# The Axiom of Consent:

## Friction Dynamics in Multi-Agent Coordination

**Murad Farzulla**[1,2]

[1]*King's College London*
[2]*Dissensus AI*

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

January 10, 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 $\alpha$, stake magnitude $\sigma$, and communication entropy $\varepsilon$. 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

# 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. 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 $5.7\times$ baseline levels [25]. 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. 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 $\sigma$, increases with information loss $\varepsilon$, and decreases with alignment $\alpha$. 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) [27] governs evolutionary selection on coordination configurations:

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

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 [38, 52, 55] provides powerful tools for incentive-compatible allocation. Recent work on commitment-enhanced communication [4] 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 [71], the Price equation [61], and evolutionary stability concepts [53] provide powerful tools for analyzing strategic interaction under selection. Vanderschraaf's work on inductive deliberation [74] shows how correlated equilibria emerge from rational learning, while Golman and Page [30] 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 [13, 75]. Organizational research reveals the dual challenge of search and coordination: hierarchical influence structures outperform flat teams in environments requiring rapid convergence [42], while network centralization improves collective adaptability when two-way information flow is preserved [8]. 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 $\alpha$ 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 [48]: 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 $d \in \mathcal{D}$ affecting agent $i \in \mathcal{A}$ with stake $s_i(d) > 0$, the **legitimacy** of the decision is determined by:

$$\text{Legitimate}(d) \iff \sum_{i \in S_d} s_i(d) \cdot \mathbb{1}[\text{Consent}_i(d)] \geq \theta \cdot \sum_{i \in S_d} s_i(d) \tag{3}$$

where $S_d = \{i \in \mathcal{A} : s_i(d) > 0\}$ is the affected set, $\mathbb{1}[\text{Consent}_i(d)]$ 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 [7, 11, 49].

**Threshold structure.** The threshold $\theta$ 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. The framework does not specify $\theta$; 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 [2, 31]: who counts in aggregation? Our answer: those with nonzero stakes in the domain.

The binary consent indicator $\mathbb{1}[\text{Consent}_i(d)]$ 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 $T_i : \mathcal{O} \to \mathbb{R}$ 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

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

**Stake magnitude ($\sigma$)** 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 ($\varepsilon$)** 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 ($\alpha$)** 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.

## 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 $\sigma(d) = 0$. $\square$ $\square$

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 $\alpha > -1$. $\square$ $\square$

**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. $\square$ $\square$

**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 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.

### 2.5.1 Relationship to Scanlon's Contractualism

T.M. Scanlon's contractualism [66] 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 $\theta$ 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.

### 2.5.2 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 [31]: 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.

### 2.5.3 The Affected Interests Principle

Robert Goodin's "all-affected interests" principle [31] 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.

### 2.5.4 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 $\theta$.

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 $\theta$; it analyzes consequences given various $\theta$ 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.

## 2.6 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.

### 2.6.1 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.

### 2.6.2 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 (exit, voice, loyalty erosion) 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.

### 2.6.3 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.

### 2.6.4 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.

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).

Table 1: The Kernel Triple Across Scales

| Component | Symbol | Interpretation | Empirical Proxies |
|---|---|---|---|
| Alignment | $\alpha$ | Correlation of target functions | Survey congruence, revealed preference |
| Stake | $\sigma$ | Magnitude of optimization at risk | Tax burden, employment dependence |
| Entropy | $\varepsilon$ | 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 |

# 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 $(\rho_S, w_S, M_S)$ [36, 56]. 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** $\mathcal{O}_S$—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** $\text{Atom}_S$ 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.

## 3.3 The Triple $(\rho_S, w_S, M_S)$

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

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

$$\rho_S(\tau, G, p) \in [0,1] \tag{7}$$

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:

$$w_S(\tau) \geq 0 \tag{8}$$

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:

$$M_S(\tau' \to \tau) = \Pr(\text{type } \tau \text{ produced from } \tau') \tag{9}$$

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

The kernel triple $(\rho_S, w_S, M_S)$ 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 | $\text{Atom}_S$ | $\rho_S$ (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/$(1 + \text{Friction})$ | Reform, evolution |
| Cultural | Belief/practice | Transmission $\times$ retention | Copying error, reinterpretation |

## 3.4 The ROM Update Equation

The kernel triple governs temporal evolution through the *replicator-optimization mechanism* (ROM) equation [27].

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

$$\frac{dp_t(\tau)}{dt} = \sum_{\tau' \in T_S} p_t(\tau') \cdot w_S(\tau') \cdot \rho_S(\tau', G_{S,t}, p_t) \cdot M_S(\tau' \to \tau) - p_t(\tau) \cdot \bar{\phi}_t \qquad (10)$$

*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 [33, 58]. 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 $\tau'$ at time $t$ (population share)
- $w_S(\tau')$: Intrinsic weight of $\tau'$ (resource advantage)
- $\rho_S(\tau', G, p)$: Survival probability of $\tau'$ given network and population state
- $M_S(\tau' \to \tau)$: Probability that $\tau'$ produces $\tau$ in transmission
- $\bar{\phi}_t$: Mean fitness, ensuring normalization

The first term sums over all types that could produce $\tau$ via transmission, weighted by prevalence, intrinsic weight, survival, and transmission probability. The second term removes the current proportion of $\tau$ 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 $(\rho, w, M)$

The axiom of consent defines friction in terms of alignment ($\alpha$), stake ($\sigma$), and entropy ($\varepsilon$). 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**:

$$\rho_S^{\text{consent}}(\tau, G, p) = \frac{L(\tau)}{1 + F(\tau)} \qquad (11)$$

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

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:

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.

| Axiom Term | Kernel Term | Interpretation |
|---|---|---|
| $\alpha$ (alignment) | Encoded in $\rho_S$ | Higher alignment $\rightarrow$ higher survival; alignment enters friction function inversely, thus survival positively |
| $\sigma$ (stake) | $w_S$ (weight) | Stakes determine influence on dynamics; higher stakes $\rightarrow$ larger weight in evolutionary pressure |
| $\varepsilon$ (entropy) | $M_S$ (mutation) | Information loss as type transition probability; high entropy $\rightarrow$ increased transmission noise |

### 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) $\tau$ representing a governance arrangement:

$$w_S(\tau) = \sum_{i \in \text{Stakeholders}(\tau)} s_i \cdot \mathbf{1}[i \text{ supports persistence of } \tau] \tag{12}$$

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 $\varepsilon$), the policies they implement diverge from intended outcomes. This divergence is structurally equivalent to mutation in transmission.

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

$$M_S(\tau' \rightarrow \tau) = M_0(\tau' \rightarrow \tau) \cdot \left(1 + \lambda \cdot \bar{\varepsilon}(\tau')\right) \tag{13}$$

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

After modulation, row-stochasticity must be restored:

$$M_S(\tau' \rightarrow \tau) \leftarrow \frac{M_S(\tau' \rightarrow \tau)}{\sum_{\tau''} M_S(\tau' \rightarrow \tau'')} \tag{14}$$

### 3.5.3 Alignment as Survival Modulator

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

$$F(\tau) = \sum_{i \in A} s_i \cdot \frac{1 + \varepsilon_i(\tau)}{1 + \alpha_i(\tau)} \tag{15}$$

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:

$$\rho_S^{\text{consent}}(\tau) = \rho_S^{\text{base}}(\tau) \cdot \exp\left(-\lambda \cdot F(\tau)\right) \tag{16}$$

where $\rho_S^{\text{base}}$ 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 The Unification: Isomorphism of Structure

The central claim of this section is now statable precisely.

**Theorem 3.2** (Axiom-Kernel Correspondence). *The axiom of consent's kernel triple $(\alpha, \sigma, \varepsilon)$ and the ROM kernel triple $(\rho_S, w_S, M_S)$ are structurally isomorphic under the consent-friction instantiation. Specifically:*

(i) **Alignment-Survival:** $\alpha$ *enters $\rho_S$ 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:** $\sigma$ *determines $w_S$; aggregate stakes of supporting stakeholders constitute the type's evolutionary weight.*

(iii) **Entropy-Mutation:** $\varepsilon$ *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.

## 3.7 Properties of the Consent-Kernel Dynamics

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

**Proposition 3.3** (Consent-Aligned Survival Advantage). *Under the consent-friction instantiation, types with higher consent alignment exhibit higher survival probability:*

$$\alpha(\tau_1) > \alpha(\tau_2) \implies \rho_S(\tau_1) > \rho_S(\tau_2) \tag{17}$$

*holding stakes, entropy, and legitimacy constant.*

*Proof.* From Eq. 6, friction $F \propto (1 + \varepsilon)/(1 + \alpha)$. For fixed $\varepsilon$, increasing $\alpha$ decreases $F$. From Eq. 11, $\rho_S = L/(1 + F)$. Decreasing $F$ (with $L$ held constant) increases $\rho_S$. □ □

**Proposition 3.4** (Friction as Selection Pressure). *Friction acts as negative selection pressure on type prevalence. Types generating high friction decrease in frequency, ceteris paribus:*

$$\frac{\partial}{\partial F(\tau)}\left[\frac{dp_t(\tau)}{dt}\right] < 0 \tag{18}$$

*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. $\qquad\square$ $\qquad\square$

**Proposition 3.5** (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 [58] 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 $\tau$. Since consent-respecting types have higher $\rho_S$, they receive higher mass at equilibrium. Full proof requires specifying regularity conditions on $L$, $F$, and $M$; see Farzulla [27] for technical details. $\qquad\square$ $\qquad\square$

## 3.8 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 3.8** (Ownership Accumulation). The ownership-perception $O_A(d,t) \in [0,1]$ of agent $A$ over domain $d$ evolves as:

$$\frac{dO_A}{dt} = \beta \cdot (1 - O_A) \cdot \mathbf{1}[A \text{ holds consent for } d] \tag{19}$$

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

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

**Definition 3.9** (Ownership-Modulated Mutation). Ownership perception modulates the mutation kernel:

$$M_S(\tau' \to \tau) = M_0(\tau' \to \tau) \cdot \exp\left(-\gamma\left(\bar{O}(\tau') - \bar{O}(\tau)\right)\right) \tag{20}$$

where $\bar{O}(\tau)$ is average ownership-perception in configuration $\tau$ 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.

## 3.9 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 3.10** (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. $\pi$ 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 3.6** (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* $T' \in T_{S'}$*:*

$$\sum_{\tau_j : \pi(\tau_j) = T'} M_S(\tau_i \to \tau_j) = \sum_{\tau_l : \pi(\tau_l) = T'} M_S(\tau_k \to \tau_l) \tag{21}$$

(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 [76] and explains why "emergence" appears mysterious: it is what coarse-graining looks like when lumpability conditions fail.

## 3.10 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 $(\rho_S, w_S, M_S)$ provides domain-general apparatus for selection-transmission dynamics at any scale.

2. **Consent instantiation:** The axiom's $(\alpha, \sigma, \varepsilon)$ maps structurally onto $(\rho_S, w_S, M_S)$ 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* [27].

# 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 [71] governs frequency dynamics in populations under selection:

$$\frac{dp_t(\tau)}{dt} = p_t(\tau) \left[ \pi(\tau, p_t) - \bar{\pi}(p_t) \right] \tag{22}$$

where $\pi(\tau, p_t)$ is the payoff to type $\tau$ 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 [33, 54, 58, 72] incorporates imperfect transmission:

$$\frac{dp_t(\tau)}{dt} = \sum_{\tau'} p_t(\tau') \cdot \pi(\tau', p_t) \cdot M(\tau' \to \tau) - p_t(\tau) \cdot \bar{\pi}(p_t) \tag{23}$$

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

The ROM equation [27] generalizes this framework through three innovations:

1. **Decomposed fitness:** Rather than a monolithic payoff $\pi$, 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:

$$\boxed{\frac{dp_t(\tau)}{dt} = \sum_{\tau' \in T_S} p_t(\tau') \cdot w_S(\tau') \cdot \rho_S(\tau', G_{S,t}, p_t) \cdot M_S(\tau' \to \tau) - p_t(\tau) \cdot \bar{\phi}_t} \quad (24)$$

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 $\pi$ 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:*

$$w_S(\tau) = \sum_{i \in Stakeholders(\tau)} s_i \cdot \mathbf{1}[i \; supports \; \tau] \quad (25)$$

$$\rho_S(\tau, G, p) = \frac{L(\tau)}{1 + F(\tau)} \quad (26)$$

$$M_S(\tau' \to \tau) \propto M_0(\tau' \to \tau) \cdot (1 + \lambda \bar{\varepsilon}(\tau')) \quad (27)$$

*the ROM equation becomes:*

$$\frac{dp_t(\tau)}{dt} = \sum_{\tau'} p_t(\tau') \cdot \sigma(\tau') \cdot \frac{L(\tau')}{1 + F(\tau')} \cdot M_S(\tau' \to \tau) - p_t(\tau) \cdot \bar{\phi}_t \quad (28)$$

*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 $\sigma$ from supporting stakeholders achieve higher weight. Together, these determine which configurations persist.

## 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 in domain $d$ at time $t$ is:

$$F(d, t) = \sigma(d) \cdot \frac{1 + \varepsilon(d, t)}{1 + \alpha(d, t)} \quad (29)$$

Differentiating with respect to time:

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

$$\frac{dF}{dt} = \frac{\partial F}{\partial \sigma} \frac{d\sigma}{dt} + \frac{\partial F}{\partial \alpha} \frac{d\alpha}{dt} + \frac{\partial F}{\partial \varepsilon} \frac{d\varepsilon}{dt} \quad (30)$$

*Substituting partial derivatives from Propositions 2.2–2.4:*

$$\boxed{\frac{dF}{dt} = \frac{1 + \varepsilon}{1 + \alpha} \frac{d\sigma}{dt} - \frac{\sigma(1 + \varepsilon)}{(1 + \alpha)^2} \frac{d\alpha}{dt} + \frac{\sigma}{1 + \alpha} \frac{d\varepsilon}{dt}} \quad (31)$$

*Proof.* Direct application of the chain rule to Eq. 29:

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

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

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

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 $\varepsilon$) 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:*

$$\frac{1 + \varepsilon^*}{1 + \alpha^*} \frac{d\sigma}{dt} = \frac{\sigma^*(1 + \varepsilon^*)}{(1 + \alpha^*)^2} \frac{d\alpha}{dt} - \frac{\sigma^*}{1 + \alpha^*} \frac{d\varepsilon}{dt} \tag{32}$$

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

**Theorem 4.4** (Lyapunov Stability)**.** *Consider friction $F$ as a candidate Lyapunov function for the consent-holding dynamics. If:*

*(i) Stakes are bounded: $\sigma(t) \leq \sigma_{\max} < \infty$*

*(ii) Alignment improves monotonically: $d\alpha/dt \geq 0$*

*(iii) Entropy is non-increasing: $d\varepsilon/dt \leq 0$*

*then $F(t)$ is non-increasing and the system converges to a friction-minimizing configuration.*

*Proof.* Under conditions (i)–(iii), examine Eq. 31. The first term is bounded by $\sigma_{\max}(1 + \varepsilon)/(1 + \alpha) \cdot d\sigma/dt$. The second term is non-positive since $-\sigma(1 + \varepsilon)/(1 + \alpha)^2 < 0$ and $d\alpha/dt \geq 0$. The third term is non-positive since $\sigma/(1 + \alpha) > 0$ and $d\varepsilon/dt \leq 0$.

If stake dynamics are neutral ($d\sigma/dt = 0$), then $dF/dt \leq 0$ strictly. Since $F \geq 0$ is bounded below and monotonically non-increasing, $F$ converges to a limit. By LaSalle's invariance principle, the system converges to the largest invariant set where $dF/dt = 0$. □ □

This theorem provides conditions under which consent-respecting configurations emerge as attractors: when institutional design improves alignment and transparency while constraining stake concentration, 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 [23], legitimacy quantifies the degree to which effective voice tracks stakes:

**Definition 4.3** (Legitimacy)**.** The **legitimacy** of a consent-holding configuration in domain $d$ at time $t$ is:

$$L(d,t) = \frac{\sum_{i \in S_d} s_i(d) \cdot \text{eff\_voice}_i(d,t)}{\sum_{i \in S_d} s_i(d)} \tag{33}$$

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

$$\rho_S(\tau) = \frac{L(\tau)}{1 + F(\tau)} \tag{34}$$

*where $\rho_S(\tau)$ is survival probability for configuration $\tau$. 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 4.6** (Legitimacy Evolution)**.** *Under the consent-friction dynamics, legitimacy evolves according to:*

$$\frac{dL}{dt} = \underbrace{\sum_{i \in S_d} \frac{s_i}{\sum_j s_j} \frac{d(\text{eff\_voice}_i)}{dt}}_{\text{voice dynamics}} + \underbrace{\sum_{i \in S_d} \frac{\text{eff\_voice}_i - L}{\sum_j s_j} \frac{ds_i}{dt}}_{\text{stake reweighting}} \tag{35}$$

*Proof.* Apply the quotient rule to Eq. 33. Let $N = \sum_i s_i \cdot \text{eff\_voice}_i$ and $D = \sum_i s_i$. Then:

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

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 4.7** (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 $\pi$ determined $a$ from the space of possible actions $\mathcal{A}_d$. This procedure $\pi$ 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 $\pi$ 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 4.8** (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 4.9** (Irreducible Friction). *The minimal achievable friction in domain d with positive stakes is:*

$$F_{\min}(d) = \frac{\sigma(d)}{2} \tag{36}$$

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

*Proof.* Substitute $\alpha = 1, \varepsilon = 0$ into Eq. 29:

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

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

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* 4.4 (Connection to Arrow's Impossibility). Arrow's theorem [3] demonstrates that no social welfare function satisfies minimal fairness axioms without dictatorship. Our Theorem 4.8 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 4.10** (ROM Convergence). *Under the ROM dynamics (Eq. 24) 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 $\tau, G, p$*

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

$$p^*(\tau) \propto \sum_{\tau'} p^*(\tau') \cdot w_S(\tau') \cdot \rho_S(\tau') \cdot M_S(\tau' \to \tau) \tag{37}$$

*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 $\tau_1, \tau_2$ 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. $\quad\square\quad\quad\square$

**Corollary 4.11** (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 3.9 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 4.5** (Exact Lumpability). A partition $\mathcal{P} = \{T_1, \ldots, T_k\}$ of type space $T_S$ is **exactly lumpable** for the ROM dynamics if for all $T_i, T_j \in \mathcal{P}$ and all $\tau, \tau' \in T_i$:

$$\sum_{\tau'' \in T_j} w_S(\tau) \cdot \rho_S(\tau) \cdot M_S(\tau \to \tau'') = \sum_{\tau'' \in T_j} w_S(\tau') \cdot \rho_S(\tau') \cdot M_S(\tau' \to \tau'') \tag{38}$$

**Theorem 4.12** (Preservation of ROM Structure). *If partition $\mathcal{P}$ is exactly lumpable for the ROM dynamics at scale $S$, then the coarse-grained dynamics at scale $S'$ (induced by $\mathcal{P}$) also satisfy ROM with kernel triple $(\rho_{S'}, w_{S'}, M_{S'})$ where:*

$$\rho_{S'}(T_i) = \rho_S(\tau) \text{ for any } \tau \in T_i \text{ (well-defined by lumpability)} \tag{39}$$

$$w_{S'}(T_i) = \sum_{\tau \in T_i} w_S(\tau) \cdot p(\tau | T_i) \tag{40}$$

$$M_{S'}(T_i \to T_j) = \sum_{\tau \in T_i} \sum_{\tau' \in T_j} p(\tau | T_i) \cdot M_S(\tau \to \tau') \tag{41}$$

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

$$\frac{dP_t(T_i)}{dt} = \sum_{\tau \in T_i} \frac{dp_t(\tau)}{dt}$$

$$= \sum_{\tau \in T_i} \left[ \sum_{\tau' \in T_S} p_t(\tau') \cdot w_S(\tau') \cdot \rho_S(\tau') \cdot M_S(\tau' \to \tau) - p_t(\tau) \cdot \bar{\phi}_t \right]$$

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

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

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 $S'$. $\square$ $\square$

*Remark* 4.6 (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 [76]: 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 4.7** (Ownership Dynamics). The ownership-perception $O_A(d, t)$ of agent $A$ over domain $d$ evolves as:

$$\frac{dO_A}{dt} = \beta \cdot (1 - O_A) \cdot \mathbf{1}[A = H(d,t)] - \gamma \cdot O_A \cdot \mathbf{1}[A \neq H(d,t)] \tag{42}$$

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 ($\beta, \gamma$ may differ) captures the empirical observation that ownership perceptions develop slowly but erode slowly as well—the "sticky" nature of perceived authority.

**Theorem 4.13** (Tenure-Transition Relationship). *Let $\tau_H$ denote the tenure of current consent-holder $H$. The probability of regime transition decreases exponentially with tenure:*

$$\Pr(transition|\tau_H) \propto \exp(-\gamma \cdot O_H(\tau_H)) \tag{43}$$

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

*Proof.* Regime transitions occur when the mutation kernel permits type change. From Eq. 20, transitions away from high-ownership configurations are suppressed by factor $\exp(-\gamma(\bar{O}(\tau') - \bar{O}(\tau)))$.

For the current holder's configuration $\tau'$, $\bar{O}(\tau') = O_H(\tau_H)$. Transition to a new configuration $\tau$ 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. $\qquad\square$

$\square$

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.

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.

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 |

2. **Design:** Institutional interventions can target specific kernel components (improve $\alpha$, reduce $\varepsilon$, redistribute $\sigma$).
3. **Unification:** The same mathematics applies across political, economic, and computational domains.

## 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. 28).

2. **Friction Dynamics:** Temporal evolution of friction depends on stake, alignment, and entropy dynamics (Eq. 31).

3. **Legitimacy Evolution:** Legitimacy changes through voice dynamics and stake reweighting (Theorem 4.6).

4. **Key Theorems:**

   - Consent-Holding Necessity (Theorem 4.7): Consent-holding is unavoidable.
   - Inevitable Friction (Theorem 4.8): Zero friction is impossible with positive stakes.
   - ROM Convergence (Theorem 4.10): System converges to consent-respecting equilibria.
   - Tenure-Transition (Theorem 4.13): Regime transition probability decays with tenure.

5. **Scale Coherence:** ROM structure is preserved under lumpable coarse-graining (Theorem 4.12).

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* [27].

# 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 $C : \mathcal{R} \to \mathcal{A}$ maps resources to controlling agents.

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

$$\sigma_j(r) = \left| \frac{\partial U_j}{\partial \text{allocation}(r)} \right| \tag{44}$$

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:

$$\alpha_{ij}(r) = \frac{\text{cov}(\nabla_r U_i, \nabla_r U_j)}{\sqrt{\text{var}(\nabla_r U_i) \cdot \text{var}(\nabla_r U_j)}} \tag{45}$$

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:

$$\varepsilon_{ij} = H(U_j | \text{signal}_{i \to j}) \tag{46}$$

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:*

$$F_j(r) = \sigma_j(r) \cdot \frac{1 + \varepsilon_{ij}(r)}{1 + \alpha_{ij}(r)} \tag{47}$$

*Total system friction is:*

$$F_{system} = \sum_{r \in \mathcal{R}} \sum_{j \neq C(r)} F_j(r) \tag{48}$$

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:

$$L(C) = \frac{\sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{A}} \sigma_j(r) \cdot v_j(r)}{\sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{A}} \sigma_j(r)} \tag{49}$$

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. 28), the survival probability of configuration $C$ is:

$$\rho(C) = \frac{L(C)}{1 + F(C)} \tag{50}$$

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:

$$C^* = \arg \max_C \sum_j U_j(C) \tag{51}$$

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:*

$$C^{**} = \arg \min_C \left[ \sum_r \sum_{j \neq C(r)} \sigma_j(r) \cdot \frac{1 + \varepsilon_{C(r),j}}{1 + \alpha_{C(r),j}(r)} \right] \tag{52}$$

*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 [46, 47]. 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:

$$\alpha_{ij} = \frac{\mathbb{E}_{s,a}[R_i(s,a) \cdot R_j(s,a)] - \mathbb{E}[R_i]\mathbb{E}[R_j]}{\sigma_{R_i} \cdot \sigma_{R_j}} \tag{53}$$

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:*

$$\mathbb{E}[F] \propto \bar{\sigma} \cdot \frac{1 + \bar{\varepsilon}}{1 + \bar{\alpha}} \tag{54}$$

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

*Implication:* MARL convergence depends on reward alignment $\alpha$ and communication capacity (inverse $\varepsilon$). Systems with misaligned rewards or bandwidth-constrained communication generate friction exceeding convergence thresholds.

### 5.1.6 Application: AI Alignment

The AI alignment problem [9, 64] can be framed as consent-friction minimization between AI systems and human principals.

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

$$\alpha_{AH} = \mathrm{corr}(U_A, U_H) \tag{55}$$

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

$$\varepsilon_{AH} = H(U_A | \text{observables}_H) \tag{56}$$

Opaque AI systems have high $\varepsilon$; interpretable systems have low $\varepsilon$.

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

$$F_{AH} = \sigma_H \cdot \frac{1 + \varepsilon_{AH}}{1 + \alpha_{AH}} \tag{57}$$

*where $\sigma_H$ is the magnitude of human stakes in AI behavior.*

*Predictions:*

1. **Interpretability reduces friction.** Investment in interpretability (reducing $\varepsilon$) 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 [26]. AI systems with embodied autonomy possess the functional properties that make unconsented rule illegitimate [22].

### 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 $\tau \in \mathcal{T}$ index coordination strategies (consent-holding configurations). Under ROM dynamics:*

$$\frac{dp_t(\tau)}{dt} = \sum_{\tau'} p_t(\tau') \cdot \sigma(\tau') \cdot \frac{L(\tau')}{1 + F(\tau')} \cdot M(\tau' \to \tau) - p_t(\tau) \cdot \bar{\phi}_t \qquad (58)$$

*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 [1, 14, 17]. Governance structures are encoded in protocols; transactions are recorded on public blockchains; and friction manifests as price volatility.

**Key Results.** In Farzulla [25], I document that infrastructure disruption events generate $5.7\times$ 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) [28] 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.

### 5.2.2 Political Legitimacy

Political systems provide the original domain for consent analysis. The kernel triple operationalizes legitimacy as stakes-weighted voice [23].

**Key Mapping.**

- **Alignment** $\alpha$: Correlation between citizen preferences and policy outcomes
- **Stakes** $\sigma$: Magnitude of citizen welfare at stake in policy decisions
- **Entropy** $\varepsilon$: 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 ($\alpha$) | Reward correlation | Holder-protocol match | Citizen-policy match |
| Stake ($\sigma$) | Utility exposure | Token value | Affected interests |
| Entropy ($\varepsilon$) | 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:

$$F = \sigma \cdot \frac{1 + \varepsilon}{1 + \alpha} \tag{59}$$

$$L = \frac{\sum_i s_i \cdot v_i}{\sum_i s_i} \tag{60}$$

$$\frac{dp(\tau)}{dt} = \sum_{\tau'} p(\tau') \cdot \sigma(\tau') \cdot \frac{L(\tau')}{1 + F(\tau')} \cdot M(\tau' \to \tau) - p(\tau) \cdot \bar{\phi} \tag{61}$$

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 ($\alpha$), stakes ($\sigma$), and entropy ($\varepsilon$). 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 [63, 65]. 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 ($\alpha$)

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:

$$\alpha_i^{\text{survey}} = \frac{\mathbf{p}_i \cdot \mathbf{p}_H}{\|\mathbf{p}_i\| \cdot \|\mathbf{p}_H\|} \tag{62}$$

**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:

$$\alpha_i^{\text{CSES}}(t) = 1 - \frac{D(\mathbf{p}_i, \mathbf{p}_{G(t)})}{D_{\max}} \tag{63}$$

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 [44]. 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 $a_i \in \mathcal{A}$ be agent $i$'s observed actions and $a_H \in \mathcal{A}$ be the consent-holder's policy actions. Behavioral alignment is:

$$\alpha_i^{\text{behavioral}} = \text{corr}(U_i(a_H), U_i(a_i^*)) \tag{64}$$

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:

$$\alpha_i^{\text{market}}(P) = \text{sign}(\Delta h_i(P)) \cdot \left(1 - e^{-\lambda|\Delta h_i(P)|}\right) \tag{65}$$

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

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

$$\alpha_i^{\text{vote}} = \sum_{p \in \text{Parties}} v_{ip} \cdot \cos(\mathbf{p}_p, \mathbf{p}_G) \tag{66}$$

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:

$$\alpha(d, t) = \frac{\sum_{i \in S_d} s_i(d) \cdot \alpha_i(d, t)}{\sum_{i \in S_d} s_i(d)} \tag{67}$$

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.

## 6.3 Operationalizing Stakes ($\sigma$)

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:

$$\sigma_i^{\text{econ}}(d) = \mathbb{E}[|W_i(O) - W_i(O')|] \tag{68}$$

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:

$$\sigma_i^{\text{PV}}(d) = \sum_{t=0}^{\infty} \delta^t \cdot \sigma_i^{\text{econ}}(d, t) \tag{69}$$

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:

$$\sigma_i^{\text{pol}}(d) = \phi(\text{proximity}_i(d), \text{reversibility}(d), \text{magnitude}(d)) \tag{70}$$

where $\phi$ 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 [31], 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:

$$\sigma_i^{\text{comp}}(d) = c(d) \cdot \pi_i(d) \tag{71}$$

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:

$$c(d) = \max_{o \in \mathcal{O}(d)} L(o) - \min_{o \in \mathcal{O}(d)} L(o) \tag{72}$$

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

## 6.4 Operationalizing Entropy ($\varepsilon$)

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:

$$\varepsilon_i^{\text{info}} = H(\mathbf{p}_i | \hat{\mathbf{p}}_i^H) = -\sum_k p_{ik} \log \frac{p_{ik}}{\hat{p}_{ik}^H} \tag{73}$$

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:

$$\varepsilon_i^{\text{uncertainty}} = H(\mathbf{p}_i | \mathcal{D}_i) = \text{entropy of posterior} \tag{74}$$

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:

$$\varepsilon_{i,H}^{\text{channel}} = 1 - \frac{I(X_i; Y_H)}{H(X_i)} \tag{75}$$

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 $\varepsilon$ 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.

$$\varepsilon^{\text{misperception}} = \frac{1}{n} \sum_i \|\hat{\mathbf{p}}_i^H - \mathbf{p}_i\|^2 \tag{76}$$

**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.

## 6.5 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:

$$F^{\text{vol}}(d, t) = \frac{\sigma_{\text{realized}}(d, t)}{\sigma_{\text{baseline}}(d)} \tag{77}$$

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

The 5.7× volatility differential documented in Farzulla [25] 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:

$$h_t = \omega + \alpha_1 \epsilon_{t-1}^2 + \beta_1 h_{t-1} + \gamma \cdot F(d, t-1) \tag{78}$$

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:

$$F^{\text{inst}}(P, t) = w_1 \cdot \text{Protest}(P, t) + w_2 \cdot \text{Litigation}(P, t) + w_3 \cdot \text{Exit}(P, t) + w_4 \cdot \text{Noncompliance}(P, t) \tag{79}$$

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 $\mathcal{M}$ is:

$$F^{\text{coord}}(\mathcal{M}) = 1 - \frac{R_{\text{achieved}}}{R_{\text{optimal}}} \tag{80}$$

where $R_{\text{achieved}}$ is realized aggregate reward and $R_{\text{optimal}}$ 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:

$$F^{\text{overhead}}(\mathcal{M}) = \frac{C_{\text{coordination}}}{C_{\text{total}}} \tag{81}$$

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

The friction estimate is:

$$\hat{F} = \hat{\sigma} \cdot \frac{1 + \hat{\varepsilon}}{1 + \hat{\alpha}} \tag{82}$$

Taylor expansion around true values yields:

$$\hat{F} \approx F + \frac{\partial F}{\partial \sigma}\eta_\sigma + \frac{\partial F}{\partial \varepsilon}\eta_\varepsilon + \frac{\partial F}{\partial \alpha}\eta_\alpha + O(\eta^2) \tag{83}$$

Substituting partial derivatives:

$$\hat{F} \approx F + \frac{1 + \varepsilon}{1 + \alpha}\eta_\sigma + \frac{\sigma}{1 + \alpha}\eta_\varepsilon - \frac{\sigma(1 + \varepsilon)}{(1 + \alpha)^2}\eta_\alpha \tag{84}$$

**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. 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.

Table 5: Measurement Apparatus Summary

| Construct | Symbol | Primary Operationalizations |
|---|---|---|
| Alignment | $\alpha$ | Survey preference correlation, revealed preference from behavior, voting patterns |
| Stakes | $\sigma$ | Monetary exposure, affected interest magnitude, task criticality |
| Entropy | $\varepsilon$ | KL divergence, channel capacity, transparency indices, misperception scores |
| Friction | $F$ | Market volatility, institutional instability, coordination failure |

## 6.7 Summary

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

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.

## 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:

$$F_{\text{latent}}(d,t) = F(d,t) \cdot \exp\left(\int_0^t \kappa(d,s)\, ds\right) \tag{85}$$

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 85 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:*

$$\mathbb{E}[\textit{Transition Magnitude}] \propto F_{latent}(t_{transition}) \propto e^{\int_0^{t_{transition}} \kappa(s)\, ds} \tag{86}$$

This explains the empirical pattern: authoritarian regimes persist through suppression but experience catastrophic transitions (revolutions, state collapse) when suppression capacity degrades. 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.

### 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 [60] 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?

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* [45].

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

$$\psi(d,t) = 1 - \frac{\text{Var(expressed preference)}}{\text{Var(authentic preference)}} \tag{87}$$

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:

$$F_{\text{observed}}(d,t) = (1 - \psi(d,t)) \cdot F_{\text{total}}(d,t) \tag{88}$$

$$F_{\text{latent}}(d, t) = \psi(d, t) \cdot F_{\text{total}}(d, t) \tag{89}$$

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.

### 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 $\alpha$ 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 $\sigma$ 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 $\varepsilon$ 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 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?

### 7.2.1 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.

### 7.2.2 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.

### 7.2.3 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 [57].

## 7.3 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.

### 7.3.1 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.

### 7.3.2 Measurement Challenges

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

**Alignment ($\alpha$).** 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 ($\sigma$).** 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 ($\varepsilon$).** 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 [6, 19, 29]. 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.

### 7.3.3 Scale-Mixing Problems

The coarse-graining analysis (Section 3.9) 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.

## 7.4 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 [12, 16, 32, 67]. We position our framework relative to major alternatives, identifying points of contact and departure.

### 7.4.1 Mechanism Design

Mechanism design theory [38, 55] 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.

### 7.4.2 Social Choice Theory

Social choice theory [3, 69] 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).

### 7.4.3 Rawlsian Contractualism

Rawlsian contractualism [62] derives principles of justice from a hypothetical original position behind a veil of ignorance. It is the dominant framework in contemporary political philoso-

phy, joined by related accounts emphasizing democratic authority [15, 20], outcome-sensitive legitimacy [59], and critiques of actual consent theories [70].

**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).

### 7.4.4 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 [5, 10] to empirical legitimacy research [50, 68, 73] 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.

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

### 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 [35] and autonomous agent rules of engagement [18] demonstrates growing recognition that AI systems require consent-respecting governance structures. 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 [39, 41, 43]

- **Climate negotiations:** Intergenerational consent and stake representation

- **Deliberative systems:** Multi-stakeholder consent-holding in complex democratic arrangements [51]

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 [21] and regulatory arbitrage [24]. 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.

# Acknowledgments

# References

[1] Darcy W. E. Allen, Chris Berg, Sinclair Davidson, and Mikayla Novak. *Cryptodemocracy: How Blockchain Can Radically Expand Democratic Governance.* Palgrave Macmillan, Cham, 2020.

[2] Gustaf Arrhenius. The boundary problem in democratic theory. In Folke Tersman, editor, *Democracy Unbound: Basic Explorations I*, pages 14–29. Stockholm University, 2005.

[3] Kenneth J. Arrow. *Social Choice and Individual Values.* Wiley, New York, 1951.

[4] Ala Avoyan and João Ramos. A road to efficiency through communication and commitment. *American Economic Review*, 113(9):2344–2374, 2023. doi: 10.1257/aer.20171014.

[5] Haris Aziz, Anna Bogomolnaia, and Hervé Moulin. Proportional participatory budgeting with additive utilities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5275–5282, 2021.

[6] Hanna Bäck and Sara Kalm. Demonstrations and political regimes. In Donatella Della Porta and Mario Diani, editors, *The Oxford Handbook of Social Movements*, pages 664–678. Oxford University Press, 2015.

[7] Ludvig Beckman. Towards a principle of most-deeply affected. *Philosophy & Social Criticism*, 47(2):147–166, 2020. doi: 10.1177/0191453720917450.

[8] Ethan Bernstein, Jesse Shore, and David Lazer. Network centralization and collective adaptability to a shifting environment. *Organization Science*, 34(2):870–889, 2023. doi: 10.1287/orsc.2022.1602.

[9] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press, Oxford, 2014.

[10] Robert Bredereck, Piotr Faliszewski, Andrzej Kaczmarczyk, Dušan Knop, and Rolf Niedermeier. Proportionality for constrained public decisions. *arXiv preprint arXiv:2409.02609*, 2024.

[11] Harry Brighouse and Marc Fleurbaey. Democracy and proportionality. *Journal of Political Philosophy*, 18(2):137–155, 2010. doi: 10.1111/j.1467-9760.2008.00316.x.

[12] James M. Buchanan. *The Limits of Liberty: Between Anarchy and Leviathan.* University of Chicago Press, Chicago, 1975.

[13] Lucian Busoniu, Robert Babuska, 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.

[14] Vitalik Buterin. Notes on blockchain governance. Blog post, 2017.

[15] Thomas Christiano. *The Constitution of Equality: Democratic Authority and Its Limits.* Oxford University Press, Oxford, 2008.

[16] Robert A. Dahl. *Polyarchy: Participation and Opposition.* Yale University Press, New Haven, 1971.

[17] Primavera De Filippi and Aaron Wright. *Blockchain and the Law: The Rule of Code.* Harvard University Press, Cambridge, MA, 2018.

[18] John C. Doyle and Marie E. Csete. Rules of engagement. *Nature*, 446(7138):860, 2007. doi: 10.1038/446860a.

[19] David Easton. *A Systems Analysis of Political Life*. Wiley, New York, 1965.

[20] David M. Estlund. *Democratic Authority: A Philosophical Framework*. Princeton University Press, Princeton, 2008.

[21] Murad Farzulla. Asymptotic protection: Derivatives and systemic risk, 2025.

[22] Murad Farzulla. Embodied autonomy and moral consideration in AI systems, 2025. Working paper. Extends themes in [26].

[23] Murad Farzulla. The doctrine of consensual sovereignty: A formal framework for legitimate governance, 2025.

[24] Murad Farzulla. The hedging paradox: Fourth stage AML and financial surveillance, 2025.

[25] Murad Farzulla. Market reaction asymmetry: TARCH-X event study of cryptocurrency governance events. *SSRN Electronic Journal*, 2025. doi: 10.2139/ssrn.5788082. Under review at Digital Finance Journal (Springer).

[26] Murad Farzulla. Relational functionalism: AI friendship and moral consideration, 2025.

[27] Murad Farzulla. The replicator-optimization mechanism: A formal foundation for selection-transmission dynamics, 2025. In submission, arXiv.

[28] Murad Farzulla and Aleksei Maksakov. ASRI: Aggregated systemic risk index for cryptocurrency markets, 2025.

[29] 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.

[30] Russell Golman and Scott E. Page. Polya's bees: A model of decentralized decision-making. *Science Advances*, 1(8):e1500253, 2015. doi: 10.1126/sciadv.1500253.

[31] Robert E. Goodin. Enfranchising all affected interests, and its alternatives. *Philosophy & Public Affairs*, 35(1):40–68, 2007.

[32] Jürgen Habermas. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Polity Press, Cambridge, 1996.

[33] K. P. Hadeler. Stable polymorphisms in a selection model with mutation. *SIAM Journal on Applied Mathematics*, 41(1):1–7, 1981.

[34] Oliver Hart and John Moore. Property rights and the nature of the firm. *Journal of Political Economy*, 98(6):1119–1158, 1990.

[35] Eduard Hartwich, Alexander Rieger, Johannes Sedlmeir, Dominik Jurek, and Gilbert Fridgen. Machine economies. *Electronic Markets*, 33(1):36, 2023. doi: 10.1007/s12525-023-00649-0.

[36] Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.

[37] Bengt Holmström. Moral hazard and observability. *Bell Journal of Economics*, 10(1):74–91, 1979.

[38] Leonid Hurwicz. Optimality and informational efficiency in resource allocation processes. *Mathematical Methods in the Social Sciences*, pages 27–46, 1960.

[39] Shiv Issar and Amit Prasad. What is algorithmic governance? *Sociology Compass*, 15(12): e12921, 2021. doi: 10.1111/soc4.12921.

[40] 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. doi: 10.1016/0304-405X(76)90026-X.

[41] Christian Katzenbach and Lena Ulbricht. Algorithmic governance. *Internet Policy Review*, 8(4), 2019. doi: 10.14763/2019.4.1425.

[42] Özgecan Koçak and Phanish Puranam. The dual challenge of search and coordination for organizational adaptation: How structures of influence matter. *Organization Science*, 34 (2):851–869, 2023. doi: 10.1287/orsc.2022.1601.

[43] Pascal D. König. Dissecting the algorithmic Leviathan: On the socio-political anatomy of algorithmic governance. *Philosophy & Technology*, 33(4):647–667, 2019. doi: 10.1007/s13347-019-00363-5.

[44] Jon A. Krosnick. Survey research. *Annual Review of Psychology*, 50:537–567, 1999.

[45] Timur Kuran. *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press, 1995.

[46] Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473, 2017.

[47] Adam Lerer and Alexander Peysakhovich. Learning social conventions in markov games. In *International Conference on Learning Representations*, 2019.

[48] Richard C. Lewontin. The units of selection. *Annual Review of Ecology and Systematics*, 1:1–18, 1970.

[49] Douglas MacKay. The all affected principle, and the weighting of votes. *Politics, Philosophy & Economics*, 19(3):235–257, 2020. doi: 10.1177/1470594X20915602.

[50] Richard Maher, Sarah Valdez, and Brendan McCaffrie. (inequality in) interest group involvement and the legitimacy of policy making. *British Journal of Political Science*, 52(4): 1587–1605, 2022. doi: 10.1017/S0007123421000326.

[51] 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. In John Parkinson and Jane Mansbridge, editors, *Deliberative Systems: Deliberative Democracy at the Large Scale*, pages 1–26. Cambridge University Press, 2012.

[52] Eric Maskin. Nash equilibrium and welfare optimality. *Review of Economic Studies*, 66(1): 23–38, 1999.

[53] John Maynard Smith and George R. Price. The logic of animal conflict. *Nature*, 246:15–18, 1973.

[54] 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. In Sebastian J. van Strien and Sjoerd M. Verduyn Lunel, editors, *Stochastic and Spatial Structures of Dynamical Systems*, pages 183–231. North-Holland, 1996.

[55] Roger B. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1): 58–73, 1981. doi: 10.1287/moor.6.1.58.

[56] Martin A. Nowak. *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press, 2006.

[57] Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990.

[58] Karen M. Page and Martin A. Nowak. Unifying evolutionary dynamics. *Journal of Theoretical Biology*, 219(1):93–98, 2002.

[59] Fabienne Peter. *Democratic Legitimacy*. Routledge, New York, 2009.

[60] Philip Pettit. *Republicanism: A Theory of Freedom and Government*. Oxford University Press, Oxford, 1997.

[61] George R. Price. Selection and covariance. *Nature*, 227:520–521, 1970.

[62] John Rawls. *A Theory of Justice*. Harvard University Press, 1971.

[63] Marcel K. Richter. Revealed preference theory. *Econometrica*, 34(3):635–645, 1966.

[64] Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

[65] Paul A. Samuelson. A note on the pure theory of consumer's behaviour. *Economica*, 5(17): 61–71, 1938.

[66] Thomas M. Scanlon. *What We Owe to Each Other*. Harvard University Press, 1998.

[67] Fritz W. Scharpf. *Governing in Europe: Effective and Democratic?* Oxford University Press, Oxford, 1999.

[68] 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.

[69] Amartya Sen. *Collective Choice and Social Welfare: Expanded Edition*. Harvard University Press, 2017.

[70] A. John Simmons. *Moral Principles and Political Obligations*. Princeton University Press, Princeton, 1979.

[71] Peter D. Taylor and Leo B. Jonker. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(1–2):145–156, 1978.

[72] Arne Traulsen and Christoph Hauert. Stochastic evolutionary game dynamics. *Reviews of Nonlinear Dynamics and Complexity*, 2:25–61, 2009. doi: 10.1002/9783527628001.ch2.

[73] Merlijn van Hulst, Laurens de Graaf, and Niels Karsten. Creating legitimacy for citizen initiatives: Representation, identity and strategic networking. *Critical Policy Studies*, 14 (2):131–151, 2020. doi: 10.1080/19460171.2018.1549561.

[74] Peter Vanderschraaf. *Learning and Coordination: Inductive Deliberation, Equilibrium, and Convention*. Routledge, 2014. doi: 10.4324/9781315054797.

[75] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021.

[76] 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 $T_P : \mathcal{O} \to \mathbb{R}$ specifying preferences over outcomes. The agent has a (potentially different) target function $T_A : \mathcal{O} \to \mathbb{R}$.

**Definition A.1** (Delegation Configuration). A **delegation configuration** $\mathcal{C} = (P, A, s_P, T_P, T_A, \mathcal{I})$ specifies:

- Principal $P$ with stake $s_P > 0$
- Agent $A$ with decision authority over domain $d$
- Target functions $T_P, T_A : \mathcal{O} \to \mathbb{R}$
- Information structure $\mathcal{I}$ specifying what $A$ observes about $T_P$

The agent selects action $a \in \mathcal{A}$ to maximize their target:

$$a^* = \arg\max_{a \in \mathcal{A}} T_A(o(a)) \tag{90}$$

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:

$$\text{Loss}_P = T_P(o(a_P^*)) - T_P(o(a_A^*)) \tag{91}$$

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 [40], 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 $\alpha$, stakes $\sigma$, and entropy $\varepsilon$.

## A.3 Alignment and Residual Loss

Define alignment as the correlation between target functions:

$$\alpha = \text{corr}(T_P, T_A) = \frac{\text{Cov}(T_P, T_A)}{\sqrt{\text{Var}(T_P) \cdot \text{Var}(T_A)}} \tag{92}$$

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:*

$$\mathbb{E}[C_R] = \frac{\sigma \cdot V(T)}{1 + \alpha} \tag{93}$$

*where $\sigma$ 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:

$$\mathbb{E}[T_P(o_A^*)] = \mathbb{E}[T_P] + \text{Cov}(T_P, T_A) \cdot \frac{\mathbb{E}[T_A(o_A^*)] - \mathbb{E}[T_A]}{\text{Var}(T_A)}$$

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:

$$\varepsilon = 1 - \frac{I(T_P; \mathcal{I})}{H(T_P)} \tag{94}$$

where $I(T_P; \mathcal{I})$ 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:*

$$C_{info} = C_{base} \cdot (1 + \varepsilon) \tag{95}$$

*Proof.* The agent optimizes $\hat{T}_P$, their estimate of the principal's target, rather than $T_P$ itself. The estimation error is:

$$\mathbb{E}[(T_P - \hat{T}_P)^2] = \text{Var}(T_P | \mathcal{I}) = \text{Var}(T_P) \cdot \varepsilon$$

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 $\alpha$, stakes $\sigma$, and entropy $\varepsilon$, total agency cost takes the form:*

$$\boxed{F = \sigma \cdot \frac{1 + \varepsilon}{1 + \alpha}} \tag{96}$$

*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 $\sigma$), 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. $\qquad\qquad\square$ $\qquad\qquad\square$

## 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:*

$$F = \sum_{i=1}^{n} F_i = \sum_{i=1}^{n} s_i \cdot \frac{1 + \varepsilon_i}{1 + \alpha_i} \tag{97}$$

*where $\alpha_i = corr(T_i, T_A)$ and $\varepsilon_i$ 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. $\qquad\qquad\square$ $\qquad\qquad\square$

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 $(\varepsilon)$ 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** $(\sigma)$: 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** [40]: Our decomposition follows their agency cost structure, with alignment corresponding to goal congruence and entropy to information asymmetry.

2. **Holmström** [37]: The informativeness principle—that monitoring improves when signals are correlated with agent effort—corresponds to our entropy reduction mechanism.

3. **Hart-Moore** [34]: 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 $\alpha, \varepsilon$

**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 $g, h$

**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:*

$$F(\sigma, \alpha, \varepsilon) = c \cdot \sigma \cdot \frac{a + \varepsilon}{b + \alpha} \tag{98}$$

*for constants $c > 0$, $a > 0$, $b > 1$.*

*Setting $c = 1$, $a = 1$, $b = 1$ yields the canonical form:*

$$\boxed{F = \sigma \cdot \frac{1 + \varepsilon}{1 + \alpha}} \tag{99}$$

*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:*

$$\tilde{F}(\sigma, \alpha, \varepsilon) = \phi\left(\sigma \cdot \frac{1 + \varepsilon}{1 + \alpha}\right) \tag{100}$$

*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 $\sigma$. 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 $\alpha$ with a pole at $\alpha = -1$. By D5, $\tilde{h}$ is increasing in $\varepsilon$.
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, $\psi$ 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 $\sigma$

**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:*

$$F = \sigma \cdot Rate[\text{Info loss}] = \sigma \cdot \frac{H(T_P | \hat{T}_P)}{I(T_P; T_A)} \tag{101}$$

*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 $1 + H(T_P | \mathcal{I})$ (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 $\sigma$ 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.

# 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** $\mathcal{S} = \mathbb{R}^m_{\geq 0}$: Resource levels for $m$ resources
- **Action space** $\mathcal{A}_i = \{-1, 0, +1\}^m$ 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 $R_i : \mathcal{S} \to \mathbb{R}$ parameterized to control alignment.

**Definition C.2** (Parameterized Rewards)**.** Agent $i$'s reward function is:

$$R_i(s) = \sum_{j=1}^{m} w_{ij} \cdot u(s_j - \tau_{ij}) \tag{102}$$

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** $\sigma$: Via $\sum_i \|w_i\|$ (aggregate weight magnitude)

- **Alignment** $\alpha$: Via $\mathrm{corr}(\tau_i, \tau_j)$ across agents (target correlation)

- **Entropy** $\varepsilon$: 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:

$$\tilde{s}_i = s + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \varepsilon \cdot I_m) \tag{103}$$

where $\varepsilon \in [0, 1]$ controls noise magnitude.

Higher $\varepsilon$ 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:
Total: 125 experimental conditions, with $k = 30$ replications each.

Table 6: Experimental Factors

| Factor | Symbol | Levels | Manipulation |
|--------|--------|--------|--------------|
| Alignment | $\alpha$ | $\{-0.8, -0.4, 0, 0.4, 0.8\}$ | Target correlation |
| Stakes | $\sigma$ | $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ | Weight magnitude |
| Entropy | $\varepsilon$ | $\{0, 0.25, 0.5, 0.75, 1.0\}$ | Observation noise |

### C.2.2 Dependent Variables (Friction Proxies)

We measure coordination failure through four operationalizations:

1. **Reward Gap** $\Delta R$: Difference between Nash equilibrium payoff and realized payoff

$$\Delta R = R^* - \frac{1}{T} \sum_{t=1}^{T} \bar{R}_t \tag{104}$$

2. **Convergence Time** $\tau_c$: Episodes until policy stabilization (change $< \delta$)

$$\tau_c = \min\{e : \|\pi_e - \pi_{e-1}\| < \delta\} \tag{105}$$

3. **Policy Variance** $\mathrm{Var}(\pi)$: Variance in joint policy across replications

$$\mathrm{Var}(\pi) = \frac{1}{k} \sum_{r=1}^{k} \|\pi_r - \bar{\pi}\|^2 \tag{106}$$

4. **Pareto Inefficiency** $\eta$: Distance from Pareto frontier

$$\eta = \min_{R^{Pareto}} \|R^{realized} - R^{Pareto}\| \tag{107}$$

### 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**: $\epsilon$-greedy, $\epsilon = 0.1 \to 0.01$ 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:

$$\frac{\partial \Delta R}{\partial \alpha} < 0, \quad \frac{\partial \tau_c}{\partial \alpha} < 0 \tag{108}$$

*Hypothesis* 2 (H2: Stakes-Friction Positive Relationship). Measured friction proxies increase with stakes:

$$\frac{\partial \Delta R}{\partial \sigma} > 0, \quad \frac{\partial \tau_c}{\partial \sigma} > 0 \tag{109}$$

*Hypothesis* 3 (H3: Entropy-Friction Positive Relationship). Measured friction proxies increase with entropy:

$$\frac{\partial \Delta R}{\partial \varepsilon} > 0, \quad \frac{\partial \tau_c}{\partial \varepsilon} > 0 \tag{110}$$

*Hypothesis* 4 (H4: Friction Function Fit). Measured friction is well-predicted by the theoretical form:

$$\Delta R \approx \beta_0 + \beta_1 \cdot \sigma \cdot \frac{1 + \varepsilon}{1 + \alpha} + \text{error} \tag{111}$$

with $R^2 > 0.7$.

## C.4 Analysis Plan

### C.4.1 Regression Specification

For each friction proxy $Y \in \{\Delta R, \tau_c, \text{Var}(\pi), \eta\}$:

$$Y_{ijk} = \beta_0 + \beta_1 F_{ijk} + \gamma_1 \alpha_i + \gamma_2 \sigma_j + \gamma_3 \varepsilon_k + \mu_{ijk} \tag{112}$$

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 $\sigma$, 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 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 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)

Positive results would establish that the friction framework captures real coordination dynamics, not merely normative intuitions. Negative results would identify boundary conditions where the framework fails, prompting theoretical refinement.