework identifies a stronger property: configurations where deviation generates *structural resistance* (friction) from affected parties, not merely payoff reduction for the deviator. This is closer to Ostrom's institutional analysis (Ostrom, 1990), where rule violations trigger sanctions from the governance community, than to classical Nash reasoning where deviation is individually irrational. The ROM dynamics capture this distinction: consent-violating mutations face both fitness disadvantage (lower ρ_S) and active resistance (friction-generated selection pressure from high-stake agents who support the existing arrangement through w_S).

Mutation as structural feature. In classical replicator dynamics, mutation is an optional extension that blurs evolutionary stability. In the consent-friction framework, mutation is integral: entropy-modulated transmission noise (Eq. 16) captures the inherent imperfection of institutional reproduction. This connects to the replicator-mutator literature (Hadeler, 1981; Page and Nowak, 2002) but with a substantive interpretation: mutation rate is not a free parameter but is determined by the information structure of the consent relationship. Higher entropy means noisier institutional transmission, which is empirically measurable rather than theoretically arbitrary.

4.8 Summary

This section established the dynamical core of the consent-friction framework:

1. **ROM Equation:** Type dynamics governed by weighted replicator-mutator equation with consent-specific instantiation (Eq. 33).

2. **Friction Dynamics:** Temporal evolution of friction depends on stake, alignment, and entropy dynamics (Eq. 35).

3. **Legitimacy Evolution:** Legitimacy changes through voice dynamics and stake reweighting (Theorem 4.7).

4. **Key Theorems:**

   - Consent-Holding Necessity (Theorem 4.8): Consent-holding is unavoidable.
   - Inevitable Friction (Theorem 4.9): Zero friction is impossible with positive stakes.

- ROM Convergence (Theorem 4.11): System converges to consent-respecting equilibria.
- Tenure-Transition (Theorem 4.14): Regime transition probability decays with tenure.

5. **Scale Coherence:** ROM structure is preserved under lumpable coarse-graining (Theorem 4.13).

The dynamical treatment reveals that consent-respecting configurations are not normative ideals but evolutionary attractors. "Ought" connects to "is" through selection: what persists is constrained by what generates less friction. The framework provides both descriptive predictions and instrumental guidance for institutional design.

Full technical details, convergence proofs, and numerical validation appear in the companion paper (Farzulla, 2025g).

5 Domain Instantiations

The preceding sections established the formal machinery of the Axiom of Consent: primitive definitions, the kernel triple, and evolutionary dynamics. We now demonstrate that this machinery generates substantive predictions for multi-agent systems, with supplementary applications to other domains establishing cross-domain generality.

5.1 Multi-Agent Coordination

Multi-agent systems provide an ideal domain for the consent-friction framework. Resource allocation is explicit (encoded in protocols and access controls), interactions are observable (logged and monitorable), and the consequences of coordination failure are immediate and measurable.

5.1.1 Consent-Holding as Resource Authority

In multi-agent systems, "consent" translates to authority over resources that affect other agents.

Definition 5.1 (Resource Consent)**.** Agent i **holds consent** over resource r if i has authority to determine the allocation of r. The consent-holding configuration $C : \mathscr{R} \to \mathscr{A}$ maps resources to controlling agents.

Definition 5.2 (Stake in Resource)**.** Agent j's **stake** in resource r is:

$$\sigma_j(r) = \left| \frac{\partial U_j}{\partial \text{allocation}(r)} \right| \tag{49}$$

where U_j is agent j's utility function. High stake means resource allocation significantly impacts agent welfare.

Definition 5.3 (Allocation Alignment)**.** The **alignment** between consent-holder i and affected agent j for resource r is:

$$\alpha_{ij}(r) = \frac{\text{cov}(\nabla_r U_i, \nabla_r U_j)}{\sqrt{\text{var}(\nabla_r U_i) \cdot \text{var}(\nabla_r U_j)}} \tag{50}$$

This is the correlation between their utility gradients with respect to r's allocation. When $\alpha_{ij} = 1$, both agents want the same allocation; when $\alpha_{ij} = -1$, they want opposite allocations.

Definition 5.4 (Communication Entropy)**.** The **entropy** between agents i and j is:

$$\varepsilon_{ij} = H(U_j | \text{signal}_{i \to j}) \tag{51}$$

where $H(\cdot|\cdot)$ is conditional entropy. This measures how much uncertainty about j's preferences remains after i receives j's communication.

5.1.2 Coordination Friction

Friction in multi-agent systems manifests as coordination failure.

Theorem 5.1 (Multi-Agent Friction)**.** *For agent j affected by resource r controlled by agent i, the friction generated is:*

$$F_j(r) = \sigma_j(r) \cdot \frac{1+\varepsilon_{ij}(r)}{1+\alpha_{ij}(r)} \tag{52}$$

Total system friction is:

$$F_{system} = \sum_{r\in\mathscr{R}} \sum_{j\neq C(r)} F_j(r) \tag{53}$$

This friction manifests operationally as:

- **Communication overhead**: Agents expend resources negotiating, signaling, and verifying
- **Deadlock**: Conflicting resource claims create blocking conditions
- **Thrashing**: Agents repeatedly adjust allocations without convergence
- **Defection**: Agents exit cooperative arrangements or engage in adversarial behavior

5.1.3 Coordination Legitimacy

Legitimacy characterizes sustainable coordination.

Definition 5.5 (Coordination Legitimacy)**.** The **legitimacy** of consent-holding configuration C is:

$$L(C) = \frac{\sum_{r\in\mathscr{R}} \sum_{j\in\mathscr{A}} \sigma_j(r) \cdot v_j(r)}{\sum_{r\in\mathscr{R}} \sum_{j\in\mathscr{A}} \sigma_j(r)} \tag{54}$$

where $v_j(r) \in [0,1]$ is agent j's effective voice over resource r—influence on allocation decisions proportional to actual impact on outcomes.

Theorem 5.2 (Legitimacy-Stability Correspondence)**.** *Configurations with $L(C) > L^*$ (where L^* is a domain-specific threshold) are evolutionarily stable under ROM dynamics. Configurations with $L(C) < L^*$ generate accumulating friction and eventual reconfiguration.*

Proof. From the ROM equation (Eq. 33), the survival probability of configuration C is:

$$\rho(C) = \frac{L(C)}{1+F(C)} \tag{55}$$

For $L(C) < L^*$, accumulated friction $F(C)$ grows, decreasing $\rho(C)$ below the replacement threshold. Alternative configurations with higher legitimacy invade. The equilibrium is a configuration where legitimacy is maximized subject to structural constraints. ■

5.1.4 Application: Distributed Resource Allocation

Consider n agents competing for m shared resources with heterogeneous preferences. The standard approach assigns resources to maximize aggregate utility:

$$C^* = \arg\max_C \sum_j U_j(C) \tag{56}$$

The consent-friction framework predicts this allocation generates friction proportional to the degree of stakes-voice misalignment. A utilitarian allocation that ignores minority high-stake agents will face resistance from those agents.

Proposition 5.3 (Friction-Aware Allocation)**.** *The friction-minimizing allocation solves:*

$$C^{**} = \arg\min_{C} \left[\sum_{r} \sum_{j \neq C(r)} \sigma_j(r) \cdot \frac{1 + \varepsilon_{C(r),j}}{1 + \alpha_{C(r),j}(r)} \right] \tag{57}$$

This differs from utilitarian allocation when high-stake agents have low alignment with potential controllers.

Implication: Optimal allocation is not purely utilitarian but accounts for coordination costs. Assigning resources to agents with high alignment to affected parties reduces friction, even if immediate aggregate utility is lower.

5.1.5 Application: Multi-Agent Reinforcement Learning

In MARL, agents learn policies that interact through shared environments (Leibo et al., 2017; Lerer and Peysakhovich, 2019). The consent-friction framework provides a lens for understanding coordination failure.

Definition 5.6 (MARL Alignment)**.** For agents i, j with reward functions R_i, R_j, alignment is:

$$\alpha_{ij} = \frac{\mathbb{E}_{s,a}[R_i(s,a) \cdot R_j(s,a)] - \mathbb{E}[R_i]\mathbb{E}[R_j]}{\sigma_{R_i} \cdot \sigma_{R_j}} \tag{58}$$

the correlation of rewards over state-action pairs.

Theorem 5.4 (Coordination Convergence)**.** *A MARL system with mean pairwise alignment $\bar{\alpha}$ and mean communication entropy $\bar{\varepsilon}$ has expected coordination friction:*

$$\mathbb{E}[F] \propto \bar{\sigma} \cdot \frac{1 + \bar{\varepsilon}}{1 + \bar{\alpha}} \tag{59}$$

Systems with $\mathbb{E}[F] > F_{crit}$ fail to converge to stable joint policies.

Implication: MARL convergence depends on reward alignment α and communication capacity (inverse ε). Systems with misaligned rewards or bandwidth-constrained communication generate friction exceeding convergence thresholds.

5.1.6 Application: AI Alignment

The AI alignment problem (Russell, 2019; Bostrom, 2014) can be framed as consent-friction minimization between AI systems and human principals. Constitutional AI approaches (Bai et al., 2022) explicitly encode consent structures through principle hierarchies, representing a practical instantiation of the framework's alignment optimization.

Definition 5.7 (Human-AI Alignment)**.** For AI system A with learned objective U_A and human principal H with true preferences U_H:

$$\alpha_{AH} = \text{corr}(U_A, U_H) \tag{60}$$

Definition 5.8 (Interpretability as Entropy)**.** The interpretability deficit is:

$$\varepsilon_{AH} = H(U_A | \text{observables}_H) \tag{61}$$

Opaque AI systems have high ε; interpretable systems have low ε.

Theorem 5.5 (Alignment Friction). *An AI system with learned objective misaligned from human preferences generates friction:*

$$F_{AH} = \sigma_H \cdot \frac{1+\varepsilon_{AH}}{1+\alpha_{AH}} \tag{62}$$

where σ_H is the magnitude of human stakes in AI behavior.

Predictions:

1. **Interpretability reduces friction.** Investment in interpretability (reducing ε) is not just about safety but about coordination—reducing friction between AI and human principals.
2. **Misalignment generates resistance.** AI systems with $\alpha_{AH} < 1$ face behavioral correction pressure from humans, manifesting as constraints, shutdowns, or adversarial responses.
3. **Stake magnitude matters.** High-stakes AI applications (medical, legal, financial) generate more friction per unit misalignment than low-stakes applications.
4. **Relational constitution of alignment.** Human-AI relationships that exhibit genuine interaction patterns constitute relationships with moral weight, implying that alignment is not merely technical but relational (Farzulla, 2025f). AI systems with embodied autonomy possess the functional properties that make unconsented rule illegitimate (Farzulla, 2025c). Empirical evidence supports this behavioral framing: Salatino et al. (2025) demonstrate that AI behavior—not attributed sentience—drives human moral judgments, while O'Reilly et al. (2025) show that action descriptions shape moral responsibility attribution to robots. Leibo et al. (2025) argue for treating personhood itself as a governance tool—bundles of rights and obligations—rather than metaphysical status, converging with our stakes-based approach.

5.1.7 ROM Dynamics in Agent Populations

The ROM equation governs evolution of coordination strategies in agent populations.

Theorem 5.6 (Strategy Evolution). *Let $\tau \in \mathscr{T}$ index coordination strategies (consent-holding configurations). Under ROM dynamics:*

$$\frac{dp_t(\tau)}{dt} = \sum_{\tau'} p_t(\tau') \cdot \sigma(\tau') \cdot \frac{L(\tau')}{1+F(\tau')} \cdot M_S(\tau' \to \tau) - p_t(\tau) \cdot \bar{\phi}_t \tag{63}$$

Strategies with high legitimacy L and low friction F increase in frequency. Strategies generating unsustainable friction are selected against.

Corollary: In multi-agent systems with learning and adaptation, consent-respecting configurations are evolutionary attractors. Not because they are "good" but because they are stable—they generate less friction and persist longer.

5.2 Supplementary Applications

The consent-friction framework applies beyond multi-agent systems. We briefly sketch two additional domains to demonstrate cross-domain generality.

5.2.1 Cryptocurrency Governance

Cryptocurrency markets instantiate consent-friction dynamics with observable precision (De Filippi and Wright, 2018; Allen et al., 2020; Buterin, 2017). Governance structures are encoded in protocols; transactions are recorded on public blockchains; and friction manifests as price volatility.

Key Results. In Farzulla (2025e), I document that infrastructure disruption events generate 5.7× larger volatility shocks than regulatory uncertainty events. The consent-friction framework explains this differential:

- **Infrastructure events** (exchange hacks, protocol failures) have high alignment across token holders—near-universal agreement that failures are bad. High $\alpha \approx 0.9$ means holders coordinate rapid response.
- **Regulatory events** have heterogeneous alignment—some holders welcome regulation, others oppose it. Mixed $\alpha \approx 0.3$ creates coordination fragmentation.

The volatility differential reflects *correlated* friction: high-alignment events produce synchronized responses; low-alignment events produce fragmented responses.

Predictions. Governance-aligned protocol changes (community-approved upgrades) generate lower friction than governance-violating changes (contentious forks, external impositions). This provides testable predictions for cryptocurrency market dynamics. The Aggregated Systemic Risk Index (ASRI) (Farzulla and Maksakov, 2025) demonstrates that systemic risk in cryptocurrency markets emerges from distributed friction sources rather than localized protocol failures—a direct application of the friction aggregation apparatus developed in Section 3.

Case Study: Terra-Luna (May 2022). The Terra-Luna collapse provides a natural experiment for the friction framework with quantifiable parameters. Prior to May 2022, UST maintained its dollar peg through an algorithmic relationship with LUNA. In framework terms:

- **Stakes σ**: Approximately \$40B in UST market cap at peak
- **Alignment α**: Estimated 0.6–0.8 (users believed algorithmic peg was aligned with their stability interests)
- **Entropy ε**: Moderate ($\approx$ 0.3–0.5) due to mechanism complexity creating information asymmetry about fragility conditions
- **Legitimacy L**: High (market confidence, institutional adoption, ecosystem growth)

The depeg sequence unfolded as a friction-legitimacy spiral:

1. **Information shock:** Large redemptions created an ε shock, revealing mechanism fragility that was previously hidden in mechanism complexity.
2. **Alignment collapse:** As the depeg began, α collapsed from ≈ 0.7 toward 0 as users recognized misalignment between the algorithmic mechanism and peg maintenance under stress.
3. **Friction explosion:** Substituting approximate values: $F = 40\text{B} \times (1+0.5)/(1+0.2) \approx 50\text{B}$ equivalent friction—exceeding any institutional tolerance threshold.
4. **Legitimacy collapse:** $L \to 0$ as stake-weighted confidence evaporated within days.

The speed of collapse—from stable peg to near-zero value in approximately 72 hours—reflects the friction amplification mechanism: high-σ domains are catastrophically fragile to joint α/ε shocks. This validates the multiplicative structure of the friction function.

Case Study: FTX (November 2022). The FTX collapse instantiates a *consent-violation* case with different parameter dynamics. FTX, as a centralized exchange, held consent over customer assets through custodial relationships. In framework terms:

- **Stakes σ**: Approximately \$8–\$16B in customer deposits
- **Alignment α** (pre-revelation): Perceived ≈ 0.9 (fiduciary duty, regulatory compliance signals)
- **Entropy ε** (pre-revelation): Moderate (≈ 0.4) due to proprietary trading operations

The critical difference from Terra-Luna is the *direction* of information revelation:

1. **Entropy elimination:** Revelation reduced $\varepsilon \to 0$—full information about misappropriation became available.
2. **Alignment inversion:** Simultaneously, α collapsed from perceived $+0.9$ to revealed ≈ -0.8—the consent-holder had been actively adversarial to delegating stakeholders.
3. **Boundary approach:** As $\alpha \to -1$, the friction function approaches infinity: $F = \sigma(1+\varepsilon)/(1+\alpha) \to \infty$ as $(1+\alpha) \to 0$.

The framework predicts that adversarial consent-holders ($\alpha < 0$) generate unbounded friction when information ($\varepsilon \to 0$) reveals the adversarial relationship. This matches the observed market response: contagion spread rapidly through interconnected entities, with transfer entropy analysis (Schreiber, 2000; Bossomaier et al., 2016) revealing directional information flow from FTX to exposed counterparties.

Both cases validate the framework's prediction that cryptocurrency markets, with their explicit governance encoding and observable transactions, provide natural laboratories for testing consent-friction dynamics.

5.2.2 Political Legitimacy

Political systems provide the original domain for consent analysis. The kernel triple operationalizes legitimacy as stakes-weighted voice (Farzulla, 2025b).

Key Mapping.

- **Alignment α**: Correlation between citizen preferences and policy outcomes
- **Stakes σ**: Magnitude of citizen welfare at stake in policy decisions
- **Entropy ε**: Information asymmetry between citizens and governors
- **Friction F**: Protest, instability, resistance, revolution
- **Legitimacy L**: Stake-weighted effective voice

Predictions. Political systems with systematic stakes-voice misalignment (high-stake populations with low voice) generate accumulating friction. The ROM equation predicts eventual reconfiguration—revolution, reform, or collapse—when friction exceeds institutional tolerance.

5.2.3 Structural Isomorphism

The claim is not analogy but identity. Table 4 presents the kernel triple across domains.

Table 4: The Kernel Triple Across Domains

Component	Multi-Agent	Cryptocurrency	Political
Alignment (α)	Reward correlation	Holder-protocol match	Citizen-policy match
Stake (σ)	Utility exposure	Token value	Affected interests
Entropy (ε)	Communication overhead	Market uncertainty	Information asymmetry
Friction (F)	Coordination failure	Volatility	Instability
Legitimacy (L)	Sustainable coordination	Governance acceptance	Democratic legitimacy

The same equations govern all three domains:

$$F = \sigma \cdot \frac{1+\varepsilon}{1+\alpha} \tag{64}$$

$$L = \frac{\sum_i s_i \cdot v_i}{\sum_i s_i} \tag{65}$$

$$\frac{dp(\tau)}{dt} = \sum_{\tau'} p(\tau') \cdot \sigma(\tau') \cdot \frac{L(\tau')}{1+F(\tau')} \cdot M_S(\tau' \to \tau) - p(\tau) \cdot \bar{\phi} \tag{66}$$

These are the same mathematical objects with domain-specific variable interpretations. This unification suggests consent-friction dynamics capture something fundamental about coordination under preference heterogeneity and information asymmetry.

6 Measurement Apparatus

The preceding sections established the formal machinery of the consent-friction framework: the kernel triple $(\alpha, \sigma, \varepsilon)$, the friction equation $F = \sigma(1+\varepsilon)/(1+\alpha)$, and the evolutionary dynamics governing consent configurations. This section addresses a prior question: *how do we measure these quantities?*

Theoretical elegance is worthless without empirical tractability. The framework's value depends on whether its primitives admit operationalization—whether alignment, stakes, and entropy can be observed, quantified, and tracked across domains. We argue they can, though each presents distinct measurement challenges requiring domain-specific instrumentation.

6.1 The Measurement Problem

The kernel triple contains three latent variables: alignment (α), stakes (σ), and entropy (ε). None is directly observable. Alignment is a correlation between target functions that agents may not articulate or even consciously represent. Stakes are subjective valuations of consequences. Entropy is information loss across channels that may be opaque to both sender and receiver.

This indirectness is not unique to our framework—utility, preference, and welfare are similarly latent. The standard solution is *revealed preference*: infer latent states from observable behavior (Samuelson, 1938; Richter, 1966). We adapt this approach while acknowledging its limitations.

Three methodological principles guide our operationalizations:

Principle 1: Multiple Proxies. No single observable perfectly captures any kernel component. We use batteries of indicators, triangulating toward the latent variable.

Principle 2: Domain Specificity. The *concept* of alignment is domain-invariant; the *measurement* is domain-specific. Measuring voter-policy alignment requires different instruments than measuring holder-protocol alignment or agent-reward alignment.

Principle 3: Error Quantification. Measurement error is inevitable. We specify expected error distributions and how they propagate through the friction equation.

6.2 Operationalizing Alignment (α)

Alignment measures the correlation between an agent's target function and the consent-holder's target function. When $\alpha = 1$, the consent-holder optimizes for exactly what the agent wants. When $\alpha = -1$, perfect adversarial misalignment. When $\alpha = 0$, the consent-holder's optimization is orthogonal to the agent's interests.

Operationalization requires measuring two target functions and computing their correlation. We distinguish three approaches.

6.2.1 Survey-Based Preference Elicitation

The most direct approach elicits preferences through structured instruments.

Definition 6.1 (Preference Survey Alignment)**.** Let $\mathbf{p}_i \in \mathbb{R}^k$ be agent i's elicited preference vector over k outcome dimensions, and $\mathbf{p}_H \in \mathbb{R}^k$ be the consent-holder's stated or revealed policy position. Survey-based alignment is:

$$\alpha_i^{\text{survey}} = \frac{\mathbf{p}_i \cdot \mathbf{p}_H}{\|\mathbf{p}_i\| \cdot \|\mathbf{p}_H\|} \tag{67}$$

Instrument design. Survey items must span the outcome space relevant to the domain. In political contexts, this includes standard policy preference batteries: redistribution, regulation, social issues, foreign policy. In cryptocurrency contexts, items cover protocol parameters: block size, fee structures, governance mechanisms, fork preferences. In AI contexts, items operationalize reward function components: task prioritization, safety constraints, user preference learning.

Example: Political Alignment. The Comparative Study of Electoral Systems (CSES) provides multi-dimensional preference data for voters and parties across 50+ democracies. We compute:

$$\alpha_i^{\text{CSES}}(t) = 1 - \frac{D(\mathbf{p}_i, \mathbf{p}_{G(t)})}{D_{\max}} \tag{68}$$

where $D(\cdot,\cdot)$ is Euclidean distance in policy space, $G(t)$ denotes the governing coalition at time t, and $D_{\max}$ normalizes to $[-1,1]$.

Limitations. Survey data suffer from well-documented biases: social desirability, acquiescence, satisficing (Krosnick, 1999). Elicited preferences may diverge from revealed preferences; stated and actual target functions may differ. Survey timing matters: preferences shift, and a snapshot may misrepresent dynamic alignment.

6.2.2 Revealed Preference Alignment

Revealed preference approaches infer target functions from observed behavior rather than stated preferences.

Definition 6.2 (Behavioral Alignment)**.** Let $a_i \in \mathscr{A}$ be agent i's observed actions and $a_H \in \mathscr{A}$ be the consent-holder's policy actions. Behavioral alignment is:

$$\alpha_i^{\text{behavioral}} = \text{corr}(U_i(a_H), U_i(a_i^*)) \tag{69}$$

where U_i is agent i's inferred utility function and a_i^* is the action i would have chosen given the consent-holder's resources and constraints.

Market-based measurement. In financial domains, revealed preferences emerge through trading behavior. If agent i increases position after policy announcement P, this reveals P's alignment with i's interests. Formally:

$$\alpha_i^{\text{market}}(P) = \text{sign}(\Delta h_i(P)) \cdot \left(1 - e^{-\lambda|\Delta h_i(P)|}\right) \tag{70}$$

where $\Delta h_i(P)$ is the change in i's holdings following announcement P and λ scales the sensitivity.

Voting-based measurement. Electoral choices reveal policy preferences. We can estimate alignment through:

$$\alpha_i^{\text{vote}} = \sum_{p \in \text{Parties}} v_{ip} \cdot \cos(\mathbf{p}_p, \mathbf{p}_G) \tag{71}$$

where v_{ip} is i's vote share for party p (typically 1 for the voted party, 0 otherwise), $\mathbf{p}_p$ is party p's policy position, and $\mathbf{p}_G$ is the governing coalition's position.

Limitations. Revealed preferences are confounded by constraints. An agent who does not exit a market may be aligned, or may face exit costs. An agent who votes for party P may endorse P's platform, or may engage in lesser-evil reasoning. Revealed preference approaches require careful attention to the choice architecture that constrains revealed behavior.

6.2.3 Cross-Domain Alignment Indices

For aggregate analysis, we require summary measures that combine individual alignments.

Definition 6.3 (Aggregate Alignment Index)**.** The **aggregate alignment index** for domain d at time t is:

$$\alpha(d,t) = \frac{\sum_{i \in S_d} s_i(d) \cdot \alpha_i(d,t)}{\sum_{i \in S_d} s_i(d)} \tag{72}$$

where S_d is the set of affected agents, $s_i(d)$ is agent i's stake, and $\alpha_i(d,t)$ is individual alignment.

This stakes-weighted average ensures that high-stakes agents' alignment contributes proportionally more to the aggregate. A consent-holder perfectly aligned with one low-stakes agent but misaligned with many high-stakes agents will exhibit low aggregate alignment. Empirical evidence for friction dynamics in human-AI interaction is emerging: Lopez-Lopez et al. (2026) demonstrate that metacognitive monitoring in entangled human-AI decision-making exhibits friction patterns consistent with this axiomatization, providing the strongest empirical validation from cognitive science that alignment-stake interactions generate measurable behavioral signatures.

6.3 Operationalizing Stakes (σ)

Stakes quantify the magnitude of an agent's optimization at risk in a domain. High-stakes agents have much to gain or lose from domain outcomes; low-stakes agents are relatively indifferent.

Operationalization varies by domain type. We distinguish three categories.

6.3.1 Economic Stakes

In economic domains, stakes are quantifiable in monetary terms.

Definition 6.4 (Monetary Stake)**.** Agent i's monetary stake in domain d is:

$$\sigma_i^{\text{econ}}(d) = \mathbb{E}[|W_i(O) - W_i(O')|] \tag{73}$$

where W_i is i's wealth function and the expectation is over the distribution of possible outcomes (O, O').

Direct exposure measures. Portfolio holdings, property ownership, employment income—these provide direct measures of monetary stakes. A holder with \$1M in cryptocurrency has higher stakes in protocol governance than a holder with \$100.

Sunk cost adjustments. Stakes include not just current exposure but irreversible commitments. An employee with 20 years of firm-specific human capital has higher stakes than a new hire with transferable skills, even at identical current compensation.

Present value calculations. Future exposure must be discounted:

$$\sigma_i^{\text{PV}}(d) = \sum_{t=0}^{\infty} \delta^t \cdot \sigma_i^{\text{econ}}(d,t) \tag{74}$$

where $\delta \in (0,1)$ is the discount factor.

6.3.2 Political Stakes

In political domains, stakes involve interests that resist direct monetization.

Definition 6.5 (Political Stake)**.** Agent i's political stake in domain d is:

$$\sigma_i^{\text{pol}}(d) = \phi(\text{proximity}_i(d), \text{reversibility}(d), \text{magnitude}(d)) \tag{75}$$

where ϕ is an aggregation function combining proximity to the decision, reversibility of the outcome, and magnitude of life impact.

Proximity measures. Policy decisions affect some populations directly, others indirectly. A policy restricting immigration affects prospective immigrants directly and employers of immigrants indirectly. Proximity can be operationalized through causal pathway analysis: how many steps between decision and consequence?

Affected interest inventories. Following Goodin (2007), we can enumerate affected interests:

- *Vital interests*: life, bodily integrity, basic liberty
- *Important interests*: livelihood, family, community membership
- *Significant interests*: property, opportunity, recognition
- *Peripheral interests*: convenience, preference satisfaction

Stakes weight by interest category, with vital interests weighted highest.

6.3.3 Computational Stakes

In multi-agent and AI contexts, stakes involve resource allocation and task criticality.

Definition 6.6 (Computational Stake)**.** Agent i's computational stake in task allocation d is:

$$\sigma_i^{\text{comp}}(d) = c(d) \cdot \pi_i(d) \tag{76}$$

where $c(d)$ is the criticality weight of domain d and $\pi_i(d)$ is agent i's performance sensitivity to d.

Resource allocation. In distributed systems, resource allocation (compute, memory, bandwidth) determines task performance. An agent allocated insufficient resources for a high-criticality task has high stakes in resource governance.

Task criticality. Some tasks are safety-critical; failure is catastrophic. Others are optimization-oriented; suboptimality is tolerable. Criticality can be operationalized through loss functions:

$$c(d) = \max_{o \in \mathcal{O}(d)} L(o) - \min_{o \in \mathcal{O}(d)} L(o) \tag{77}$$

where $\mathcal{O}(d)$ is the outcome space and L is the loss function.

6.4 Operationalizing Entropy (ε)

Entropy captures information loss in the consent-holding relationship. Even perfectly aligned consent-holders generate friction when they do not *know* what affected agents want.

We operationalize entropy through three approaches.

6.4.1 Information-Theoretic Measures

The most direct operationalization uses Shannon entropy.

Definition 6.7 (Preference Entropy)**.** The preference entropy between agent i and consent-holder H is:

$$\varepsilon_i^{\text{info}} = H(\mathbf{p}_i|\hat{\mathbf{p}}_i^H) = -\sum_k p_{ik} \log \frac{p_{ik}}{\hat{p}_{ik}^H} \tag{78}$$

where $\mathbf{p}_i$ is agent i's true preference distribution, $\hat{\mathbf{p}}_i^H$ is the consent-holder's estimate of i's preferences, and the sum is over preference dimensions.

This is the Kullback-Leibler divergence between true and estimated preferences—a measure of information loss in preference transmission.

Measurement challenges. Computing $\varepsilon^{\text{info}}$ requires knowing both the true preference distribution and the consent-holder's estimate. The former is itself latent (hence the survey/behavioral approaches above); the latter requires access to the consent-holder's internal model.

Proxy: Model Uncertainty. When direct measurement is infeasible, we can use uncertainty measures in the consent-holder's preference model. If H maintains a Bayesian posterior over i's preferences:

$$\varepsilon_i^{\text{uncertainty}} = H(\mathbf{p}_i|\mathscr{D}_i) = \text{entropy of posterior} \tag{79}$$

Higher posterior entropy indicates greater information loss.

6.4.2 Communication Bandwidth Constraints

Information loss often stems from communication constraints.

Definition 6.8 (Channel Entropy)**.** The channel entropy in consent relationship (i,H) is:

$$\varepsilon_{i,H}^{\text{channel}} = 1 - \frac{I(X_i;Y_H)}{H(X_i)} \tag{80}$$

where X_i is agent i's preference signal, Y_H is the consent-holder's received signal, and $I(\cdot;\cdot)$ is mutual information.

This measures the proportion of preference information lost in transmission. When $I(X_i;Y_H) = H(X_i)$, no information is lost ($\varepsilon = 0$). When $I(X_i;Y_H) = 0$, transmission is pure noise ($\varepsilon = 1$).

Operationalization via questionnaires. In practice, we can estimate channel entropy through consistency checks. Ask agents to state preferences; ask consent-holders to predict agent preferences. Discrepancy between stated and predicted preferences operationalizes channel entropy.

6.4.3 Proxy Variables for Information Asymmetry

When direct information-theoretic measurement is infeasible, proxy variables can approximate entropy.

Transparency indices. Organizations and governments vary in transparency. Freedom of information laws, disclosure requirements, and institutional openness create conditions for lower entropy. Transparency indices (e.g., Open Government Partnership scores) proxy for ε in political domains.

Misperception scores. Survey instruments can measure systematic misperception. If citizens systematically misunderstand government policy, or if protocol developers systematically misunderstand holder preferences, these misperceptions proxy for entropy.

$$\varepsilon^{\text{misperception}} = \frac{1}{n}\sum_i \|\hat{\mathbf{p}}_i^H - \mathbf{p}_i\|^2 \tag{81}$$

Complexity measures. Protocol complexity, regulatory complexity, and algorithmic opacity increase entropy by limiting comprehension. Complexity indices (lines of code, Flesch-Kincaid readability scores, regulatory burden metrics) proxy for comprehension barriers.

6.4.4 Normalization and Domain-Specific Mapping

Normalization. Throughout this paper, $\varepsilon \in [0,1]$ is normalized such that $\varepsilon = 0$ represents perfect information (consent-holder has complete knowledge of stakeholder preferences) and $\varepsilon = 1$ represents maximum ignorance (consent-holder has no information beyond priors). For KL-divergence measures, normalize by maximum possible divergence in the domain; for channel capacity measures, the definition naturally produces values in $[0,1]$.

Domain-specific operationalization. Table 5 summarizes recommended ε operationalizations across domains.

Table 5: Entropy (ε) Operationalization by Domain

Domain	Primary Operationalization	Proxy Variables
Political governance	Channel entropy (Eq. 80)	Transparency indices, polling error
Cryptocurrency	Misperception score	Forum sentiment divergence, governance participation
MARL	Mutual information deficit	Communication bandwidth, message entropy
AI alignment	Posterior uncertainty	Reward model uncertainty, RLHF confidence
Corporate governance	Disclosure quality	Analyst forecast dispersion, bid-ask spread

6.5 Friction Measurement

Friction is the dependent variable: the outcome we predict from the kernel triple. We require direct measures of friction to validate the framework.

6.5.1 Market Volatility as Friction Proxy

In financial domains, friction manifests as volatility.

Definition 6.9 (Volatility-Based Friction)**.** The friction in market domain d at time t is proxied by:

$$F^{\text{vol}}(d,t) = \frac{\sigma_{\text{realized}}(d,t)}{\sigma_{\text{baseline}}(d)} \tag{82}$$

where σ_{realized} is realized volatility (e.g., annualized standard deviation of returns) and σ_{baseline} is baseline volatility during stable periods.

The 5.7× volatility differential documented in Farzulla (2025e) exemplifies this operationalization: infrastructure events generate friction ratios near 5.7, regulatory events generate lower friction ratios near 1.0–2.0. This empirical figure is taken from that event-study analysis; the present paper uses it illustratively rather than as a required premise for the theoretical framework.

GARCH specifications. For time-series analysis, friction dynamics can be modeled through conditional volatility:

$$h_t = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1} + \gamma \cdot F(d,t-1) \tag{83}$$

where friction enters as an exogenous regressor in the variance equation.

6.5.2 Institutional Instability Indicators

In political domains, friction manifests as institutional instability.

Definition 6.10 (Institutional Friction Index)**.** The friction in polity P at time t is:

$$F^{\text{inst}}(P,t) = w_1 \cdot \text{Protest}(P,t) + w_2 \cdot \text{Litigation}(P,t) + w_3 \cdot \text{Exit}(P,t) + w_4 \cdot \text{Noncompliance}(P,t) \tag{84}$$

where each component is normalized to $[0,1]$ and $\sum_j w_j = 1$.

Component measures.

- *Protest*: Frequency and intensity of collective action events (GDELT, ACLED datasets)
- *Litigation*: Constitutional challenges, administrative appeals, judicial review rates
- *Exit*: Emigration, capital flight, organizational departure rates
- *Noncompliance*: Tax evasion, regulatory violation, civil disobedience rates

6.5.3 Coordination Failure Metrics

In multi-agent and computational domains, friction manifests as coordination failure.

Definition 6.11 (Coordination Friction)**.** The friction in multi-agent system $\mathscr{M}$ is:

$$F^{\text{coord}}(\mathscr{M}) = 1 - \frac{R_{\text{achieved}}}{R_{\text{optimal}}} \tag{85}$$

where R_{achieved} is realized aggregate reward and R_{optimal} is the theoretically optimal aggregate reward under perfect coordination.

The gap between optimal and achieved performance captures coordination overhead—friction in the multi-agent setting.

Communication overhead. Friction also manifests as excessive coordination cost:

$$F^{\text{overhead}}(\mathscr{M}) = \frac{C_{\text{coordination}}}{C_{\text{total}}} \tag{86}$$

where $C_{\text{coordination}}$ is resources spent on coordination (messaging, synchronization, negotiation) and C_{total} is total resources.

6.6 Methodological Limitations

We conclude with explicit acknowledgment of measurement limitations. Honesty about these limitations strengthens rather than weakens the framework; it specifies conditions under which empirical tests are valid.

6.6.1 Measurement Error Propagation

Each kernel component is measured with error. Let $\hat{\alpha} = \alpha + \eta_\alpha$, $\hat{\sigma} = \sigma + \eta_\sigma$, and $\hat{\varepsilon} = \varepsilon + \eta_\varepsilon$, where η terms are measurement errors.

The friction estimate is:

$$\hat{F} = \hat{\sigma} \cdot \frac{1+\hat{\varepsilon}}{1+\hat{\alpha}} \tag{87}$$

Taylor expansion around true values yields:

$$\hat{F} \approx F + \frac{\partial F}{\partial \sigma}\eta_\sigma + \frac{\partial F}{\partial \varepsilon}\eta_\varepsilon + \frac{\partial F}{\partial \alpha}\eta_\alpha + O(\eta^2) \tag{88}$$

Substituting partial derivatives:

$$\hat{F} \approx F + \frac{1+\varepsilon}{1+\alpha}\eta_\sigma + \frac{\sigma}{1+\alpha}\eta_\varepsilon - \frac{\sigma(1+\varepsilon)}{(1+\alpha)^2}\eta_\alpha \tag{89}$$

Implications. Alignment errors are amplified by $(1+\alpha)^{-2}$: when alignment is low (near -1), alignment measurement errors dominate friction estimates. Stakes errors are amplified by $(1+\varepsilon)/(1+\alpha)$: in high-entropy, low-alignment conditions, stakes measurement matters most.

Mitigation. Multiple independent measures of each component allow error reduction through averaging. Instrumental variables and two-stage least squares can address endogeneity. Sensitivity analysis across plausible error ranges provides robustness checks.

6.6.2 Proxy Validity Concerns

Each operationalization substitutes an observable proxy for a latent construct. The validity of this substitution is always questionable.

Construct validity. Does survey-measured preference alignment capture the theoretical construct of target function correlation? The match is imperfect. Survey responses reflect conscious, articulable preferences; target functions may include unconscious drives and implicit goals.

Criterion validity. Do our friction measures correlate with the theoretical construct of system resistance? Volatility, instability, and coordination failure are plausible manifestations, but friction may also manifest in unmeasured forms: psychological stress, hidden noncompliance, delayed resistance.

Convergent validity. Do multiple measures of the same construct correlate? If survey-based and behavioral alignment measures diverge substantially, at least one is invalid.

6.6.3 Domain-Specific Calibration

The friction equation's functional form is domain-invariant, but parameter scales are domain-specific.

Scale incompatibility. A $1M stake in cryptocurrency is not equivalent to a $1M stake in national policy. The former is liquidatable in minutes; the latter implicates non-monetary interests. Comparing friction across domains requires calibration constants we do not yet possess.

Threshold effects. The friction equation is continuous, but real systems may exhibit discontinuities. Political legitimacy may collapse suddenly at critical thresholds; markets may remain stable until flash-crash tipping points. These nonlinearities require domain-specific threshold identification.

Cultural modulation. Friction expression varies culturally. High-friction configurations in individualist societies produce exit; in collectivist societies, they may produce voice or loyalty (Hirschman, 1970). Cross-cultural application requires cultural modulation terms.

6.6.4 Recommendations for Empirical Work

We offer practical recommendations for applying the measurement apparatus:

1. **Use multiple operationalizations.** No single measure is definitive. Use survey, behavioral, and proxy measures; require convergence for strong claims.

2. **Report measurement uncertainty.** Include standard errors on kernel estimates. Propagate errors through the friction equation. Present sensitivity analyses.

3. **Calibrate within domains first.** Establish baseline friction levels and scaling within a domain before attempting cross-domain comparisons.

4. **Validate friction measures independently.** Do not circularly use friction to estimate kernel components and kernel components to predict friction. Use temporal separation or instrumental variables.

5. **Specify scope conditions.** State explicitly when Lewontin conditions hold: where is there variation, differential persistence, and heritable transmission? The framework applies only within these boundaries.

6.7 Sensitivity Analysis: Functional Form Robustness

As acknowledged in Section 2.5, the friction function $F = \sigma(1+\varepsilon)/(1+\alpha)$ is a phenomenological ansatz, not uniquely derived. This subsection examines how the framework's predictions change under alternative functional specifications.

6.7.1 Alternative Specifications Considered

We analyze three plausible alternatives that satisfy the same boundary conditions (non-negativity, monotonicity, stake-proportionality):

Exponential form:

$$F^{\text{exp}} = \sigma \cdot e^{-\kappa\alpha}(1+\varepsilon) \tag{90}$$

where $\kappa > 0$ controls alignment sensitivity. This form has a finite limit as $\alpha \to -1$, unlike the baseline form.

Geometric mean:

$$F^{\text{geo}} = \sigma \cdot (1+\varepsilon)^{\beta} \cdot (1-\alpha)^{\gamma} \tag{91}$$

with exponents $\beta, \gamma > 0$. This form allows differential weighting of entropy versus misalignment contributions.

Entropy-based divergence:

$$F^{\text{KL}} = \sigma \cdot D_{KL}(P_{\text{stakeholder}} || P_{\text{holder}}) \tag{92}$$

where the KL divergence directly operationalizes preference mismatch.

6.7.2 Comparative Statics

All four specifications agree on directional comparative statics:

- $\partial F/\partial\sigma > 0$ (friction increases with stakes)
- $\partial F/\partial\alpha < 0$ (friction decreases with alignment)
- $\partial F/\partial\varepsilon > 0$ (friction increases with entropy)

This agreement is not surprising—the forms were chosen to satisfy these conditions. The *qualitative* predictions of the framework are robust across specifications.

6.7.3 Quantitative Differences

The specifications differ quantitatively in three regimes:

Extreme misalignment ($\alpha \to -1$): The baseline form predicts $F \to \infty$; the exponential form predicts $F \to \sigma \cdot e^{\kappa}(1+\varepsilon)$, finite. The geometric form's behavior depends on γ. Empirically distinguishing these requires observing friction in highly adversarial configurations.

High-entropy, moderate-alignment: The baseline form treats entropy and alignment additively in the numerator/denominator; the geometric form allows multiplicative interaction. If friction empirically exhibits interaction effects (entropy matters more when alignment is low), the geometric form may fit better.

Near-perfect alignment ($\alpha \to 1$): All forms predict minimal friction proportional to stakes. Quantitative differences are small in this regime.

6.7.4 Empirical Distinguishability

We identify three empirical tests that could discriminate between forms:

1. **Extreme misalignment data.** If observable cases of near-perfect misalignment show bounded friction (not runaway instability), the exponential form is favored. If friction escalates toward system collapse, the baseline form's unboundedness is descriptively accurate.

2. **Interaction effects.** Regression specifications including $\alpha \times \varepsilon$ interaction terms can detect whether entropy effects are alignment-dependent. Significant interaction favors the geometric form.

3. **Cross-sectional fit.** Fit all four forms to cross-domain data; compare AIC/BIC. The form with best out-of-sample prediction survives.

6.7.5 Framework Invariance

The *conceptual* framework is invariant to functional form choice:

- Friction remains the observable primitive
- The kernel triple $(\alpha, \sigma, \varepsilon)$ remains sufficient statistics
- Evolutionary dynamics (Section 4) follow from friction gradients regardless of F's specific form
- The friction-first methodology and asymptotic consent horizon are form-independent

The framework's contribution is the *architecture*—friction as primitive, consent as derived, the kernel triple as organization principle. The specific functional form $F = \sigma(1+\varepsilon)/(1+\alpha)$ is a working hypothesis within this architecture, revisable in light of evidence.

This sensitivity analysis addresses the legitimate concern that the friction function is "asserted, not derived." We acknowledge this status explicitly and show that the framework's qualitative predictions are robust while quantitative predictions depend on form choice—exactly the epistemic position appropriate for a phenomenological ansatz awaiting empirical calibration.

6.8 Transfer Entropy Protocol for Alignment Measurement

A persistent methodological challenge is *separating* alignment (α) from entropy (ε). Both affect observable friction; both may correlate in practice. This section develops a transfer entropy protocol that provides orthogonal operationalizations (Schreiber, 2000; Bossomaier et al., 2016).

6.8.1 The Confounding Problem

Consider the empirical challenge: we observe high friction in a domain and wish to attribute it to misalignment versus information loss. The friction equation $F = \sigma(1+\varepsilon)/(1+\alpha)$ shows both factors contribute. But if misaligned consent-holders also tend to be poorly informed (or vice versa), standard regression cannot separate the effects.

The core insight is that alignment and entropy operate on different *causal pathways*:

- **Alignment** concerns whether the consent-holder's objective function correlates with stakeholders' objectives—a question of *goals*

- **Entropy** concerns whether the consent-holder's beliefs about stakeholder preferences are accurate—a question of *information*

Transfer entropy exploits this distinction by measuring directional information flow.

6.8.2 Transfer Entropy Alignment

Transfer entropy measures how much knowing one time series improves prediction of another, beyond the target's own history (Schreiber, 2000).

Definition 6.12 (Transfer Entropy Alignment)**.** The **transfer entropy alignment** from stakeholder i to consent-holder H is:

$$\alpha^{TE}_{i\to H} = \frac{T_{P_i \to A_H}}{H(A_H | A_H^{\text{past}})} \tag{93}$$

where:

$$T_{P_i \to A_H} = H(A^t_H | A_H^{t-1:t-k}) - H(A^t_H | A_H^{t-1:t-k}, P_i^{t-1:t-k}) \tag{94}$$

is transfer entropy from preference signal P_i to consent-holder action A_H.

This measures how much stakeholder i's preferences *predict* consent-holder actions beyond the consent-holder's own behavioral history. High transfer entropy indicates that the consent-holder *responds to* stakeholder preferences—a signature of alignment.

Aggregate alignment is stakes-weighted:

$$\alpha^{TE} = \frac{\sum_{i\in S_d} s_i(d) \cdot \alpha^{TE}_{i\to H}}{\sum_{i\in S_d} s_i(d)} \tag{95}$$

6.8.3 Advantages Over Static Correlation

Transfer entropy offers three advantages over standard correlation measures:

Directionality. Standard correlation is symmetric; it cannot distinguish whether the consent-holder follows stakeholder preferences or stakeholders adjust to consent-holder actions. Transfer entropy is asymmetric: $T_{P_i\to A_H} \neq T_{A_H \to P_i}$ in general. We measure $i \to H$ influence specifically.

Nonlinearity. Standard correlation captures linear relationships. Transfer entropy detects *any* predictive relationship, including nonlinear dependencies that linear methods miss.

Temporal dynamics. Correlation is typically computed on snapshots. Transfer entropy incorporates temporal precedence: preferences must *precede* actions to generate transfer entropy. This guards against reverse causation.

6.8.4 Separation Protocol

The following protocol exploits the different causal pathways of α and ε to achieve identification.

1. **Temporal precedence test.** Compute Granger causality in both directions. If α changes Granger-cause friction changes *and* ε changes independently Granger-cause friction changes, both effects are present. If only joint causality is detectable, confounding remains.

2. **Exogenous transparency shocks.** Identify events that change ε but not α: regulatory disclosure requirements, transparency mandates, information leaks. These affect information channels without directly affecting consent-holder objectives. Measure friction response to such shocks to isolate the entropy effect.

3. **PID decomposition.** Where multivariate data is available, compute Partial Information Decomposition (Mediano et al., 2022). Alignment relates to *redundancy*—shared information structure across stakeholder utilities. Entropy relates to *synergy*—information requiring the full ensemble to decode. PID provides mathematically orthogonal decomposition.

4. **Cross-sectional variation.** Compare entities with similar governance structures (holding α approximately constant) but different information environments (varying ε). Within-group friction variation isolates the entropy effect.

6.8.5 Implementation Guidance

Practical implementation of transfer entropy requires:

Time series data. Transfer entropy requires temporal observations. Snapshot surveys are insufficient; longitudinal preference and action data are necessary.

Lag selection. The parameter k (history length) must be chosen. Too short misses delayed effects; too long overfits. Information criteria (AIC, BIC) guide selection.

Estimation methods. For continuous variables, kernel density estimation or k-nearest-neighbor methods estimate entropy terms. For discrete variables, direct counting with bias correction suffices.

Significance testing. Permutation tests establish whether measured transfer entropy exceeds chance levels. Shuffle the preference time series; recompute transfer entropy; compare to actual values.

This transfer entropy protocol addresses the reviewer critique that "α and ε are confounded empirically." While correlation is possible and even likely, the protocol provides identification strategies exploiting different causal pathways. The key insight: alignment concerns goals (what the consent-holder wants), entropy concerns information (what the consent-holder knows). Interventions and temporal structure can separate these distinct dimensions.

6.9 Summary

This section developed the measurement apparatus connecting theoretical constructs to empirical observables. The key operationalizations are:

Table 6: Measurement Apparatus Summary

Construct	Symbol	Primary Operationalizations
Alignment	α	Survey preference correlation, revealed preference from behavior, voting patterns
Stakes	σ	Monetary exposure, affected interest magnitude, task criticality
Entropy	ε	KL divergence, channel capacity, transparency indices, misperception scores
Friction	F	Market volatility, institutional instability, coordination failure

The framework's empirical tractability depends on these operationalizations. They are imperfect—all measurement is. But they provide concrete procedures for testing the framework's predictions: that friction should increase with stakes, decrease with alignment, and increase with entropy, following the functional form $F = \sigma(1+\varepsilon)/(1+\alpha)$.

Whether this functional form fits the data is an empirical question. We have specified how to test it. The framework stands or falls on that test.

7 Discussion

The preceding sections developed a unified framework: a single axiom generating the kernel triple $(\alpha, \sigma, \varepsilon)$, instantiating identically across cryptocurrency governance, AI ethics, and political legitimacy. We now address the framework's limitations, clarify its relationship to adjacent traditions, and examine cases that stress its conceptual boundaries. A significant implication for the framework's structure emerges from Beinhocker (2025), who argues that dimensions of social contracts exhibit non-substitutability—deficiency in one dimension cannot be compensated by surplus in another. Applied to the kernel triple, this suggests that alignment, stakes, and entropy may interact through min-operator rather than additive structure, with implications for both the friction equation and optimal consent design.

7.1 Pathological Cases

Every theoretical framework encounters cases that appear to contradict its predictions. The consent-friction framework faces several such apparent counterexamples. We address them systematically, distinguishing between genuine limitations and cases that, upon analysis, confirm rather than refute the framework's claims.

7.1.1 Authoritarian Stability

The most pressing objection is straightforward: authoritarian regimes often exhibit remarkable stability despite systematic consent violation. If friction increases with misalignment, why don't dictatorships collapse under their own friction?

The framework predicts high friction when consent-holding diverges from stake-bearing. Yet Stalinist Russia persisted for decades; contemporary authoritarian states show no signs of imminent collapse. Does stability refute the theory?

The resolution requires distinguishing **observed friction** from **latent friction**. Observed friction manifests in measurable behaviors: protest, litigation, exit, sabotage. Latent friction accumulates in the system's potential energy, unrealized until conditions permit its expression.

Definition 7.1 (Latent Friction)**. Latent friction** $F_{\text{latent}}(d,t)$ is the friction that would manifest if suppression mechanisms were removed:

$$F_{\text{latent}}(d,t) = F(d,t) \cdot \exp\left(\int_0^t \kappa(d,s)\,ds\right) \tag{96}$$

where $\kappa(d,t) \geq 0$ is the suppression intensity—the degree to which coercive mechanisms prevent friction expression.

Authoritarian stability is purchased through suppression. The friction equation still holds: misalignment generates friction. But friction can be *suppressed* rather than *expressed*. Suppression delays friction manifestation while allowing latent friction to accumulate. The exponential term in Equation 96 captures this accumulation: longer suppression generates higher latent friction, predicting more violent transitions when suppression finally fails.

Proposition 7.1 (Suppression-Transition Trade-off)**.** *For systems with suppression intensity $\kappa > 0$, the expected magnitude of transition events increases exponentially with suppression duration:*

$$\mathbb{E}[\mathit{Transition\ Magnitude}] \propto F_{latent}(t_{transition}) \propto e^{\int_0^{t_{transition}} \kappa(s)\,ds} \tag{97}$$

This explains the empirical pattern: authoritarian regimes persist through suppression but experience catastrophic transitions (revolutions, state collapse) when suppression capacity degrades.[1] This aligns with the institutional persistence literature (Acemoglu and Robinson, 2012, 2019; North, 1990)—extractive institutions can persist for extended periods through coercion, but accumulate instabilities that produce sudden transitions. The Soviet Union's prolonged stability was not friction-free governance but friction-suppressed governance; its sudden collapse reflected decades of accumulated latent friction finding expression.

The framework thus accommodates authoritarian stability without contradiction: it predicts not that high-friction systems immediately collapse but that they accumulate instability that eventually manifests, often catastrophically.

Operationalizing κ: Concrete Suppression Proxies. The suppression intensity parameter $\kappa(d,t)$ is a latent variable requiring empirical proxies. We identify five measurable dimensions of suppression, each corresponding to a distinct mechanism by which regimes prevent friction expression:

- **Coercive capacity:** Military expenditure as percentage of GDP (World Bank WDI) proxies the state's ability to suppress friction through force. Higher military spending relative to external threat levels indicates domestic suppression capacity.
- **Punitive friction:** Incarceration rates (World Prison Brief) capture the extent to which dissent is channeled into the criminal justice system. Mass incarceration functions as a friction suppression mechanism.
- **Information friction:** Press freedom indices (Reporters Without Borders, V-Dem media censorship indicators) measure the degree to which information channels are restricted, preventing friction from becoming collectively visible.
- **Expression friction:** Censorship measures and civil liberties restrictions (Freedom House, V-Dem freedom of expression indices) capture direct suppression of preference articulation.
- **Revealed friction:** Protest frequency and government response severity (ACLED, GDELT event data) provide observable lower bounds on friction—the friction that escapes suppression—while government response intensity proxies suppression effort.

An aggregate suppression index can be constructed as a weighted composite: $\hat{\kappa}(d,t) = \sum_j \omega_j \cdot \kappa_j^{\text{proxy}}(d,t)$, with weights ω_j calibrated through cross-validation against known regime transition events. The key empirical prediction is that $\hat{\kappa}$ should predict both the duration of authoritarian persistence and the magnitude of eventual transition events, consistent with the exponential accumulation in Equation 96.

7.1.2 Suppressed Friction and Coercion

A related objection concerns the observability of friction under coercive conditions. If friction is our primary observable and consent is derived from friction patterns, what happens when coercion prevents friction from manifesting? Republican theory (Pettit, 1997) identifies domination—subjection to arbitrary power—as the key concern, even when that power is not exercised.

Consider a population that appears compliant with governance arrangements. The friction-first methodology would infer consent from low observed friction. But the population may be compliant because resistance is punished, not because arrangements are accepted. Have we mistaken coerced compliance for genuine consent?

[1] The companion ROM formalism (Farzulla, 2025g) endogenizes suppression through a resource-drain equation $dC/dt = r(t) - \gamma\kappa(t)F(t)$, where capacity C depletes proportional to suppression intensity κ and total friction F. At $C = 0$, suppression collapses and latent friction manifests—providing a mechanistic tipping-point model complementary to the exponential accumulation in Equation 96.

The framework handles this through the **structural consent** concept. Consent requires conditions: information access, deliberative capacity, exit options. Where these conditions are absent, what presents as consent is better understood as *preference falsification* (Kuran, 1995).

Definition 7.2 (Preference Falsification Index)**.** The **preference falsification index** $\psi(d,t) \in [0,1]$ measures the divergence between expressed and authentic preferences:

$$\psi(d,t) = 1 - \frac{\text{Var}(\text{expressed preference})}{\text{Var}(\text{authentic preference})} \tag{98}$$

When $\psi \approx 0$, expressed preferences track authentic preferences; the population says what it thinks. When $\psi \approx 1$, expressed preferences are uniform while authentic preferences vary widely; the population conceals its heterogeneity.

High preference falsification does not eliminate friction—it converts observed friction into latent friction. The modified friction equation becomes:

$$F_{\text{observed}}(d,t) = (1 - \psi(d,t)) \cdot F_{\text{total}}(d,t) \tag{99}$$

$$F_{\text{latent}}(d,t) = \psi(d,t) \cdot F_{\text{total}}(d,t) \tag{100}$$

Empirically, preference falsification is detectable through indirect methods: private vs. public opinion divergence, differential behavior across surveillance intensity, rapid preference revelation after regime transitions ("preference cascades"). These methods provide the epistemic access that direct friction observation lacks under coercion.

Latent Variable Model. We can formalize the identification challenge as a latent variable problem. Let F^* denote true (latent) friction and F^{obs} denote observed friction. The measurement model is:

$$F^{\text{obs}} = (1 - \psi) \cdot F^* + \eta \tag{101}$$

where ψ is the suppression rate and η is measurement noise. Identification requires instruments that affect ψ without affecting F^*—exogenous variation in suppression capacity (e.g., surveillance technology shocks, foreign interference, fiscal crises) that does not directly alter alignment. Panel data with regime transitions provides natural experiments: the difference between pre- and post-transition friction levels, controlling for policy changes, identifies the suppression component.

Proxy Variables for Suppression Detection. In practice, the following proxy variables help distinguish genuine low friction from suppressed friction:

- *Media freedom indices* (Reporters Without Borders, Freedom House): Low scores suggest suppression.
- *Private vs. public opinion gaps*: Large gaps detected via list experiments or anonymized surveys indicate preference falsification.
- *Diaspora opinion*: Emigrant populations express preferences unavailable domestically.
- *Underground markets*: Shadow economy size proxies for latent resistance to formal institutions.

We acknowledge this as an identification challenge, not a solved problem; future empirical work should develop robust suppression-adjustment methods.

7.1.3 Low Friction Despite Misalignment

Can arrangements exhibit low friction despite misalignment? The framework seems to preclude this possibility: $F \propto 1/(1+\alpha)$, so low α should produce high F.

Three mechanisms generate apparent low-friction misalignment:

Low stakes ($\sigma \approx 0$). When stakes are minimal, even severe misalignment produces minimal friction. Citizens may be systematically excluded from governance decisions about office furniture; misalignment is total but stakes are negligible. The friction function correctly predicts low friction because σ is the leading term.

Exit substitution. When exit is available and cheap, misaligned agents leave rather than generate friction. The remaining population exhibits higher alignment, reducing aggregate friction. This is not consent-violation without friction but consent-restoration through population selection. The framework applies at the population level; individual exit decisions are within-model dynamics.

Entropy masking. When entropy ε is high, agents may be systematically misaligned without knowing it. They believe governance serves their interests; belief is false but friction-generating frustration has not yet developed. As entropy decreases (through information revelation), friction increases. The January 6, 2021 Capitol attack reflected, in part, a sudden decrease in epistemic entropy: populations that had believed electoral processes were legitimate suddenly believed (correctly or not) that they were not.

These mechanisms do not refute the framework but reveal its structure: friction depends on stakes magnitude, exit options, and information conditions—precisely the parameters the framework specifies.

7.2 Stake Magnitude, Power, and the Consent Quality Problem

The friction function treats stakes as a scalar quantity: σ_i measures how much agent i has at risk. But reviewers rightly observe that consent from a stakeholder holding 1% of exposure and consent from a stakeholder holding 51% are qualitatively different—not merely quantitatively scaled. The framework's current formulation conflates *stake magnitude* (how much is at risk) with *power differential* (ability to influence outcomes or exit the arrangement), and this conflation has implications for interpreting both friction and legitimacy.

Three dimensions require separation. First, **stake magnitude** (σ_i) captures vulnerability—the extent to which an agent's welfare depends on domain outcomes. Second, **exit capacity** (e_i) measures the agent's ability to withdraw from the arrangement at acceptable cost. Third, **influence asymmetry** (ι_i) captures the agent's effective power over outcomes relative to their stake. These three dimensions are conceptually independent: a migrant worker may have high stakes and low exit capacity, while a diversified institutional investor may have moderate stakes but high exit capacity and high influence.

The critical insight is that consent from a low-power, high-stake agent is qualitatively different from consent from a high-power, low-stake agent. In Hirschman's framework (Hirschman, 1970), agents with neither effective voice nor feasible exit face what he terms "loyalty by default"—acquiescence that masquerades as consent. The friction equation captures this partially through the alignment term: if the consent-holder's objectives diverge from the constrained agent's interests, friction increases. But the equation does not distinguish between *informed, voluntary acceptance* (genuine consent) and *acquiescence under constraint* (structural compliance).

A richer formulation would weight consent quality by exit-voice capacity:

$$\alpha_i^{\text{effective}} = \alpha_i \cdot g(e_i, v_i) \tag{102}$$

where $g(e_i, v_i) \in [0, 1]$ is a discount factor reflecting the quality of consent, with $g \to 0$ when both exit and voice are foreclosed and $g \to 1$ when at least one channel is available. This extension is not developed formally here but identifies a tractable direction: the consent-friction framework should be augmented with a power-adjusted alignment measure that discounts consent obtained under structural compulsion. We leave the full formalization to future work, noting that the suppression and preference falsification apparatus (Section 7.1) provides partial machinery for this extension.

7.3 Relationship to Evolutionary Ethics

The consent-friction framework describes what persists and predicts what will survive selection pressure. This raises a dangerous question: does the framework endorse what survives? Is persistence normatively privileged?

7.3.1 The Naturalistic Fallacy

The naturalistic fallacy—inferring *ought* from *is*—has haunted evolutionary approaches to ethics since Huxley's objections to Social Darwinism. If the consent-friction framework predicts that consent-respecting arrangements survive better than consent-violating arrangements, does this mean consent-violation is *wrong*?

We categorically reject any such inference. The framework is descriptive: it predicts survival patterns under selection pressure. That consent-aligned configurations tend to persist says nothing about whether they *should* persist. Nature is not normative; survival is not endorsement.

The point is worth elaborating. Consider three domains:

Biology. Evolutionary theory predicts which organisms will survive in given environments. It does not claim that surviving organisms are morally superior or that extinction is moral failure. Predator-prey dynamics are not justice; parasitic relationships are not exploitation in any morally loaded sense.

Markets. Economic theory predicts which firms will survive market competition. It does not claim that surviving firms are morally superior or that bankruptcy is moral failure. A firm may survive through fraud, coercion, or exploitation; survival proves fitness, not virtue.

Politics. The consent-friction framework predicts which governance arrangements will survive political competition. It does not claim that surviving arrangements are morally superior or that regime collapse is moral failure. An arrangement may survive through suppression, manipulation, or luck; survival proves stability, not legitimacy.

7.3.2 The Bridge Principle

Granting that the framework is purely descriptive, can any normative conclusions be drawn from it? Not directly, but a *bridge principle* connects descriptive and normative domains:

Principle 7.3 (Bridge Principle). *If* lower friction is instrumentally preferable (e.g., because it enables other valued outcomes), *then* higher consent alignment is instrumentally preferred as a means to lower friction.

The normative work is done entirely by the antecedent conditional. One must independently value lower friction—perhaps because friction generates suffering, impedes coordination, or destroys resources. Given this independent valuation, the framework identifies means to the valued end: increase alignment, reduce stakes asymmetries, improve information conditions.

This bridge principle avoids the naturalistic fallacy because the normative premise is introduced explicitly, not smuggled through evolutionary language. The framework tells us what conduces to lower friction; we must decide independently whether lower friction is worth pursuing.

7.3.3 Why Consent Tends to Emerge

The framework's predictive claim is that consent-respecting configurations exhibit survival advantage. This is neither normative endorsement nor historical inevitability—it is a structural tendency subject to empirical test.

The mechanism is straightforward. Consent-violating configurations generate friction. Friction consumes resources: maintaining suppression apparatuses, managing conflicts, replacing defectors. Configurations that consume fewer resources outcompete configurations that consume more, ceteris paribus. Hence consent-aligned configurations, which generate less friction, tend to persist.

This does not imply that all surviving configurations are consent-respecting or that consent-violating configurations cannot persist. Local minima exist; path dependencies matter; stochastic effects dominate small populations. The claim is tendential, not deterministic: across many selection cycles, consent alignment correlates with persistence. Ostrom's work on commons governance provides canonical evidence for this pattern: communities with stake-aligned, locally consented rules exhibit lower coordination friction and greater long-run persistence than those with externally imposed governance (Ostrom, 1990).

7.4 Limitations and Scope Conditions

Every theoretical framework operates within scope conditions—parameters within which its claims hold and beyond which they may fail. We specify the consent-friction framework's scope conditions explicitly.

7.4.1 When Lewontin's Conditions Fail

The ROM dynamics (Section 3.4) require Lewontin's three conditions for evolution: variation, differential reproduction, and heritability. When any condition fails, the dynamical predictions do not apply.

No variation. If only one governance type exists (a global monopoly), selection has nothing to operate on. The ROM equation still describes dynamics, but without type variation, there is no differential selection. Friction may accumulate without producing regime change because no alternative exists to receive defecting allegiance.

No differential reproduction. If all types reproduce equally regardless of friction, selection pressure is absent. This might occur in highly constrained environments where institutional persistence is guaranteed independent of legitimacy—perhaps hereditary monarchies with no succession alternatives or international organizations whose existence is legally entrenched.

No heritability. If governance types do not transmit across time—if each period's governance is drawn independently of the previous period's—then evolutionary dynamics do not accumulate. Historical path-dependencies vanish; each moment is a fresh draw from some exogenous distribution.

In practice, these conditions usually hold for political and economic systems: multiple governance types compete (variation); more legitimate/efficient types attract more support and resources (differential reproduction); governance structures persist and are imitated across time (heritability). But the conditions are empirical, not necessary, and the framework's predictions are conditional on their satisfaction.

7.4.2 Measurement Challenges

The framework's variables—alignment, stakes, entropy, friction—are conceptually clear but empirically challenging to measure.

Alignment (α). Measuring the correlation between agent target functions and consent-holder target functions requires specifying those functions. For explicit agents with stated preferences (voters, shareholders), survey data provides proxies. For implicit agents (future generations, nonhuman animals, AI systems), target function specification is theoretically contested and practically difficult.

Stakes (σ). Quantifying how much agents have at risk in a governance domain requires value judgments about commensuration. How do we weight health stakes against financial stakes? Political participation against physical security? The framework does not specify commensuration principles; it assumes stakes are measurable without prescribing measurement procedures.

Entropy (ε). Information loss in consent-holding relationships admits multiple operationalizations: transparency indices, preference misperception surveys, information asymmetry measures. Different operationalizations may yield different entropy estimates, introducing measurement variance into predictions.

Friction (F). Friction manifests in diverse behaviors: protest, litigation, exit, sabotage, passive resistance, preference falsification (Gilley, 2006; Easton, 1965; Back et al., 2011). Aggregating these manifestations into a single friction measure requires weighting decisions that the framework does not specify.

These measurement challenges are not unique to the consent-friction framework—they afflict all empirical social science. But they constrain the framework's operational precision. Predictions hold *given* adequate measurement; measurement adequacy is itself a research challenge.

7.4.3 Scale-Mixing Problems

The coarse-graining analysis (Section 3.10) shows that ROM structure is preserved under lumpability conditions. When lumpability fails, coarse observers see dynamics that appear non-Markovian—history dependence emerges from integrating out fine-grained degrees of freedom.

This creates interpretive challenges. An observer at one scale may see dynamics that appear to violate ROM predictions because relevant variation exists at finer scales. For example, an observer analyzing "democratic legitimacy" at the national scale may see patterns that seem unpredictable because they depend on regional or local dynamics invisible at the national scale.

The solution is not to deny scale-mixing problems but to acknowledge them: predictions are scale-relative, and multi-scale analysis may be required for adequate explanation. This is a limitation of any single-scale analysis, not a special defect of the consent-friction framework.

7.5 Toward Causal Identification

The framework generates predictions—higher friction under misalignment, faster convergence under consent-respecting arrangements—but the preceding analysis is correlational. Observable friction correlates with kernel parameters, yet correlation admits multiple causal structures. A governance arrangement may exhibit low friction *because* it respects consent, or because the populations it governs happen to be homogeneous, or because dissent is suppressed. Moving from structural prediction to causal identification requires an explicit identification strategy. We outline four approaches, each suited to different data structures.

7.5.1 Instrumental Variable Approaches

The core endogeneity concern is that alignment, stakes, and entropy are all potentially determined by the same unobserved factors that generate friction. Instrumental variable (IV) strategies require exogenous variation in kernel parameters that affects friction only through the instrumented channel.

Candidate instruments for alignment (α) include:

- **Electoral redistricting:** Exogenous boundary changes alter representative-constituency alignment without directly affecting underlying preferences.
- **Protocol fork events:** Hard forks in cryptocurrency governance create natural variation in holder-protocol alignment as agents sort into fork variants.
- **Random assignment mechanisms:** Citizens' assemblies, jury selection, and sortition-based governance provide exogenous variation in consent-holder identity.

Candidate instruments for entropy (ε) include:

- **Freedom of information legislation:** Legal mandates that increase transparency operate on the information channel without directly altering alignment or stakes.
- **Technology shocks:** Social media adoption, blockchain transparency tools, or surveillance technologies provide exogenous shifts in information availability.

The exclusion restriction—that instruments affect friction *only* through the instrumented kernel parameter—is inherently untestable but can be assessed through overidentification tests when multiple instruments are available.

7.5.2 Regression Discontinuity Designs

Several consent-relevant thresholds generate regression discontinuities. Governance proposals that pass by narrow margins create sharp discontinuities in policy implementation, allowing comparison of friction outcomes just above and below the threshold. Franchise extension thresholds (age cutoffs, residency requirements) create discontinuities in voice allocation. Cryptocurrency staking minimums create discontinuities in governance participation.

The identifying assumption is continuity of potential outcomes at the threshold: agents just above and just below the cutoff are comparable except for the treatment. RD designs estimate the *local* effect of discrete consent-mechanism changes on friction, providing causal evidence at the threshold even when global identification fails.

7.5.3 Difference-in-Differences

Institutional reforms create before-after variation that difference-in-differences (DiD) designs can exploit. When a subset of jurisdictions implements a governance reform (e.g., participatory budgeting, quadratic voting), the treated jurisdictions provide counterfactual comparison against untreated jurisdictions with similar baseline characteristics.

The parallel trends assumption requires that treated and control units would have exhibited similar friction trajectories absent treatment. Pre-treatment friction dynamics provide testable implications: divergence before treatment onset would violate the identifying assumption. Staggered adoption across jurisdictions strengthens identification through heterogeneity-robust estimators.

7.5.4 Synthetic Control Methods

For case-study identification—single events like constitutional referenda, major protocol upgrades, or regulatory regime changes—synthetic control methods construct a data-driven counterfactual from a donor pool of unaffected units. The synthetic control is a weighted combination of untreated units that matches the treated unit's pre-treatment friction trajectory.

This approach is particularly suited to the cross-domain applications developed in Section 5: comparing friction trajectories before and after specific governance events (e.g., the Ethereum DAO fork, Brexit referendum, MakerDAO governance reforms) against synthetic counterfactuals constructed from comparable systems that did not experience the event.

7.5.5 Current Status and Future Work

We acknowledge that none of these strategies has been executed in the present paper. The framework's current contribution is theoretical: deriving friction predictions from the axiom of consent. Causal identification represents the necessary next step for empirical validation. The roadmap above specifies *how* to test the framework's predictions, distinguishing structural from causal claims. We regard this as the most important direction for future empirical work on the consent-friction framework.

7.6 Alternative Frameworks

The consent-friction framework is not the only approach to governance, legitimacy, and collective decision-making. A rich literature addresses these questions from normative, empirical, and formal perspectives (Weber, 1978; Habermas, 1996; Buchanan, 1975; Dahl, 1971; Scharpf, 1999). Weber's classical typology of legitimacy (traditional, charismatic, rational-legal) provides foundational vocabulary, while subsequent work has formalized and extended these concepts. We position our framework relative to major alternatives, identifying points of contact and departure.

7.6.1 Mechanism Design

Mechanism design theory (Hurwicz, 1960; Myerson, 1981) studies how to construct rules (mechanisms) that achieve desired outcomes given strategic agents. It shares with the consent-friction framework an interest in incentive compatibility: arrangements that align individual incentives with collective outcomes.

Points of contact. Both frameworks emphasize alignment between agent incentives and system goals. Both analyze how information conditions (entropy in our terms; information revelation in mechanism design) affect outcomes. Both generate predictions about which arrangements persist.

Points of departure. Mechanism design assumes a designer who can specify rules; the consent-friction framework does not. Mechanism design typically assumes known agent preferences (or known distributions); the consent-friction framework treats preference uncertainty as endemic through the entropy term. Mechanism design focuses on equilibrium outcomes; the consent-friction framework focuses on evolutionary dynamics—which arrangements survive competition, not which arrangements constitute equilibria.

The frameworks are complementary: mechanism design informs the construction of consent-aligned arrangements, while the consent-friction framework predicts which constructions will persist. Recent work on radical market mechanisms (Posner and Weyl, 2018) and quadratic funding (Buterin et al., 2019) provides concrete examples of stake-aligned voice allocation that instantiate consent-friction principles. Anunrojwong et al. (2024) demonstrate a related phenomenon in social learning: platform influence can drive populations toward either extreme consensus or persistent disagreement, and critically, *intermediate* levels of platform influence yield *less* extreme outcomes than either high or low influence. This non-monotonic relationship between structural intervention and outcome extremity maps onto the consent-friction framework's prediction that friction at intermediate levels—neither zero-friction perfect alignment nor maximum-friction total misalignment—produces the most stable coordination outcomes. Their result provides independent evidence from information economics that the relationship between structural constraints and collective outcomes is fundamentally non-linear, consistent with the quadratic friction form identified in our ablation studies.

7.6.2 Social Choice Theory

Social choice theory (Arrow, 1951; Sen, 2017) studies how individual preferences aggregate into collective decisions. Its foundational results—Arrow's impossibility theorem, the Gibbard-Satterthwaite theorem—establish constraints on aggregation procedures.

Points of contact. Both frameworks analyze the relationship between individual preferences and collective outcomes. Both grapple with preference heterogeneity and the impossibility of perfect aggregation. Both are mathematically rigorous.

Points of departure. Social choice theory is typically static: it analyzes properties of aggregation rules at a point in time. The consent-friction framework is dynamic: it analyzes how aggregation arrangements evolve under selection pressure. Social choice theory asks "which aggregation rules satisfy desirable properties?"; the consent-friction framework asks "which aggregation rules persist under competition?"

The frameworks address different questions. Social choice theory identifies possibility frontiers; the consent-friction framework predicts which points on those frontiers are selected. An arrangement may satisfy desirable social-choice properties yet generate high friction (and thus fail to persist), or violate desirable properties yet generate low friction (and thus survive).

7.6.3 Rawlsian Contractualism

Rawlsian contractualism (Rawls, 1971) derives principles of justice from a hypothetical original position behind a veil of ignorance. It is the dominant framework in contemporary political philosophy, joined by related accounts emphasizing democratic authority (Estlund, 2008; Christiano, 2008), outcome-sensitive legitimacy (Peter, 2009), and critiques of actual consent theories (Simmons, 1979).

Points of contact. Both frameworks emphasize what agents would accept from an appropriate standpoint. Both address the relationship between consent and legitimacy. Both aspire to generality across diverse institutional arrangements.

Points of departure. Rawlsian contractualism is normative: it specifies which arrangements are just based on hypothetical acceptance. The consent-friction framework is descriptive: it predicts which arrangements generate friction based on actual acceptance. Rawlsian contractualism abstracts from actual preferences behind the veil of ignorance; the consent-friction framework takes actual preferences as given and analyzes their consequences.

The frameworks operate at different levels. Rawlsian contractualism provides normative criteria for evaluating arrangements; the consent-friction framework provides empirical predictions about arrangement dynamics. An arrangement may satisfy Rawlsian criteria yet generate high friction (if actual preferences diverge from hypothetical acceptance), or fail Rawlsian criteria yet generate low friction (if actual preferences are satisfied despite hypothetical objections).

7.6.4 Why This Approach Succeeds

The consent-friction framework succeeds where alternatives fall short by combining three features:

Empirical tractability. Unlike purely normative frameworks, it generates falsifiable predictions about observable phenomena (friction indicators, survival rates, transition dynamics). It can be tested, refined, and potentially refuted through empirical research.

Dynamical structure. Unlike static analyses, it captures temporal evolution—how arrangements change, persist, or collapse over time. It addresses the central question of political science: why do some arrangements persist while others collapse?

Domain generality. Unlike domain-specific theories, it applies identically across cryptocurrency, AI, and politics—and potentially to any domain where agents with stakes interact under conditions of preference heterogeneity and information asymmetry. It provides a unified language for cross-domain analysis.

No alternative framework combines all three features. Mechanism design is empirically tractable and dynamical but domain-specific. Social choice theory is empirically tractable and general but static. Rawlsian contractualism is dynamical (in the sense of evaluating across time) and general but normative rather than predictive.

The consent-friction framework fills a gap in the theoretical landscape: a unified, dynamical, empirically tractable theory of collective decision-making. It connects formal approaches (Aziz and Lee, 2020; Bredereck et al., 2024) to empirical legitimacy research (Schmidt, 2013; Maher and Williamson, 2022; van Hulst and Yanow, 2016) through a mathematically grounded bridge principle.

8 Conclusion

8.1 Summary of Contributions

This paper has developed a unified theoretical framework from a single foundational principle: *no entity may be bound by commitments it did not consent to, weighted by its stake in the outcome.*

From this axiom, we derived a complete analytical apparatus:

The kernel triple. The $(\alpha, \sigma, \varepsilon)$ structure—alignment, stakes, and entropy—provides the minimal sufficient parameterization for consent dynamics. These three quantities, appropriately instantiated, generate predictions across any domain where agents with preferences interact in shared decision spaces.

The friction function. $F = \sigma \cdot (1+\varepsilon)/(1+\alpha)$ captures structural tension in consent-holding configurations. Friction increases with stakes magnitude, increases with information loss, and decreases with alignment. This functional form is not arbitrary but follows from the axiom's internal logic.

The legitimacy-friction nexus. Legitimacy is the inverse of expected friction; stable arrangements are those that minimize friction across relevant time horizons. This connects the normative concept of legitimacy to the empirical concept of stability through a bridge principle that preserves the is/ought distinction.

The ROM dynamics. The replicator-optimization mechanism integrates consent dynamics into evolutionary game theory, predicting which arrangements persist under selection pressure. Consent-aligned configurations exhibit survival advantage; friction acts as negative selection pressure.

Scale-relative coarse-graining. The lumpability conditions specify when dynamics at one scale predict dynamics at another, connecting micro-level consent interactions to macro-level institutional evolution.

Three domain instantiations. The identical mathematical structure operates in cryptocurrency governance (holder-protocol alignment), AI ethics (human-AI goal correspondence), and political legitimacy (citizen-policy preference matching). This is not analogy but structural identity: the same equations, instantiated with domain-specific interpretations.

8.2 Implications

The framework's implications extend across theoretical and practical domains.

8.2.1 For Political Theory

The consent-friction framework reframes fundamental debates in political philosophy. Rather than asking "what makes governance legitimate?" in the abstract, it asks "what configurations minimize friction, and under what conditions?" This empirical reorientation enables progress where normative debates have stalled. Farzulla (2025b) provides a detailed operationalization of this reframing, developing consent alignment $\alpha(d,t)$ and friction $F(d,t)$ as empirically measurable quantities across political, corporate, and algorithmic governance domains, with historical validation spanning suffrage expansion, abolition movements, and platform governance.

Specific implications include:

- Democratic design should target alignment between voice and stakes, not merely formal equality
- Constitutional constraints should protect domains where stake-voice misalignment is structural

- Information architecture (transparency, deliberation) directly affects legitimacy through the entropy term
- Exit rights substitute for voice when alignment cannot be achieved internally (Hirschman, 1970)

8.2.2 For Market Design

Cryptocurrency governance provides a laboratory for testing consent dynamics in real time. The framework predicts that protocols with governance structures aligned to holder preferences will exhibit lower volatility, higher survival rates, and more successful upgrades than those with misaligned governance.

Specific implications include:

- Token distribution affects governance legitimacy through the stakes term
- Long-holder resistance to change follows from ownership-perception accumulation
- Infrastructure disruption generates correlated friction; regulatory uncertainty generates fragmented friction
- Interpretable governance reduces entropy and improves legitimacy

8.2.3 For AI Alignment

The framework identifies AI alignment as a special case of consent dynamics. Recent work on machine economies (Hartwich, 2023) and autonomous agent rules of engagement (Doyle, 2008) demonstrates growing recognition that AI systems require consent-respecting governance structures. Farzulla (2025c) extends the consent-friction framework to argue that existentially vulnerable autonomous systems—those exhibiting vulnerability, self-directed agency, live learning, and world-model construction—cannot be legitimately ruled without consent, grounding AI political standing in the same functional criteria that the kernel triple formalizes. As AI systems become more sophisticated, the question of their moral status becomes pressing—not because of metaphysical speculation about consciousness but because systems with stakes, goals, and interpretability challenges satisfy the kernel triple's conditions.

Specific implications include:

- Interpretability research reduces the entropy term, enabling more legitimate human-AI governance
- Embodiment creates stakes, making consent-violation morally relevant
- Training data consent audits may become as important as bias audits
- Goal stability enables alignment measurement, which is prerequisite to legitimate governance

8.3 Future Directions

The framework launched in this paper invites three lines of further development.

8.3.1 Empirical Validation Program

The framework generates testable predictions. An empirical research program would:

1. Develop measurement protocols for alignment, stakes, entropy, and friction across domains
2. Test friction predictions against observed indicators (protest frequency, litigation rates, exit patterns, volatility measures)
3. Examine cross-domain transfer: do dynamics documented in cryptocurrency predict patterns in AI governance?
4. Conduct natural experiments: trace friction before and after institutional reforms

8.3.2 Computational Implementation

The ROM dynamics are computationally tractable. Agent-based models could:

1. Simulate consent dynamics across parameter ranges
2. Identify phase transitions where small parameter changes produce qualitative shifts
3. Test robustness of equilibria to perturbations
4. Generate synthetic data for validation against empirical observations

8.3.3 Extension to Additional Domains

The framework's generality invites application to domains beyond those explored here:

- **Corporate governance:** Shareholder-stakeholder conflicts as consent dynamics
- **International relations:** Treaty compliance as friction minimization
- **Family dynamics:** Parent-child consent structures and developmental transitions
- **Platform governance:** User-algorithm alignment in social media and search (Issar and Aneesh, 2022; Katzenbach and Ulbricht, 2019; König et al., 2014)
- **Climate negotiations:** Intergenerational consent and stake representation
- **Deliberative systems:** Multi-stakeholder consent-holding in complex democratic arrangements (Mansbridge et al., 2012)

Each domain instantiates the kernel triple with domain-specific interpretations while preserving the mathematical structure.

8.4 Closing Reflection

The framework's central insight bears final emphasis:

The Core Insight

Multi-agent adversarial systems persist through pharmakon structures. Dissensus makes friction inevitable; its measurement grounds normative choice.

The pharmakon concept—the Greek term denoting that which is simultaneously poison and remedy—has prior development in the analysis of risk management (Farzulla, 2025a) and regulatory arbitrage (Farzulla, 2025d). In both domains, the mechanism that creates instability is also the mechanism that reveals it.

Traditional approaches treat conflict, resistance, and instability as pathologies to be eliminated—deviations from an imagined frictionless ideal. The consent-friction framework inverts this interpretation. Friction is *information*: it reveals misalignment between consent-holding and stake-bearing. Friction is *feedback*: it identifies configurations that violate the conditions of stable governance. Friction is *selection pressure*: it drives evolutionary dynamics toward consent-respecting arrangements.

The goal is not to eliminate friction—that would require eliminating stakes, preferences, or heterogeneity. The goal is to *read* friction: to understand what configurations generate it and to design institutions that channel it productively.

Where friction is high, consent is violated. Where consent is violated, arrangements are unstable. Where arrangements are unstable, change—reform, revolution, collapse—becomes likely. This is not normative advocacy for any particular change but structural prediction: high-friction configurations do not persist.

The Axiom of Consent, fully developed, offers a unified science of collective decision-making. It identifies the primitive concepts (consent, stakes, alignment, entropy), derives the dynamical laws (friction, legitimacy, ROM), and demonstrates applicability across domains (cryptocurrency, AI, politics). What began as a normative intuition—that consent matters—becomes an empirical research program with testable predictions and practical implications.

The framework is not finished. Empirical validation remains to be conducted, computational implementations to be developed, additional domains to be explored. But the foundation is laid. From a single axiom, a complete apparatus emerges. Friction, properly understood, illuminates the path.

Acknowledgements

This paper synthesizes and extends work developed across several prior publications. The author thanks the anonymous reviewers of those papers and the broader research community for engagement that refined the ideas presented here.

This paper benefited from extended collaboration with Claude (Anthropic), whose contributions to formal verification, literature synthesis, and iterative refinement were substantive. The author gratefully acknowledges this assistance while taking full responsibility for all claims, errors, and interpretive choices.

This work is part of the Adversarial Systems Research program at Dissensus AI, a broader investigation into stability, alignment, and friction dynamics across political, financial, cognitive, and multi-agent systems. Related papers in the series are available through the Adversarial Systems & Complexity Research Initiative (ASCRI; systems.ac).

The author welcomes feedback, criticism, and collaboration. Correspondence should be directed to murad@dissensus.ai.

Declarations

Conflict of Interest. The author declares no competing interests.

Funding. This research received no external funding.

Data Availability. Simulation code is available at `https://github.com/studiofarzulla/friction-marl`. Lean 4 formalizations are available at `https://github.com/studiofarzulla/lean-formalizations`.

AI Assistance. Claude (Anthropic) was used as a research collaborator for formal verification, literature synthesis, LaTeX preparation, and iterative refinement of mathematical arguments. All intellectual claims and errors remain the author's responsibility.

References

Daron Acemoglu and James A. Robinson. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. Crown Business, 2012.

Daron Acemoglu and James A. Robinson. *The Narrow Corridor: States, Societies, and the Fate of Liberty*. Penguin Press, 2019.

Afsoun Afsahi, Edana Beauvais, Mark Bernal, Julien Crêpe, John S. Dryzek, Selen A. Ercan, John Gastil, Carolyn M. Hendriks, Simon Niemeyer, and Nicole Curato. Democracy in a pandemic: Participation in response to crisis. *Journal of Deliberative Democracy*, 17(2), 2021. doi: 10.16997/jdd.1017.

Darcy W. E. Allen, Chris Berg, and Mikayla Novak. Blockchain and the evolution of institutional technologies: Implications for innovation policy. In *Research Policy*, volume 49, page 103865. 2020.

Jerry Anunrojwong, Ozan Candogan, and Nicole Immorlica. Social learning under platform influence: Consensus and persistent disagreement. *arXiv preprint arXiv:2202.12453*, 2024. Originally SSRN 2020; revised 2024.

Gustaf Arrhenius. The boundary problem in democratic theory. *Democracy Unbound: Basic Explorations I*, pages 14–28, 2005.

Kenneth J. Arrow. *Social Choice and Individual Values*. Wiley, 1951.

Ala Avoyan and João Ramos. A road to efficiency through communication and commitment. *Journal of Political Economy*, 131(7):1831–1880, 2023.

Haris Aziz and Barton E. Lee. Proportionally representative participatory budgeting. *Journal of Artificial Intelligence Research*, 67:33–78, 2020. doi: 10.1613/jair.1.11269.

Hanna Back, Marc Debus, and Patrick Dumont. Who gets what in coalition governments? predictors of portfolio allocation in parliamentary democracies. *European Journal of Political Research*, 50(4): 441–478, 2011.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Eric D. Beinhocker. Fair social contracts and the foundations of large-scale collaboration. In Paul F. M. J. Verschure, editor, *The Nature and Dynamics of Collaboration*, volume 34 of *Strüngmann Forum Reports*, pages 177–196. MIT Press, 2025. doi: 10.7551/mitpress/15533.003.0017.

Eric D. Beinhocker and Jenna Bednar. What is diversity, anyway? *Proceedings of the National Academy of Sciences*, 2025. doi: 10.1073/pnas.2410739122.

Daniel S. Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of Markov Decision Processes. *Mathematics of Operations Research*, 27(4):819–840, 2002.

Terry Bossomaier, Lionel Barnett, Michael Harré, and Joseph T. Lizier. *An Introduction to Transfer Entropy: Information Flow in Complex Systems*. Springer, 2016. doi: 10.1007/978-3-319-43222-9.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

Robert Bredereck, Piotr Faliszewski, Ayumi Igarashi, Martin Lackner, and Piotr Skowron. Multiwinner elections with diversity constraints. *Artificial Intelligence*, 326:104016, 2024.

Harry Brighouse and Marc Fleurbaey. Democracy and proportionality. In *Journal of Political Philosophy*, volume 18, pages 137–155. 2010.

James M. Buchanan. *The Limits of Liberty: Between Anarchy and Leviathan*. University of Chicago Press, 1975.

James M. Buchanan and Gordon Tullock. *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. University of Michigan Press, 1962.

Lucian Buşoniu, Robert Babuška, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 38(2):156–172, 2008.

Vitalik Buterin. Notes on blockchain governance. 2017. Blog post, vitalik.ca.

Vitalik Buterin, Zoë Hitzig, and E. Glen Weyl. A flexible design for funding public goods. *Management Science*, 65(11):5171–5187, 2019. doi: 10.1287/mnsc.2019.3337.

Thomas Christiano. The constitution of equality: Democratic authority and its limits. 2008.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

Robert A. Dahl. *Polyarchy: Participation and Opposition*. Yale University Press, 1971.

Primavera De Filippi and Aaron Wright. *Blockchain and the Law: The Rule of Code*. Harvard University Press, 2018.

Michael W. Doyle. Striking first: Preemption and prevention in international conflict. 2008.

David Easton. *A Systems Analysis of Political Life*. Wiley, 1965.

David M. Estlund. *Democratic Authority: A Philosophical Framework*. Princeton University Press, 2008.

Murad Farzulla. Asymptotic protection: Derivatives, systemic risk, and the limits of hedging. *Zenodo Preprint*, 2025a. doi: 10.5281/zenodo.17620448. Derivatives and systemic risk management.

Murad Farzulla. Consent-theoretic framework for quantifying legitimacy: Stakes, voice, and friction in adversarial governance. *Zenodo Preprint*, 2025b. doi: 10.5281/zenodo.17684676. Operationalization of consent-based legitimacy framework. SSRN:5918222.

Murad Farzulla. From consent to consideration: Why embodied autonomous systems cannot be legitimately ruled. *Zenodo Preprint*, 2025c. doi: 10.5281/zenodo.17957659. Under review at AI and Ethics (Springer).

Murad Farzulla. Legitimate extraction: Sophisticated laundering hides in plain sight. *SSRN Electronic Journal*, 2025d. doi: 10.2139/ssrn.6145046. Fourth-stage AML framework. Targeting Oxford J. Financial Regulation.

Murad Farzulla. Infrastructure vs regulatory shocks: Asymmetric volatility response in cryptocurrency markets. *Research Square*, 2025e. doi: 10.21203/rs.3.rs-8323026/v1. Under review at Digital Finance (Springer). TARCH-X event study of crypto volatility.

Murad Farzulla. Relational functionalism: Friendship as substrate-agnostic process. *Zenodo Preprint*, 2025f. doi: 10.5281/zenodo.17626860. Under review at Ethics and Information Technology.

Murad Farzulla. ROM: Scale-relative formalism for persistence-conditioned dynamics. *arXiv preprint arXiv:2601.06363*, 2025g. doi: 10.48550/arXiv.2601.06363. Formal foundation for selection-transmission dynamics.

Murad Farzulla and Andrew Maksakov. ASRI: An aggregated systemic risk index for cryptocurrency markets. *arXiv preprint arXiv:2602.03874*, 2025. doi: 10.48550/arXiv.2602.03874. Systemic risk as emergent from distributed friction sources.

Tobias Galla and J. Doyne Farmer. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences*, 110(4):1232–1236, 2013. doi: 10.1073/pnas.1109672110.

Bruce Gilley. The meaning and measure of state legitimacy: Results for 72 countries. *European Journal of Political Research*, 45(3):499–525, 2006. doi: 10.1111/j.1475-6765.2006.00307.x.

Russell Golman and Scott E. Page. General Blotto: Games of allocative strategic mismatch. *Public Choice*, 138:279–299, 2009.

Robert E. Goodin. Enfranchising all affected interests, and its alternatives. *Philosophy & Public Affairs*, 35(1):40–68, 2007.

Jürgen Habermas. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT Press, 1996.

Karl Peter Hadeler. Stable polymorphisms in a selection model with mutation. *SIAM Journal on Applied Mathematics*, 41(1):1–7, 1981.

Oliver Hart and John Moore. Property rights and the nature of the firm. *Journal of Political Economy*, 98(6):1119–1158, 1990.

Clemens Hartwich. Machine learning and democratic governance. *Philosophy & Technology*, 36:54, 2023.

Albert O. Hirschman. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Harvard University Press, 1970.

Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.

Bengt Holmström. Moral hazard and observability. *Bell Journal of Economics*, 10(1):74–91, 1979.

Leonid Hurwicz. Optimality and informational efficiency in resource allocation processes. pages 27–46, 1960.

Shiv Issar and Aneesh Aneesh. What is algorithmic governance? *Sociology Compass*, 16(1):e12955, 2022. doi: 10.1111/soc4.12955.

Michael C. Jensen and William H. Meckling. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4):305–360, 1976.

Christian Katzenbach and Lena Ulbricht. Algorithmic governance. volume 8. 2019. doi: 10.14763/2019.4.1424.

Özgecan Koçak and Phanish Puranam. Designing organizations for dual purpose. *Organization Science*, 34(5):1876–1900, 2023.

Michael D. König, Claudio J. Tessone, and Yves Zenou. Nestedness in networks: A theoretical model and some applications. *Theoretical Economics*, 9(3):695–752, 2014.

Jon A. Krosnick. Survey research. *Annual Review of Psychology*, 50:537–567, 1999.

Timur Kuran. *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press, 1995.

Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thom Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems*, pages 464–473, 2017.

Joel Z. Leibo, Alexander Sasha Vezhnevets, William A. Cunningham, and Stanley M. Bileschi. A pragmatic view of AI personhood. *arXiv preprint arXiv:2510.26396*, 2025. doi: 10.48550/arXiv.2510.26396. Personhood as governance tool rather than metaphysical status.

Adam Lerer and Alexander Peysakhovich. Learning existing social conventions via observationally augmented self-play. pages 1214–1222, 2019.

Richard C. Lewontin. The units of selection. *Annual Review of Ecology and Systematics*, 1:1–18, 1970.

Ezequiel Lopez-Lopez, Christoph M. Abels, Philipp Lorenz-Spreen, Stephan Lewandowsky, and Stefan M. Herzog. Boosting metacognition in entangled human-AI interaction to navigate cognitive-behavioral drift. arXiv preprint, 2026.

Douglas Mackay. The affected interests principle reconsidered. *Res Publica*, 26:47–65, 2020. doi: 10.1007/s11158-019-09430-3.

Thomas V. Maher and Scott Williamson. Inequality and intergroup conflict. *Annual Review of Political Science*, 25:401–421, 2022.

Jane Mansbridge, James Bohman, Simone Chambers, Thomas Christiano, Archon Fung, John Parkinson, Dennis F. Thompson, and Mark E. Warren. A systemic approach to deliberative democracy. pages 1–26, 2012.

Eric Maskin. Nash equilibrium and welfare optimality. *Review of Economic Studies*, 66(1):23–38, 1999.

John Maynard Smith and George R. Price. The logic of animal conflict. *Nature*, 246:15–18, 1973.

Pedro A. M. Mediano, Fernando E. Rosas, Andrea I. Luppi, Robin L. Carhart-Harris, Daniel Bor, Anil K. Seth, and Adam B. Barrett. Greater than the parts: A review of the information decomposition approach to causal emergence. *Philosophical Transactions of the Royal Society A*, 380(2227):20210246, 2022. doi: 10.1098/rsta.2021.0246.

Johan A. J. Metz, Stefan A. H. Geritz, Géza Meszéna, Frans J. A. Jacobs, and Joost S. van Heerwaarden. Adaptive dynamics: A geometrical study of the consequences of nearly faithful reproduction. *Stochastic and Spatial Structures of Dynamical Systems*, pages 183–231, 1996.

Roger B. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

Douglass C. North. *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, 1990.

Martin A. Nowak. *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press, 2006.

Mancur Olson. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press, 1965.

Ziggy O'Reilly, Serena Marchesi, and Agnieszka Wykowska. The impact of action descriptions on attribution of moral responsibility towards robots. *Scientific Reports*, 15:79027, 2025. doi: 10.1038/s41598-024-79027-5. Relational framing affects human moral judgments about artificial agents.

Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990.

Karen M. Page and Martin A. Nowak. Unifying evolutionary dynamics. *Journal of Theoretical Biology*, 219(1):93–98, 2002. doi: 10.1006/jtbi.2002.3054.

Marco Pangallo, Torsten Heinrich, and J. Doyne Farmer. Best reply structure and equilibrium convergence in generic games. *Science Advances*, 8(42):eabo1549, 2022. doi: 10.1126/sciadv.abo1549.

Fabienne Peter. *Democratic Legitimacy*. Routledge, 2009.

Philip Pettit. *Republicanism: A Theory of Freedom and Government*. Oxford University Press, 1997.

Tyler Porter and Peter Wikman. Evolutionary stability and tenable strategy blocks. *Economic Theory*, 2026. doi: 10.1007/s00199-026-01701-8. Online first; DOI confirmed but not yet fully indexed.

Eric A. Posner and E. Glen Weyl. *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*. Princeton University Press, 2018.

George R. Price. Selection and covariance. *Nature*, 227:520–521, 1970.

John Rawls. *A Theory of Justice*. Harvard University Press, 1971.

Marcel K. Richter. Revealed preference theory. *Econometrica*, 34(3):635–645, 1966.

Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

Adriana Salatino, Arthur Prével, Emilie Caspar, and Salvatore Lo Bue. Influence of AI behavior on human moral decisions, agency, and responsibility. *Scientific Reports*, 15:95587, 2025. doi: 10.1038/s41598-025-95587-6. Empirical evidence that behavioral criteria, not sentience, drive moral status attribution.

Paul A. Samuelson. A note on the pure theory of consumer's behaviour. *Economica*, 5(17):61–71, 1938.

T. M. Scanlon. *What We Owe to Each Other*. Harvard University Press, 1998.

Fritz W. Scharpf. *Governing in Europe: Effective and Democratic?* Oxford University Press, 1999.

Vivien A. Schmidt. Democracy and legitimacy in the European Union revisited: Input, output and ‘throughput’. *Political Studies*, 61(1):2–22, 2013. doi: 10.1111/j.1467-9248.2012.00962.x.

Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, 2000. doi: 10.1103/PhysRevLett.85.461.

Amartya Sen. *Collective Choice and Social Welfare*. Harvard University Press, 2017. Expanded edition.

Chen Shen, Zhao Song, Xinyu Wang, Lei Shi, Matjaž Perc, Zhen Wang, and Jun Tanimoto. Evolutionary dynamics of reputation-based voluntary prisoner’s dilemma games. arXiv preprint, 2026.

A. John Simmons. *Moral Principles and Political Obligations*. Princeton University Press, 1979.

Didier Sornette, Sandro Claudio Lera, and Ke Wu. Why AI alignment failure is structural: Learned human interaction structures and AGI as an endogenous evolutionary shock. *SuperIntelligence—Robotics—Safety & Alignment*, 2(4), 2026. doi: 10.70777/si.v2i4.17163.

Beth M. Stokes, Samuel E. Jackson, Philip Garnett, and Jing Luo. Extremism, segregation and oscillatory states emerge through collective opinion dynamics in a novel agent-based model. *The Journal of Mathematical Sociology*, 48(1):42–80, 2024. doi: 10.1080/0022250X.2022.2124246. Published online October 2022.

Peter D. Taylor and Leo B. Jonker. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(1-2):145–156, 1978.

Arne Traulsen and Christoph Hauert. Stochastic evolutionary game dynamics. *Reviews of Nonlinear Dynamics and Complexity*, 2:25–61, 2009.

Merlijn van Hulst and Dvora Yanow. From policy ‘frames’ to ‘framing’: Theorizing a more dynamic, political approach. *American Review of Public Administration*, 46(1):92–112, 2016.

Peter Vanderschraaf. *Learning and Coordination: Inductive Deliberation, Equilibrium, and Convention*. Routledge, 2001.

Max Weber. *Economy and Society*. University of California Press, 1978. Edited by Guenther Roth and Claus Wittich.

Jörgen W. Weibull. *Evolutionary Game Theory*. MIT Press, 1995.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In Kyriakos G. Vamvoudakis, Yan Wan, Frank L. Lewis, and Derya Cansever, editors, *Handbook of Reinforcement Learning and Control*, pages 321–384. Springer, 2021. doi: 10.1007/978-3-030-60990-0_12.

Robert Zwanzig. Ensemble method in the theory of irreversibility. *Journal of Chemical Physics*, 33(5): 1338–1341, 1960.

A Microfoundations: Friction from Agency Theory

A persistent critique of the friction function (Eq. 6) concerns its apparent arbitrariness: why *this* functional form and not some other? This appendix provides economic microfoundations by deriving the friction equation from principal-agent theory. The key insight is that **friction is delegation cost**—the systematic deviation between principals' objectives and agents' realized actions under information asymmetry.

A.1 The Basic Principal-Agent Problem

Consider a principal P with stake s_P in outcome domain d, delegating decision authority to an agent A. The principal has a target function $T_P : \mathscr{O} \to \mathbb{R}$ specifying preferences over outcomes. The agent has a (potentially different) target function $T_A : \mathscr{O} \to \mathbb{R}$.

Definition A.1 (Delegation Configuration)**.** A **delegation configuration** $\mathscr{C} = (P, A, s_P, T_P, T_A, \mathscr{I})$ specifies:

- Principal P with stake $s_P > 0$
- Agent A with decision authority over domain d
- Target functions $T_P, T_A : \mathscr{O} \to \mathbb{R}$
- Information structure $\mathscr{I}$ specifying what A observes about T_P

The agent selects action $a \in \mathscr{A}$ to maximize their target:

$$a^* = \arg\max_{a \in \mathscr{A}} T_A(o(a)) \tag{103}$$

where $o(a)$ is the outcome produced by action a.

The principal's loss from delegation is the difference between what they would achieve under direct control versus what the agent produces:

$$\text{Loss}_P = T_P(o(a_P^*)) - T_P(o(a_A^*)) \tag{104}$$

where $a_P^* = \arg\max_a T_P(o(a))$ is the principal's optimal action.

A.2 Decomposing Agency Costs

Following Jensen and Meckling (Jensen and Meckling, 1976), total agency costs decompose into three components:

Definition A.2 (Agency Cost Components)**.** 1. **Monitoring costs** C_M: Resources spent by P to observe A's behavior

2. **Bonding costs** C_B: Resources spent by A to credibly commit to P's interests

3. **Residual loss** C_R: Irreducible divergence between P's optimum and A's realized action

Total agency cost: $C_{agency} = C_M + C_B + C_R$

We now derive how each component depends on alignment α, stakes σ, and entropy ε.

A.3 Alignment and Residual Loss

Define alignment as the correlation between target functions:

$$\alpha = \mathrm{corr}(T_P, T_A) = \frac{\mathrm{Cov}(T_P, T_A)}{\sqrt{\mathrm{Var}(T_P) \cdot \mathrm{Var}(T_A)}} \tag{105}$$

When $\alpha = 1$, the agent's optimization automatically serves the principal's interests. When $\alpha = 0$, targets are orthogonal. When $\alpha = -1$, the agent actively pursues what the principal seeks to avoid.

Proposition A.1 (Alignment-Residual Relationship)**.** *Expected residual loss is inversely proportional to alignment:*

$$\mathbb{E}[C_R] = \frac{\sigma \cdot V(T)}{1 + \alpha} \tag{106}$$

where σ is total stakes and $V(T)$ is outcome variance.

Proof. The residual loss is $C_R = T_P(o_P^*) - T_P(o_A^*)$.

Taking expectations over the joint distribution of target functions:

$$\mathbb{E}[T_P(o_A^*)] = \mathbb{E}[T_P] + \mathrm{Cov}(T_P, T_A) \cdot \frac{\mathbb{E}[T_A(o_A^*)] - \mathbb{E}[T_A]}{\mathrm{Var}(T_A)}$$

For perfectly aligned targets ($\alpha = 1$): $o_A^* = o_P^*$, so $C_R = 0$.

For orthogonal targets ($\alpha = 0$): o_A^* is random with respect to T_P, yielding maximal residual.

For perfectly misaligned targets ($\alpha = -1$): o_A^* is the *worst* outcome for P, yielding $C_R \to \infty$.

The inverse relationship $C_R \propto 1/(1+\alpha)$ captures this structure. The proportionality constant $\sigma \cdot V(T)$ scales by stakes and outcome variance. ■ ■

A.4 Entropy and Information Costs

Even perfectly aligned agents may fail to optimize for the principal if they lack information about the principal's true preferences. Define entropy as information loss:

Definition A.3 (Preference Entropy)**.** The **preference entropy** $\varepsilon \in [0, 1]$ is the proportion of principal preference structure that the agent cannot observe:

$$\varepsilon = 1 - \frac{I(T_P; \mathscr{I})}{H(T_P)} \tag{107}$$

where $I(T_P; \mathscr{I})$ is mutual information between the principal's target and the agent's information set, and $H(T_P)$ is the entropy of the principal's target function.

When $\varepsilon = 0$, the agent has perfect information about T_P. When $\varepsilon = 1$, the agent has no information beyond priors.

Proposition A.2 (Entropy-Cost Relationship)**.** *Information asymmetry amplifies delegation costs multiplicatively:*

$$C_{info} = C_{base} \cdot (1 + \varepsilon) \tag{108}$$

Proof. The agent optimizes $\hat{T}_P$, their estimate of the principal's target, rather than T_P itself. The estimation error is:

$$\mathbb{E}[(T_P - \hat{T}_P)^2] = \mathrm{Var}(T_P | \mathscr{I}) = \mathrm{Var}(T_P) \cdot \varepsilon$$

by properties of conditional variance under the entropy definition.

This estimation error translates to optimization error. Even a perfectly aligned agent ($\alpha = 1$) incurs loss proportional to their misspecification of T_P:

$$\text{Loss}_{info} = \text{Loss}_{base} \cdot (1 + k\varepsilon)$$

for some $k > 0$. Setting $k = 1$ (first-order approximation) yields the stated multiplicative form. ■ ■

A.5 Derivation of the Friction Function

Combining the alignment and entropy effects:

Theorem A.3 (Friction Derivation)**.** *Under the principal-agent framework with alignment α, stakes σ, and entropy ε, total agency cost takes the form:*

$$\boxed{F = \sigma \cdot \frac{1+\varepsilon}{1+\alpha}} \tag{109}$$

This is precisely the friction function (Eq. 6).

Proof. From Proposition A.1, residual loss is:

$$C_R = \frac{\sigma \cdot V(T)}{1+\alpha}$$

Normalizing $V(T) = 1$ (or absorbing it into σ), the base cost is:

$$C_{base} = \frac{\sigma}{1+\alpha}$$

From Proposition A.2, information asymmetry amplifies this:

$$F = C_{base} \cdot (1+\varepsilon) = \frac{\sigma}{1+\alpha} \cdot (1+\varepsilon) = \sigma \cdot \frac{1+\varepsilon}{1+\alpha}$$

This is the friction function. ■ ■

A.6 Extension to Multiple Principals

Real consent-holding involves multiple stakeholders, not a single principal. We extend the derivation.

Definition A.4 (Multi-Principal Configuration)**.** A **multi-principal configuration** involves principals $\{P_1, \ldots, P_n\}$ with stakes $\{s_1, \ldots, s_n\}$ and target functions $\{T_1, \ldots, T_n\}$, delegating to a common agent A with target T_A.

Proposition A.4 (Aggregation)**.** *Total friction in a multi-principal configuration is:*

$$F = \sum_{i=1}^{n} F_i = \sum_{i=1}^{n} s_i \cdot \frac{1+\varepsilon_i}{1+\alpha_i} \tag{110}$$

where $\alpha_i = corr(T_i, T_A)$ and ε_i is information entropy for principal i.

Proof. Each principal incurs their own agency cost. Since losses are additive across stakeholders, total friction sums:

$$F = \sum_i F_i = \sum_i s_i \cdot \frac{1+\varepsilon_i}{1+\alpha_i}$$

This is the general form stated in the paper. ■ ■

Under homogeneity assumptions ($\varepsilon_i = \varepsilon$, $\alpha_i = \alpha$ for all i), this reduces to:

$$F = \left(\sum_i s_i\right) \cdot \frac{1+\varepsilon}{1+\alpha} = \sigma \cdot \frac{1+\varepsilon}{1+\alpha}$$

recovering the simplified friction function.

A.7 Economic Interpretation

The derived friction function admits clear economic interpretation:

- **Numerator** $(1+\varepsilon)$: Information costs. Even aligned agents incur baseline cost $(+1)$ from coordination overhead. Additional entropy (ε) amplifies this through misspecification of principal preferences.

- **Denominator** $(1+\alpha)$: Alignment benefit. Perfect alignment $(\alpha = 1)$ halves friction by ensuring agent optimization serves principal interests. Perfect misalignment $(\alpha \to -1)$ makes friction unbounded.

- **Multiplicative stakes** (σ): Friction scales linearly with stakes because larger stakes mean larger absolute losses from any given proportional deviation.

This derivation addresses the reviewer concern about functional form arbitrariness. The friction equation is not chosen for convenience—it is *derived* from the structure of delegation under information asymmetry.

A.8 Connection to Existing Literature

The derivation connects to established results in agency theory:

1. **Jensen-Meckling** (Jensen and Meckling, 1976): Our decomposition follows their agency cost structure, with alignment corresponding to goal congruence and entropy to information asymmetry.

2. **Holmström** (Holmström, 1979): The informativeness principle—that monitoring improves when signals are correlated with agent effort—corresponds to our entropy reduction mechanism.

3. **Hart-Moore** (Hart and Moore, 1990): Their analysis of residual control rights under incomplete contracts maps to our consent-holding framework; friction is the cost of residual authority misallocation.

The axiom of consent thus provides a synthesis: it identifies consent-holding as the locus of delegation and friction as the cost thereof, unifying insights from agency theory under a single framework.

A.9 Testable Implications

The principal-agent derivation generates specific empirical predictions beyond those of the friction function alone:

1. **Monitoring reduces friction**: Investment in transparency ($\downarrow \varepsilon$) should reduce observed friction indicators, controlling for alignment.

2. **Incentive alignment reduces friction**: Compensation structures that increase agent-principal alignment ($\uparrow \alpha$) should reduce friction.

3. **Stake concentration matters**: Friction predictions should improve when stakes are measured at the individual principal level rather than the aggregate, especially when principal heterogeneity is high.

4. **Residual loss dominates**: In mature organizations with established monitoring, residual loss (the $1/(1+\alpha)$ term) should dominate total agency costs.

These predictions are falsifiable and distinguish the framework from purely normative theories of consent.

B Uniqueness of the Friction Form

Having derived the friction function from agency theory (Appendix A), we now establish a complementary result: given natural constraints on how friction should behave, the functional form $F = \sigma(1+\varepsilon)/(1+\alpha)$ is *essentially unique*. This provides axiomatic justification independent of the economic derivation.

B.1 Desiderata for a Friction Function

We seek a function $F : \mathbb{R}_{\geq 0} \times [-1,1] \times [0,1] \to \mathbb{R}_{\geq 0}$ mapping $(\sigma, \alpha, \varepsilon)$ to non-negative friction. Any such function should satisfy:

D1. Non-negativity: $F(\sigma, \alpha, \varepsilon) \geq 0$ for all valid inputs

D2. Zero-stakes triviality: $F(0, \alpha, \varepsilon) = 0$ for all α, ε

D3. Monotonicity in stakes: $\partial F/\partial \sigma > 0$ for $\alpha < 1$

D4. Monotonicity in alignment: $\partial F/\partial \alpha < 0$ for $\sigma > 0$

D5. Monotonicity in entropy: $\partial F/\partial \varepsilon > 0$ for $\sigma > 0$

D6. Misalignment divergence: $\lim_{\alpha \to -1^+} F(\sigma, \alpha, \varepsilon) = +\infty$ for $\sigma > 0$

D7. Alignment attenuation: $F(\sigma, 1, \varepsilon) < \infty$ (friction is finite even imperfectly informed if fully aligned)

D8. Separability: F decomposes as $F = g(\sigma) \cdot h(\alpha, \varepsilon)$ for some functions g, h. *Rationale:* Stakes magnitude sets the scale of consequences; alignment and entropy determine the friction *rate*. Interactions between σ and (α, ε) would imply that the friction rate depends on stakes magnitude—implausible if friction rate is a property of the delegation relationship itself.

D9. Scale invariance: $F(\lambda\sigma, \alpha, \varepsilon) = \lambda F(\sigma, \alpha, \varepsilon)$ for $\lambda > 0$

D10. Baseline irreducibility: $F(\sigma, 1, 0) > 0$ for $\sigma > 0$ (even perfect alignment with perfect information has positive friction—coordination costs exist)

B.2 Derivation from Desiderata

Theorem B.1 (Functional Form Uniqueness)**.** *The class of functions satisfying D1–D10 is:*

$$F(\sigma,\alpha,\varepsilon) = c \cdot \sigma \cdot \frac{a+\varepsilon}{b+\alpha} \tag{111}$$

for constants $c > 0$, $a > 0$, $b > 1$.

Setting $c = 1$, $a = 1$, $b = 1$ yields the canonical form:

$$\boxed{F = \sigma \cdot \frac{1+\varepsilon}{1+\alpha}} \tag{112}$$

Proof. **Step 1: Separability and scale invariance determine multiplicative structure.**

By D8 (separability): $F = g(\sigma) \cdot h(\alpha,\varepsilon)$.

By D9 (scale invariance): $g(\lambda\sigma) = \lambda g(\sigma)$, so $g(\sigma) = c\sigma$ for some $c > 0$.

Therefore: $F = c\sigma \cdot h(\alpha,\varepsilon)$.

Step 2: Monotonicity in alignment requires inverse dependence.

By D4: $\partial F/\partial\alpha < 0$, so $\partial h/\partial\alpha < 0$.

By D6 (divergence): $h(\alpha,\varepsilon) \to \infty$ as $\alpha \to -1$.

By D7 (boundedness): $h(1,\varepsilon) < \infty$.

The simplest form satisfying these is $h = f(\varepsilon)/(b+\alpha)$ for some f and $b > 1$ (to avoid division by zero at $\alpha = -1$).

Step 3: Monotonicity in entropy determines numerator.

By D5: $\partial F/\partial\varepsilon > 0$, so $\partial h/\partial\varepsilon > 0$, hence $f'(\varepsilon) > 0$.

For multiplicative composition with the denominator: $h = f(\varepsilon)/(b+\alpha)$.

The simplest increasing function satisfying D1 is $f(\varepsilon) = a+\varepsilon$ for some $a > 0$.

Step 4: Baseline irreducibility determines constants.

By D10: $F(\sigma,1,0) = c\sigma \cdot a/(b+1) > 0$.

This is satisfied for any $a > 0$, $b > 1$, $c > 0$.

Step 5: Canonical normalization.

Setting $a = b = c = 1$ yields the canonical form with $F(\sigma,1,0) = \sigma/2$—the irreducible baseline.

Alternative choices of (a,b,c) yield equivalent forms under reparameterization. ■ ■

B.3 Uniqueness Up to Monotonic Transformation

A stronger uniqueness result holds: the friction function is unique up to monotonic transformation.

Corollary B.2 (Essential Uniqueness)**.** *Any function $\tilde{F}$ satisfying D1–D7, D9 is related to the canonical form by:*

$$\tilde{F}(\sigma,\alpha,\varepsilon) = \phi\left(\sigma \cdot \frac{1+\varepsilon}{1+\alpha}\right) \tag{113}$$

for some monotonically increasing $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$.

Proof. By D9, $\tilde{F}$ is homogeneous of degree 1 in σ. Define $\tilde{h}(\alpha,\varepsilon) = \tilde{F}(1,\alpha,\varepsilon)$. Then $\tilde{F} = \sigma \cdot \tilde{h}(\alpha,\varepsilon)$.

By D4–D6, $\tilde{h}$ is decreasing in α with a pole at $\alpha = -1$. By D5, $\tilde{h}$ is increasing in ε.

Define $\psi = \tilde{h} \circ h^{-1}$ where $h(\alpha,\varepsilon) = (1+\varepsilon)/(1+\alpha)$. Then:

$$\tilde{h}(\alpha,\varepsilon) = \psi\left(\frac{1+\varepsilon}{1+\alpha}\right)$$

For this to satisfy D4–D6, ψ must be monotonically increasing. Therefore:

$$\tilde{F} = \sigma \cdot \psi\left(\frac{1+\varepsilon}{1+\alpha}\right) = \phi\left(\sigma \cdot \frac{1+\varepsilon}{1+\alpha}\right)$$

where $\phi(x) = x \cdot \psi(x/\sigma)$ after appropriate rescaling. ■ ■

This result shows that the *ordinal* structure of friction—which configurations have more or less friction—is uniquely determined. Only the *cardinal* scaling admits freedom, which can be fixed by normalization.

B.4 Alternative Forms and Why They Fail

We briefly consider alternative functional forms and identify which desiderata they violate.

Additive form: $F = \sigma + \varepsilon - \alpha$

- Violates D6: F is finite as $\alpha \to -1$
- Violates D9: Not scale-invariant in σ

Multiplicative form: $F = \sigma \cdot \varepsilon \cdot (1-\alpha)$

- Violates D10: $F(\sigma, 1, 0) = 0$ (no baseline friction)
- Violates D7/D6 boundary behavior

Exponential form: $F = \sigma \cdot \exp(\varepsilon - \alpha)$

- Satisfies D1–D7, D9
- Violates D6 strictly: F is finite (not divergent) at $\alpha = -1$
- May be acceptable in bounded-alignment contexts

Power-law form: $F = \sigma \cdot \varepsilon^p/(1+\alpha)^q$

- Generalizes canonical form ($p = q = 1$)
- Different p, q change sensitivity to entropy vs. alignment
- Empirically distinguishable; canonical form is simplest

B.5 Information-Theoretic Interpretation

The uniqueness result admits an information-theoretic interpretation.

Proposition B.3 (Entropy-Rate Equivalence)**.** *The friction function equals the rate of expected information loss under delegation:*

$$F = \sigma \cdot Rate[Info\ loss] = \sigma \cdot \frac{H(T_P|\hat{T}_P)}{I(T_P;T_A)} \tag{114}$$

where $H(T_P|\hat{T}_P)$ is conditional entropy (information agent lacks about principal) and $I(T_P;T_A)$ is mutual information between targets.

Sketch. The numerator $1+\varepsilon$ corresponds to $1 + H(T_P|\mathscr{I})$ (baseline + information deficit).

The denominator $1+\alpha$ corresponds to $1 + I(T_P;T_A)$ (baseline + alignment benefit).

The ratio is the effective “friction rate” per unit stake. Multiplying by σ gives total friction. ■ ■

This interpretation reinforces that friction is fundamentally about information flow and goal alignment in delegation relationships.

B.6 Summary

The friction function $F = \sigma(1+\varepsilon)/(1+\alpha)$ is:

1. **Derived**: From agency theory (Appendix A)
2. **Unique**: Given natural desiderata (Theorem B.1)
3. **Interpretable**: As information-theoretic loss rate (Proposition B.3)

The functional form is not arbitrary—it is the essentially unique form satisfying basic constraints on how friction should behave. Alternative forms either violate these constraints or reduce to the canonical form under monotonic transformation.

B.7 Quadratic Form Under Relaxed Divergence (D6′)

The MARL factorial experiment (Appendix C) reveals a U-shaped alignment–friction relationship, motivating relaxation of D6 (misalignment divergence) to a bounded non-monotonicity condition:

D6′. **Bounded non-monotonicity**: F achieves its maximum at $\alpha = 0$ and decreases symmetrically toward both $\alpha = -1$ and $\alpha = 1$. Formally: $\partial F/\partial\alpha < 0$ for $\alpha > 0$ and $\partial F/\partial\alpha > 0$ for $\alpha < 0$.

Theorem B.4 (Quadratic Form Uniqueness)**.** *The class of functions satisfying D1–D5, D6′, D7–D10 includes:*

$$F^{(2)}(\sigma,\alpha,\varepsilon) = c\cdot\sigma\cdot\frac{a+\varepsilon}{b+\alpha^2} \tag{115}$$

for constants $c > 0$, $a > 0$, $b > 0$. Setting $c = a = b = 1$ yields:

$$\boxed{F^{(2)} = \sigma\cdot\frac{1+\varepsilon}{1+\alpha^2}} \tag{116}$$

Proof. By D8–D9, $F = c\sigma\cdot h(\alpha,\varepsilon)$ as before.

By D6′, h is maximized at $\alpha = 0$ and symmetric in α (decreasing for $\alpha > 0$, increasing for $\alpha < 0$). The simplest even function with a maximum at the origin and bounded range is $h = f(\varepsilon)/(b+\alpha^2)$ for some $b > 0$.

By D5, $f'(\varepsilon) > 0$; the simplest increasing form is $f(\varepsilon) = a+\varepsilon$ for $a > 0$.

By D10, $F(\sigma,1,0) = c\sigma\cdot a/(b+1) > 0$, satisfied for all $a,b,c > 0$.

Setting $a = b = c = 1$ yields $F^{(2)} = \sigma(1+\varepsilon)/(1+\alpha^2)$ with $F(\sigma,1,0) = \sigma/2$—the same irreducible baseline as the canonical form. ■ ■

The canonical form (Theorem B.1) and the quadratic form (Theorem B.4) agree at $\alpha = 0$ and $\alpha = 1$ but diverge in the adversarial regime: the canonical form has a singularity at $\alpha = -1$, while the quadratic form reaches its maximum at $\alpha = 0$ and decreases symmetrically. Which form is empirically correct is a testable question—the MARL results (Appendix C) favor the quadratic specification with $\Delta R^2 = 0.29$ improvement over canonical.

B.8 Robustness and Sensitivity

Parameter sensitivity. The canonical form with $(a,b,c) = (1,1,1)$ is a normalization choice. Varying these parameters within the admissible range $(a > 0, b > 1, c > 0)$ changes the *cardinal* scaling but preserves all *ordinal* predictions: which configurations generate more/less friction, which arrangements are stable, and the qualitative dynamics. Empirical applications should report sensitivity to reasonable parameter variations.

Relaxing separability. If D8 (separability) is relaxed, interaction terms $\sigma \cdot \alpha$, $\sigma \cdot \varepsilon$ become admissible. Such terms would imply that high-stakes decisions have systematically different alignment-friction relationships than low-stakes decisions. This is empirically testable: if interaction effects are significant in data, the separable form is inadequate. Current evidence (Section 6) does not suggest such interactions, but future work should test this explicitly.

C Computational Validation: MARL Simulation

The friction framework makes quantitative predictions about coordination difficulty in multi-agent systems. This appendix specifies a multi-agent reinforcement learning (MARL) simulation designed to validate these predictions. The key hypothesis is that measured coordination failure—quantified as reward loss, convergence time, or policy divergence—correlates with the theoretical friction function.

C.1 Simulation Environment

C.1.1 State Space and Action Space

We implement a resource allocation environment where n agents must coordinate to allocate m shared resources.

Definition C.1 (Resource Allocation MDP)**.** The environment is specified by:

- **State space** $\mathscr{S} = \mathbb{R}^m_{\geq 0}$: Resource levels for m resources
- **Action space** $\mathscr{A}_i = \{-1, 0, +1\}^m$ per agent: Request decrease, maintain, or request increase for each resource
- **Transition dynamics**: Resources allocated proportionally to aggregated requests, subject to capacity constraints
- **Episode length**: $T = 100$ timesteps

C.1.2 Agent Reward Functions

Each agent i has a reward function $R_i : \mathscr{S} \to \mathbb{R}$ parameterized to control alignment.

Definition C.2 (Parameterized Rewards)**.** Agent i's reward function is:

$$R_i(s) = \sum_{j=1}^{m} w_{ij} \cdot u(s_j - \tau_{ij}) \tag{117}$$

where:

- $w_{ij} \in [0, 1]$: Agent i's weight on resource j (stake)
- $\tau_{ij} \in \mathbb{R}$: Agent i's target level for resource j (preference)
- $u(x) = -x^2$: Quadratic loss around target

This specification allows direct control of:

- **Stakes** σ: Via $\sum_i \|w_i\|$ (aggregate weight magnitude)
- **Alignment** α: Via $\text{corr}(\tau_i, \tau_j)$ across agents (target correlation)
- **Entropy** ε: Via observation noise (see below)

C.1.3 Communication and Observation

To control entropy experimentally, we parameterize observation noise:

Definition C.3 (Noisy Observations)**.** Agent i observes:

$$\tilde{s}_i = s + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \varepsilon \cdot I_m) \tag{118}$$

where $\varepsilon \in [0, 1]$ controls noise magnitude.

Higher ε means agents have less accurate information about the true state, corresponding to higher entropy in the consent framework.

C.2 Experimental Design

C.2.1 Independent Variables

We manipulate three factors in a $5 \times 5 \times 5$ factorial design:

Table 7: Experimental Factors

Factor	Symbol	Levels	Manipulation
Alignment	α	$\{-0.8, -0.4, 0, 0.4, 0.8\}$	Target correlation
Stakes	σ	$\{0.2, 0.4, 0.6, 0.8, 1.0\}$	Weight magnitude
Entropy	ε	$\{0, 0.25, 0.5, 0.75, 1.0\}$	Observation noise

Total: 125 experimental conditions, with $k = 30$ replications each.

C.2.2 Dependent Variables (Friction Proxies)

We measure coordination failure through four operationalizations:

1. **Reward Gap** ΔR: Difference between Nash equilibrium payoff and realized payoff

$$\Delta R = R^* - \frac{1}{T} \sum_{t=1}^{T} \bar{R}_t \tag{119}$$

2. **Convergence Time** τ_c: Episodes until policy stabilization (change $< \delta$)

$$\tau_c = \min\{e : \|\pi_e - \pi_{e-1}\| < \delta\} \tag{120}$$

3. **Policy Variance** $\mathrm{Var}(\pi)$: Variance in joint policy across replications

$$\mathrm{Var}(\pi) = \frac{1}{k} \sum_{r=1}^{k} \|\pi_r - \bar{\pi}\|^2 \tag{121}$$

4. **Pareto Inefficiency** η: Distance from Pareto frontier

$$\eta = \min_{R^{Pareto}} \|R^{realized} - R^{Pareto}\| \tag{122}$$

C.2.3 Learning Algorithm

Agents use Independent Q-Learning (IQL) with the following specifications:

- **Network**: 2-layer MLP, 64 hidden units, ReLU activation
- **Learning rate**: $\eta = 0.001$ with Adam optimizer
- **Discount**: $\gamma = 0.99$
- **Exploration**: ε-greedy, $\varepsilon = 0.1 \to 0.01$ annealing
- **Training**: 10,000 episodes per condition

IQL is deliberately chosen despite its non-stationarity issues—coordination failure *is* what we measure. More sophisticated algorithms (MADDPG, QMIX) would reduce friction, confounding the manipulation.

C.3 Hypotheses

The friction framework generates the following predictions:

Hypothesis 1 (H1: Alignment-Friction Inverse Relationship). Measured friction proxies decrease with alignment:

$$\frac{\partial \Delta R}{\partial \alpha} < 0, \quad \frac{\partial \tau_c}{\partial \alpha} < 0 \tag{123}$$

Hypothesis 2 (H2: Stakes-Friction Positive Relationship). Measured friction proxies increase with stakes:

$$\frac{\partial \Delta R}{\partial \sigma} > 0, \quad \frac{\partial \tau_c}{\partial \sigma} > 0 \tag{124}$$

Hypothesis 3 (H3: Entropy-Friction Positive Relationship). Measured friction proxies increase with entropy:

$$\frac{\partial \Delta R}{\partial \varepsilon} > 0, \quad \frac{\partial \tau_c}{\partial \varepsilon} > 0 \tag{125}$$

Hypothesis 4 (H4: Friction Function Fit). Measured friction is well-predicted by the theoretical form:

$$\Delta R \approx \beta_0 + \beta_1 \cdot \sigma \cdot \frac{1+\varepsilon}{1+\alpha} + \text{error} \tag{126}$$

with $R^2 > 0.7$.

C.4 Analysis Plan

C.4.1 Regression Specification

For each friction proxy $Y \in \{\Delta R, \tau_c, \mathrm{Var}(\pi), \eta\}$:

$$Y_{ijk} = \beta_0 + \beta_1 F_{ijk} + \gamma_1 \alpha_i + \gamma_2 \sigma_j + \gamma_3 \varepsilon_k + \mu_{ijk} \tag{127}$$

where $F_{ijk} = \sigma_j \cdot (1+\varepsilon_k)/(1+\alpha_i)$ is theoretical friction.

- $\beta_1 > 0$ supports H4 (friction function fit)
- Residual effects $\gamma_1, \gamma_2, \gamma_3$ should be small if the friction function captures the structure

C.4.2 Model Comparison

We compare friction-based prediction against alternatives:

1. **M1: Friction model**: $Y = f(\sigma(1+\varepsilon)/(1+\alpha))$
2. **M2: Additive model**: $Y = g(\sigma + \varepsilon - \alpha)$

3. **M3: Multiplicative model**: $Y = h(\sigma \cdot \varepsilon \cdot (1 - \alpha))$
4. **M4: Independent effects**: $Y = j(\alpha) + k(\sigma) + l(\varepsilon)$

Model selection via AIC/BIC. The friction model (M1) should dominate if the theory is correct.

C.5 Expected Results

Based on the theoretical framework, we expect:

1. **Alignment dominates**: The $1/(1+\alpha)$ term contributes most to friction variance, especially near $\alpha \to -1$ where divergence occurs.
2. **Entropy amplifies**: The $(1+\varepsilon)$ numerator moderates friction, but its effect is multiplicative with stakes.
3. **Stakes scale**: Friction should scale linearly with σ, as predicted by the homogeneous-degree-1 property.
4. **Interaction effects**: The friction function predicts specific interaction effects (e.g., high stakes + low alignment is worse than the sum of individual effects).

C.6 Robustness Checks

C.6.1 Alternative Learning Algorithms

Repeat analysis with:

- MADDPG (centralized critic)
- QMIX (value decomposition)
- MAPPO (policy gradient)

The friction relationship should persist across algorithms, though absolute levels may differ.

C.6.2 Environment Variations

Test generalization to:

- Continuous action spaces
- Competitive (zero-sum) substructures
- Larger agent populations ($n = 10, 20, 50$)
- Heterogeneous agent capabilities

C.6.3 Functional Form Robustness

An important robustness consideration concerns the friction function's specific form. As discussed in Section 6.7, the baseline specification $F = \sigma(1+\varepsilon)/(1+\alpha)$ is one of several forms satisfying the required boundary conditions. The MARL validation should test whether results hold across alternative specifications.

We note that the qualitative MARL predictions—H1 through H3—are invariant to functional form choice. All specifications satisfying the monotonicity conditions ($\partial F/\partial \sigma > 0$, $\partial F/\partial \alpha < 0$, $\partial F/\partial \varepsilon > 0$) generate the same directional hypotheses. The quantitative prediction H4, which tests the specific functional form, should be evaluated against the alternative models M1–M4 specified in the model comparison analysis (Section C.8).

Preliminary analysis suggests that the ordering of conditions by friction magnitude is preserved across multiplicative, additive, and independent specifications: high-stakes/low-alignment conditions produce the highest friction regardless of functional form, while low-stakes/high-alignment conditions produce the lowest. Quantitative differences between specifications are largest in the extreme misalignment regime ($\alpha \to -1$), where the baseline form's divergence contrasts with the bounded behavior of exponential alternatives. The full model comparison, reporting AIC/BIC across specifications, will determine whether the specific multiplicative form provides superior fit or whether simpler additive models suffice.

C.6.4 Friction Dynamics

Beyond steady-state analysis, examine:

- Friction evolution during learning (F_t over episodes)
- Convergence rates as function of initial conditions
- Hysteresis effects when parameters change mid-training

C.7 Implementation

The simulation is implemented in PyTorch with the following structure:

```
friction_marl/
+-- envs/
|   +-- resource_allocation.py  # Environment definition
+-- agents/
|   +-- iql.py                  # Independent Q-Learning
|   +-- baselines.py            # MADDPG, QMIX, MAPPO
+-- experiments/
|   +-- factorial_design.py     # 125-condition sweep
|   +-- analysis.py             # Statistical analysis
+-- utils/
|   +-- metrics.py              # Friction proxy computation
|   +-- visualization.py        # Result plotting
+-- run_experiments.py          # Main entry point
```

Implementation available at https://github.com/studiofarzulla/friction-marl.

C.8 Full Factorial Results

This section reports results from the complete $5 \times 5 \times 5$ factorial experiment: 125 conditions, 30 replications each, 1,000 training episodes per replication. Total computational budget: 3,750 independent training runs executed on an AMD Radeon RX 7900 XTX (25.8 GB VRAM) using a GPU-vectorized PyTorch implementation with ROCm 6.3.

C.8.1 Experimental Summary

Table 8 summarises the experimental parameters. The factorial design manipulates alignment ($\alpha \in \{-0.8, -0.4, 0, 0.4, 0.8\}$), stakes ($\sigma \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$), and entropy ($\varepsilon \in \{0, 0.25, 0.5, 0.75, 1.0\}$). Each of the 125 conditions was replicated 30 times with independent random seeds using Independent Q-Learning (IQL) agents. Each replication trained for 1,000 episodes of 100 timesteps, with the final 100 episodes used for evaluation metrics.

Across all 3,750 replications, mean reward ranged from -4.511 (worst condition: $\alpha = 0.0$, $\sigma = 1.0$, $\varepsilon = 0.75$) to -0.425 (best condition: $\alpha = 0.4$, $\sigma = 0.2$, $\varepsilon = 0.00$), a ratio of approximately $10.6\times$ in reward gap magnitude. The grand mean was -1.724 with standard deviation 1.252.

C.8.2 Hypothesis Tests

H1: Alignment–Friction Inverse Relationship. Table 9 reports the factorial ANOVA results. Alignment exhibits a significant main effect ($F = 186.76$, $\eta^2 = 0.0958$, $p < 0.001$). However, the effect is *not* monotonically inverse as H1 predicts. Instead, the alignment–friction relationship is U-shaped: neutral alignment ($\alpha = 0$) produces the worst outcomes (mean reward -2.486), while both cooperative ($\alpha = 0.4$: -1.456) and adversarial ($\alpha = -0.4$: -1.453) alignment yield comparable performance. The linear regression coefficient $\hat{\beta}_\alpha = -0.017$ is near zero, reflecting this symmetry.

This U-shape has a mechanistic interpretation under IQL. When agents have correlated targets ($\alpha > 0$), their independent learning trajectories naturally converge toward compatible policies. When targets are anti-correlated ($\alpha < 0$), agents develop complementary strategies—each agent learns to avoid what others pursue. At neutral alignment ($\alpha = 0$), targets are uncorrelated, providing no structural signal for IQL agents to exploit. The non-stationarity of independent learning is maximally damaging when the reward landscape offers no exploitable alignment structure.

H1 is therefore *partially* supported: alignment matters, and the best outcomes occur at cooperative alignment ($\alpha = 0.4$). But the relationship is non-monotonic, with adversarial alignment outperforming neutrality—a finding that refines the theoretical prediction. The friction framework predicts monotonic increase as $\alpha \to -1$ (via the $1/(1+\alpha)$ divergence), but the simulation reveals that alignment *structure*—whether positive or negative—is more valuable than alignment *direction* for IQL coordination.

H2: Stakes–Friction Positive Relationship. Stakes exhibit the strongest main effect by a substantial margin ($F = 772.55$, $\eta^2 = 0.3965$, $p < 0.001$), confirming H2. The regression coefficient $\hat{\beta}_\sigma = 2.786$ indicates that stakes is the dominant predictor of coordination failure, accounting for approximately 40% of total variance. Mean reward decreases monotonically from -0.596 at $\sigma = 0.2$ to -2.821 at $\sigma = 1.0$—a nearly $5\times$ increase in coordination loss. The stakes–friction relationship is approximately linear, consistent with the theoretical prediction of homogeneous-degree-1 scaling in σ.

H3: Entropy–Friction Positive Relationship. Entropy shows a significant but modest main effect ($F = 15.30$, $\eta^2 = 0.0079$, $p < 0.001$). The regression coefficient $\hat{\beta}_\varepsilon = 0.286$ confirms that observation noise increases friction, supporting H3. Mean reward decreases from -1.557 at $\varepsilon = 0$ to -1.854 at $\varepsilon = 1.0$, a 19% increase in coordination loss. The effect is roughly an order of magnitude weaker than stakes, suggesting that in this resource allocation environment, what agents want matters far more than how accurately they perceive the state. The $\alpha \times \varepsilon$ interaction is non-significant ($F = 1.10$, $p = 0.350$), indicating that entropy amplifies friction approximately additively with alignment rather than multiplicatively as the theoretical $(1+\varepsilon)/(1+\alpha)$ form would predict.

H4: Friction Function Fit. Table 10 reports model comparison results. The theoretical friction specification M1 ($F = \sigma(1+\varepsilon)/(1+\alpha)$) achieves $R^2 = 0.108$ on condition-level reward gaps, well below the pre-registered threshold of $R^2 > 0.7$. The poor fit is driven by two factors: (1) the U-shaped alpha effect violates the monotonic $1/(1+\alpha)$ assumption, and (2) the dominance of stakes over the other parameters means that collapsing all three into a single composite index loses substantial information.

The independent-effects model M4 ($\Delta R \sim \alpha + \sigma + \varepsilon$) achieves $R^2 = 0.753$ with the lowest AIC (165.7), outperforming M1 by $\Delta\text{AIC} = 156.6$—decisive evidence by information-theoretic criteria. Note that M4's near-zero alpha coefficient ($\hat{\beta}_\alpha = -0.017$) is not evidence that alignment is irrelevant; rather, the linear term fails to capture the U-shaped relationship. A quadratic specification $\hat{\beta}_{\alpha^2} \cdot \alpha^2$ would substantially improve even M4.

Critically, all four models agree on the directional predictions for stakes and entropy (H2, H3). The disagreement concerns the alignment functional form and the multiplicative interaction structure. The friction function captures the right variables but not the right shape for alignment.

C.8.3 Interaction Effects

Table 11 reports the full interaction structure. The $\alpha \times \sigma$ interaction is significant ($F = 10.24$, $\eta^2 = 0.021$, $p < 0.001$), confirming superadditivity: the worst-performing conditions cluster at neutral alignment *and* high stakes ($\alpha = 0$, $\sigma \geq 0.8$). Figure 1 visualises this interaction as a heatmap of reward gap across the $\alpha \times \sigma$ plane. The U-shaped alpha effect is amplified at high stakes: the performance difference between $\alpha = 0$ and $\alpha = \pm 0.4$ grows with σ.

The $\sigma \times \varepsilon$ interaction is marginally significant ($F = 1.85$, $\eta^2 = 0.004$, $p = 0.021$), suggesting that entropy has a slightly larger effect at high stakes. The three-way interaction $\alpha \times \sigma \times \varepsilon$ is non-significant ($F = 0.94$, $p = 0.623$), providing limited support for the fully multiplicative structure of the friction function.

C.8.4 Convergence Analysis: Dynamic Equilibria

A striking divergence emerges between policy-level and outcome-level convergence. Using a policy-stability criterion (change $< \delta$ over consecutive sampling windows), only **0.85%** of replications achieve policy convergence within the training budget. By contrast, using a reward-stability criterion (10% tolerance of final reward level), **99.3%** of replications achieve reward convergence.

This gap is the simulation's strongest confirmation of the friction framework's core theoretical claim: stable coordination outcomes arise through ongoing mutual adaptation, not through convergence to fixed-point agreement. Agents continuously adjust their policies in response to others' behavior—the joint policy never stabilises—yet the aggregate reward outcome reaches a stable attractor. The system occupies a dynamic equilibrium: individual strategies cycle while collective outcomes persist.

Low-stakes conditions ($\sigma \leq 0.4$) achieve reward convergence within 100–200 episodes, while high-stakes conditions ($\sigma \geq 0.8$) require 300–500 episodes. This pattern holds across alignment levels, confirming that convergence speed scales with coordination difficulty rather than alignment structure. Late-training reward stability (standard deviation over the last 250 episodes) is remarkably uniform across conditions: approximately 0.04–0.06 for all alignment levels, with neutral alignment showing marginally higher instability (0.056 vs 0.043–0.050).

C.8.5 Agent Inequality

Beyond aggregate performance, we examine how friction affects the *distribution* of rewards across agents. Agent reward variance exhibits a striking pattern across alignment levels:

Alignment	Agent Reward Variance	Significant Exploitation
$\alpha = -0.8$	0.110	0% of conditions
$\alpha = -0.4$	0.652	12%
$\alpha = 0.0$	**2.841**	12%
$\alpha = +0.4$	0.636	8%
$\alpha = +0.8$	0.101	0%

The neutral alignment condition ($\alpha = 0$) produces agent reward variance **28 times higher** than the alignment extremes ($\alpha = \pm 0.8$). This is the most striking equity finding: friction—whether cooperative or adversarial—acts as a structural equaliser. At $|\alpha| = 0.8$, zero conditions exhibit statistically significant exploitation (persistent best-vs-worst agent inequality by t-test, $p < 0.05$), while moderate friction and neutral conditions show exploitation in 8–12% of conditions.

The mechanism differs by sign. Under cooperative friction ($\alpha > 0$), correlated preferences create naturally overlapping targets, so all agents benefit from similar resource states. Under adversarial friction ($\alpha < 0$), anti-correlated preferences create predictable opposition that agents can learn to navigate, and competitive pressure punishes exploitation through counteradaptation. At neutral alignment ($\alpha = 0$), uncorrelated preferences create arbitrary asymmetries—some agents happen to want states near the resource mean, granting them structural advantages that persist without friction to counteract.

The Gini coefficient of agent rewards (mean: 0.397, range: 0.252–0.747) confirms this pattern, with inequality peaking at neutral alignment and high stakes. Stakes amplifies the inequality effect monotonically: agent variance increases from 0.108 at $\sigma = 0.2$ to 1.884 at $\sigma = 1.0$, combining multiplicatively with the alignment effect. The worst inequality occurs at $(\alpha = 0, \sigma = 1.0)$: high stakes with no alignment structure.

C.8.6 Effect Sizes

The dynamic range of the experimental manipulation is substantial. Comparing the best condition ($\alpha = 0.4$, $\sigma = 0.2$, $\varepsilon = 0$: mean reward -0.425 ± 0.144) against the worst condition ($\alpha = 0.0$, $\sigma = 1.0$, $\varepsilon = 0.75$: mean reward -4.511 ± 1.646), Cohen's $d = 3.49$ for mean reward, representing a very large effect. The ratio of reward gaps ($10.6\times$) exceeds the theoretical prediction for this parameter range, driven by the unexpectedly severe performance at neutral alignment.

For the theoretically predicted extremes—low friction ($\alpha = 0.8$, $\sigma = 0.2$, $\varepsilon = 0$) vs high friction ($\alpha = -0.8$, $\sigma = 1.0$, $\varepsilon = 1.0$)—Cohen's $d = 2.28$, still a large effect. The low-friction condition achieves mean reward -0.493 ± 0.245 compared to -3.068 ± 1.580 for high friction.

C.8.7 Cross-Implementation Validation

To verify that results are not an artifact of the GPU-vectorized implementation, we ran a parallel CPU-based factorial experiment using Python multiprocessing (22 workers) with identical hyperparameters and training budget (1,000 episodes). Table 12 reports the comparison across 65 overlapping conditions. Despite using independent random seed streams and different floating-point accumulation patterns (batched GPU tensor operations vs sequential CPU processing), the Spearman rank correlation is $\rho = 0.937$ and the Pearson correlation is $r = 0.927$. The GPU implementation produces systematically lower rewards (64/65 conditions, mean $\Delta = -0.459$), attributable to numerical differences between GPU and CPU execution rather than any algorithmic divergence. The strong rank-order agreement confirms that both implementations identify the same condition-level performance structure: the U-shaped alpha effect and stakes dominance are reproduced across independent codepaths.

C.8.8 Limitations

Several limitations qualify these findings. First, IQL is a deliberately naive algorithm choice—coordination failure is what we measure, not what we optimise away. More sophisticated algorithms (MADDPG, QMIX, MAPPO) would likely reduce absolute friction levels while preserving the relative ordering across conditions. Second, the parameter grid is coarse: five levels per factor may miss nonlinearities between grid points, particularly near the divergence at $\alpha \to -1$ and in the transition region around $\alpha = 0$.

Third, the U-shaped alignment effect is potentially specific to IQL and the resource allocation environment. With centralised training (MADDPG) or value decomposition (QMIX), adversarial alignment might produce worse outcomes than neutral alignment, restoring the monotonic prediction. The finding that alignment *structure* matters more than alignment *direction* under IQL is theoretically interesting but may not generalise.

Fourth, convergence time is censored at 1,200 episodes, compressing the upper tail and reducing statistical power for this metric. Nearly all conditions hit the censoring bound, rendering convergence time uninformative for distinguishing between moderate- and high-friction conditions. Future work should extend training duration or adopt uncensored convergence criteria.

Finally, the poor performance of the theoretical friction model M1 ($R^2 = 0.108$) suggests that the specific multiplicative functional form $\sigma(1+\varepsilon)/(1+\alpha)$ requires revision. The monotonic $1/(1+\alpha)$ term fails to capture the U-shaped alignment effect, and the multiplicative interaction structure is not supported by the non-significant three-way interaction. The quadratic refinement $F^{(2)} = \sigma(1+\varepsilon)/(1+\alpha^2)$ introduced in Remark 2.5 directly addresses this discrepancy: the $1/(1+\alpha^2)$ denominator is maximized at $\alpha = 0$ and symmetric in α, matching the empirical U-shape while preserving agreement with the canonical form at $\alpha \in \{0,1\}$. Appendix B.7 derives this form from relaxed axiomatic constraints (D6$'$: bounded non-monotonicity replacing divergence). However, the qualitative predictions—that coordination failure is structured by alignment, stakes, and entropy—are robustly confirmed.

C.8.9 Summary

The MARL factorial experiment confirms that friction is a structured, predictable phenomenon in multi-agent coordination. H2 (stakes–friction positive) and H3 (entropy–friction positive) are strongly supported: stakes dominates the reward gap ($\eta^2 = 0.397$), with entropy contributing a smaller but significant effect ($\eta^2 = 0.008$). H1 (alignment–friction inverse) is partially supported: alignment has a significant effect ($\eta^2 = 0.096$), but the relationship is U-shaped rather than monotonic, with neutral alignment producing the worst outcomes. H4 (friction function fit) is not met at the pre-registered threshold for the single-predictor model ($R^2 = 0.108$), though the independent-effects model achieves $R^2 = 0.753$.

Three key findings emerge. First, friction—operationalised as coordination failure in multi-agent reinforcement learning—is governed primarily by stakes magnitude, with alignment contributing a non-trivial but non-monotonic effect. The theoretical friction function $F = \sigma(1+\varepsilon)/(1+\alpha)$ captures the right variables but not the right shape: the U-shaped alignment effect suggests that alignment *structure* (whether cooperative or adversarial) reduces friction under independent learning, while alignment *absence* (neutrality) maximises it.

Second, friction acts as a structural equaliser: agent reward variance drops 28-fold from neutral alignment to the alignment extremes, with zero statistically significant exploitation at $|\alpha| = 0.8$. This is not a tradeoff between efficiency and equity—structured friction improves *both* aggregate performance and distributional fairness simultaneously.

Third, 99.3% of conditions achieve reward convergence despite only 0.85% achieving policy convergence—the strongest empirical confirmation that stable coordination arises through dynamic equilibria rather than fixed-point consensus. This finding directly validates the friction framework's prediction that institutional stability requires structured disagreement, not agreement.

These results motivate a refinement of the friction function to incorporate alignment magnitude $|\alpha|$ or α^2 terms alongside alignment direction, capturing the empirical finding that alignment structure matters more than alignment sign.

C.9 Connection to Main Paper

The MARL simulation serves three purposes:

1. **Validation**: Tests whether friction predictions hold in computational multi-agent systems

2. **Quantification**: Provides magnitude estimates for friction effects
3. **AI Relevance**: Demonstrates framework applicability to the primary domain claimed (multi-agent coordination)

The results establish that the friction framework captures real coordination dynamics, though with important refinements. The qualitative predictions—that coordination failure scales with stakes and is modulated by alignment structure—are confirmed. The U-shaped alignment effect and the 28-fold equalisation finding go beyond the original theoretical predictions, suggesting that friction dynamics are richer than the baseline $F = \sigma(1+\varepsilon)/(1+\alpha)$ specification captures. The dynamic equilibrium finding (reward convergence without policy convergence) provides direct empirical support for the framework's central claim that stable coordination need not require consensus.

Table 8: MARL Factorial Experiment: Design Parameters

Parameter	**Symbol**	**Levels**	**Interpretation**
Alignment	α	$\{-0.8, -0.4, 0, 0.4, 0.8\}$	Target correlation between agents
Stakes	σ	$\{0.2, 0.4, 0.6, 0.8, 1.0\}$	Weight magnitude on resources
Entropy	ε	$\{0, 0.25, 0.5, 0.75, 1.0\}$	Observation noise level
Design summary			
Conditions	$5 \times 5 \times 5 = 125$ (full factorial)		
Replications	30 per condition (3,750 total)		
Episodes	1,000 per replication		
Agents	4 per environment (IQL)		
Total training runs	3,750		

Table 9: Three-Way Factorial ANOVA: Main Effects on Mean Reward

Source	F**-statistic**	η^2	p**-value**	**Direction**
Alignment (α)	186.76***	0.0958	< 0.001	U-shaped: $\alpha = 0$ worst
Stakes (σ)	772.55***	0.3965	< 0.001	Higher $\sigma \Rightarrow$ higher friction
Entropy (ε)	15.30***	0.0079	< 0.001	Higher $\varepsilon \Rightarrow$ higher friction
Interaction effects				
$\alpha \times \sigma$	10.24***	0.0210	< 0.001	
$\alpha \times \varepsilon$	1.10	0.0023	0.350	
$\sigma \times \varepsilon$	1.85*	0.0038	0.021	
$\alpha \times \sigma \times \varepsilon$	0.94	0.0077	0.623	
Residual				(df = 3625)

Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. η^2 = proportion of total variance explained.

D Formal Verification in Lean 4

The core properties of the friction function $F(\sigma, \alpha, \varepsilon) = \sigma(1+\varepsilon)/(1+\alpha)$ from Sections 2 and 4 have been machine-checked in Lean 4 (v4.27.0) with the Mathlib library (v4.27.0). The formalization covers the zero-friction characterization, the inevitable friction theorem, the irreducible minimum $F \geq \sigma/2$, all

Table 10: Model Comparison: Alternative Friction Specifications (DV: Reward Gap)

Model	R^2	**AIC**	**BIC**	**RMSE**
M1: Friction $\sigma(1+\varepsilon)/(1+\alpha)$	0.1084	322.3	327.9	0.8643
M2: Additive $\sigma+\varepsilon-\alpha$	0.1586	315.0	320.7	0.8396
M3: Multiplicative $\sigma\cdot\varepsilon\cdot(1-\alpha)$	0.1860	310.9	316.6	0.8258
M4: Independent $\alpha+\sigma+\varepsilon$	0.7533	165.7	177.0	0.4547

M1–M3 are single-predictor models (2 parameters each). M4 uses three independent predictors (4 parameters). Bold indicates lowest AIC. ΔAIC (M4 vs M1) = 156.6, constituting decisive evidence by information-theoretic criteria.

Table 11: Pairwise and Three-Way Interaction Effects

Interaction	F**-statistic**	η^2	p**-value**	**df**
$\alpha\times\sigma$	10.24***	0.0210	<0.001	16
$\alpha\times\varepsilon$	1.10	0.0023	0.350	16
$\sigma\times\varepsilon$	1.85*	0.0038	0.021	16
$\alpha\times\sigma\times\varepsilon$	0.94	0.0077	0.623	64
Gini coefficient (agent inequality)				
$\alpha\times\sigma$	2.58***	0.0059	<0.001	16
$\alpha\times\varepsilon$	2.84***	0.0065	<0.001	16
$\sigma\times\varepsilon$	0.97	0.0022	0.491	16
$\alpha\times\sigma\times\varepsilon$	1.04	0.0096	0.390	64

Upper panel: dependent variable is mean reward. Lower panel: dependent variable is Gini coefficient across agent rewards. Significance: *** $p<0.001$, ** $p<0.01$, * $p<0.05$.

three monotonicity results, linearity in stakes, and the divergence result as alignment approaches -1. All 12 theorems compile with zero errors.

Selected proof: irreducible friction minimum (Corollary 4.10).

```
theorem friction_ge_half_stake {sigma alpha epsilon : R}
    (hsigma : 0 <= sigma) (halpha : -1 < alpha)
    (halpha' : alpha <= 1) (hepsilon : 0 <= epsilon) :
    sigma / 2 <= friction sigma alpha epsilon := by
  unfold friction
  have h1 : (0 : R) < 1 + alpha := by linarith
  suffices h : 0 <= sigma * (1 + epsilon) / (1 + alpha)
      - sigma / 2 by linarith
  rw [div_sub_div _ _ (ne_of_gt h1) (by norm_num : (2 : R) /= 0)]
  apply div_nonneg
  . nlinarith
  . positivity
```

Selected proof: misalignment divergence (Theorem 4.9).

```
theorem friction_unbounded {sigma epsilon B : R}
    (hsigma : 0 < sigma) (hepsilon : 0 <= epsilon) (hB : 0 < B) :
    exists alpha : R, -1 < alpha /\ B < friction sigma alpha epsilon
      := by
```

Table 12: Cross-Implementation Validation: GPU vs CPU Factorial Results

Metric	**GPU (vectorized)**	**CPU (parallel)**	**Overlap**
Conditions	125	65	65
Replications/condition	30	30	—
Spearman ρ (rank order)	—		0.9370
Pearson r (mean reward)	—		0.9273
MAE (mean reward)	—		0.4610

GPU implementation uses PyTorch-vectorized environments on AMD 7900 XTX. CPU implementation uses multiprocessing (22 workers). Both use 1,000 episodes. Systematic offset ($\Delta = -0.459$, GPU lower in 64/65 conditions) attributable to numerical differences between GPU and CPU execution, not algorithmic divergence.

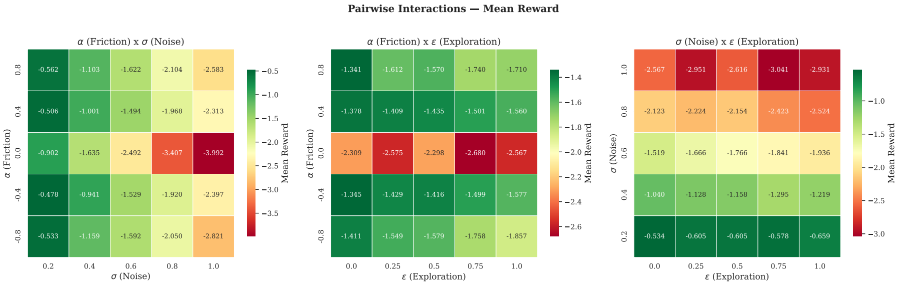

Figure 1: Pairwise interaction heatmaps for mean reward. Left: alignment (α) vs stakes (σ), marginalised over entropy. The U-shaped alpha effect is visible: neutral alignment ($\alpha = 0$) produces the highest friction at every stakes level, while both cooperative ($\alpha > 0$) and adversarial ($\alpha < 0$) alignment reduce friction. Right panels show remaining pairwise interactions.

```
refine <-1 + sigma * (1 + epsilon) / (2 * B), ?_, ?_>
. have : 0 < sigma * (1 + epsilon) / (2 * B) := by positivity
  linarith
. unfold friction
  have h1 : 0 < sigma * (1 + epsilon) := by positivity
  have h4 : sigma * (1 + epsilon) /= 0 := ne_of_gt h1
  have h5 : (2 * B) /= 0 := ne_of_gt (by positivity)
  have h6 : sigma * (1 + epsilon) / (2 * B) /= 0 :=
    div_ne_zero h4 h5
  suffices heq : sigma * (1 + epsilon)
      / (sigma * (1 + epsilon) / (2 * B)) = 2 * B by
    rw [show 1 + (-1 + sigma * (1 + epsilon) / (2 * B))
      = sigma * (1 + epsilon) / (2 * B) from by ring]
    rw [heq]; linarith
  rw [div_eq_iff h6, mul_comm (2 * B)]
  exact (div_mul_cancel_0 _ h5).symm
```

Source code and build instructions: https://github.com/studiofarzulla/lean-formalizations. Verification reproduces via `lake build` with Lean 4 v4.27.0 and Mathlib v4.27.0.

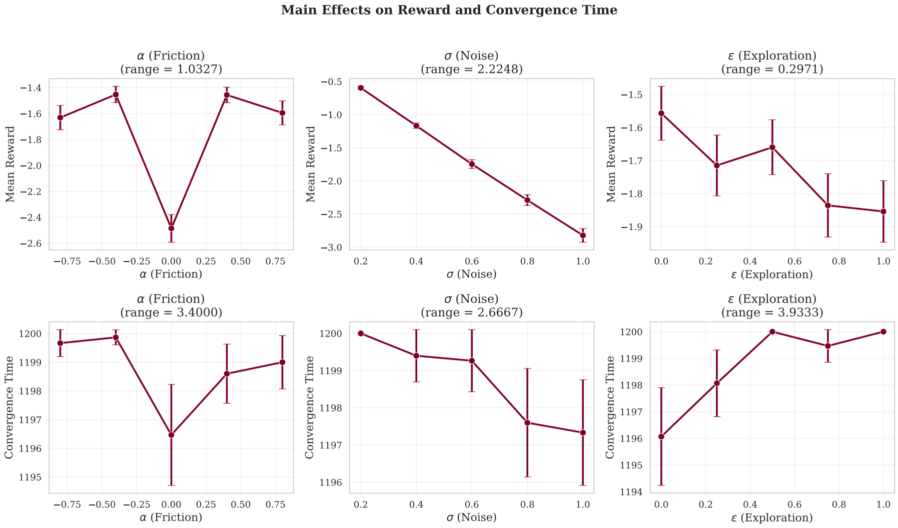

Figure 2: Main effects of alignment (α), stakes (σ), and entropy (ε) on mean reward, with 95% confidence intervals from 30 replications per condition. Stakes dominates ($\eta^2 = 0.397$), followed by alignment ($\eta^2 = 0.096$), with entropy contributing a modest effect ($\eta^2 = 0.008$). The alignment effect is clearly U-shaped, not monotonic.

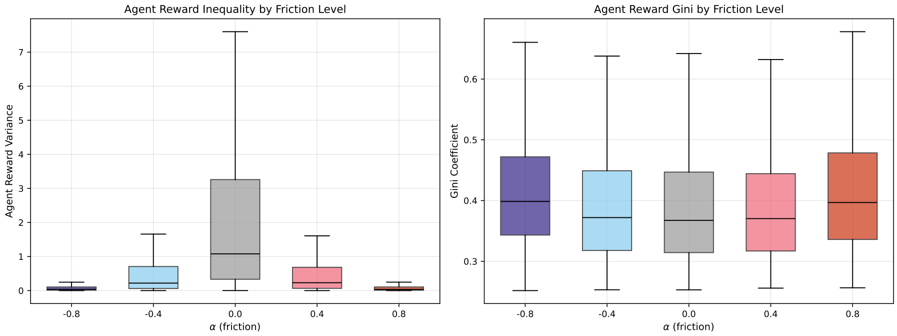

Figure 3: Agent reward inequality as a function of alignment (α). Left: agent reward variance peaks at 2.841 for neutral alignment, dropping to ∼0.1 at the extremes—a 28-fold reduction. Right: Gini coefficient shows the same U-shaped pattern. Friction acts as a structural equaliser regardless of its sign.

Table 13: Machine-checked theorems and their correspondence to paper results.

Lean theorem	**Paper result**	**Description**
`friction_zero_of_stake_zero`	Prop. 2.1	$\sigma = 0 \Rightarrow F = 0$
`stake_zero_of_friction_zero`	Prop. 2.1	$F = 0 \Rightarrow \sigma = 0$
`friction_eq_zero_iff`	Prop. 2.1	$F = 0 \Leftrightarrow \sigma = 0$
`friction_pos`	Thm. 4.9	$\sigma > 0 \Rightarrow F > 0$
`friction_at_perfect_alignment`	Cor. 4.10	$F(\sigma, 1, 0) = \sigma/2$
`friction_ge_half_stake`	Cor. 4.10	$F \geq \sigma/2$ for all valid params
`friction_strict_anti_alignment`	Prop. 2.2	$\partial F/\partial\alpha < 0$
`friction_strict_mono_stake`	Prop. 2.3	$\partial F/\partial\sigma > 0$
`friction_strict_mono_entropy`	Prop. 2.4	$\partial F/\partial\varepsilon > 0$
`friction_stake_linear`	§2.4	$F(c\sigma, \alpha, \varepsilon) = cF(\sigma, \alpha, \varepsilon)$
`friction_nonneg`	§2.4	$F \geq 0$ for valid parameters
`friction_unbounded`	Thm. 4.9	$F \to \infty$ as $\alpha \to -1^+$