Dissensus AI Working Paper Series

DAI-2503

The Replicator-Optimization Mechanism

Computational Unity Across Physical and Abstract Substrates

Murad Farzulla[1,2,*]

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

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

February 2026

Abstract

What persists is what we observe, and what we observe is what has survived selection—persistence under pressure is what fitness ultimately reduces to, and this tautology has more teeth than it first appears. Equilibria in the classical sense are not observed in practice but only approached asymptotically, which means that what we actually see is perpetual motion, adjustment, and friction, where friction is the energy dissipated in the gap between current configurations and the equilibria they cannot reach.

What this paper offers is a demonstration that four independent fields—physics, biology, economics, and cultural evolution—have converged on the same mathematical machinery for describing these dynamics. The convergence is not metaphorical but structural: a rigorous isomorphism of mathematical form, where the same fitness landscapes, selection operators, and transmission kernels appear independently because they describe something real about how persistent systems behave. We synthesize these convergent results into the Replicator-Optimization Mechanism (ROM): a unified apparatus for persistence-conditioned dynamics that can be instantiated at any scale.

The primary application is political philosophy, where we instantiate ROM with friction from stake-voice mismatch as the primitive quantity, legitimacy as survival probability, and belief-transfer as the mutation kernel modulator. What emerges is not so much a new formalism as a translation manual—one showing that political philosophy's debates about consent and legitimacy are at bottom debates about friction and selection, and that the formal tools to make progress already exist in adjacent fields.

The core algebraic results—simplex preservation, survival monotonicity, moving equilibrium existence, and the impossibility of static equilibrium under varying friction—have been machine-checked in Lean 4 with the Mathlib library (28 theorems, zero `sorry` placeholders; Appendix F).

Keywords: persistence, selection, friction, scale-relativity, convergent dynamics, replicator equation, consent, institutional evolution, formal verification

1 Introduction

1.1 The Thesis

Every persistent pattern optimizes for persistence, which sounds like a tautology until you notice how much work it does: what we observe is what has survived, and what has survived is what was fit to survive, and fitness when you strip away the domain-specific language reduces to persistence under selection pressure, which means that the tautology has teeth after all.

The second thesis is perhaps less obvious but equally important: equilibria in the classical sense do not exist in practice, or at least not as stable resting points that systems actually reach. Classical economics assumes markets clear, classical game theory assumes players converge to Nash equilibrium, classical political philosophy assumes legitimate states achieve stable consent, and yet none of these equilibria are observed in the world—what we see instead is perpetual motion, friction, adjustment, selection, and then more friction, with systems approaching equilibria asymptotically without ever arriving.

The third thesis follows naturally from the first two: friction is the measurable signature of this disequilibrium, the energy dissipated by systems that cannot reach equilibrium but keep trying anyway. This dissipation manifests differently depending on the substrate—as heat in thermodynamics, as volatility in markets, as protest and exit in politics, as coordination failure in multi-agent systems—but the underlying dynamic is the same, which suggests that friction is not pathology but rather the universal signature of systems optimizing under constraint.

What this paper argues is that four independent scientific fields have converged on the same formal machinery for describing these dynamics, and this convergence is not analogy or metaphor but a structural isomorphism—the same mathematical machinery appearing in different substrates because it describes something real about how persistent systems actually behave.

1.2 The Convergence

Consider what each field has independently discovered:

Physics. Statistical mechanics describes systems that never reach equilibrium but fluctuate around it, dissipating energy as friction (Kubo, 1966). The Mori-Zwanzig formalism (Mori, 1965; Zwanzig, 1961) shows how coarse-graining introduces memory effects—the influence of unobserved degrees of freedom on observed dynamics. Effective field theory (Weinberg, 1979) treats the choice of "fundamental" units as scale-relative: what counts as atomic depends on your energy scale. The renormalization group (Wilson, 1971; Kadanoff, 1966) provides the mathematical apparatus for relating descriptions at different scales.

Biology. The replicator equation (Taylor and Jonker, 1978) describes how type frequencies change under selection. The Price equation (Price, 1970) partitions evolutionary change into selection and transmission components in a manner explicitly substrate-neutral. Multi-level selection theory (Okasha, 2006) shows that selection operates at multiple scales simultaneously, with the "level" of selection being a parameter of analysis rather than a fact about the world. Porter and Wikman (2026) provide updated formal conditions for evolutionary stability in economic contexts, extending the classical ESS concept with refinements directly applicable to the persistence dynamics formalized here.

Economics. Institutional economics (North, 1990) describes how rules persist, evolve, and are selected. Evolutionary game theory shows how strategies propagate through populations with variation and differential survival (Weibull, 1995). Recent work on learning dynamics (Pangallo et al., 2019; Galla and Farmer, 2013) shows that convergence to equilibrium is the exception, not the rule: most learning processes oscillate or exhibit chaos. Shen et al. (2026) extend these dynamics to reputation-based voluntary participation games, demonstrating how voluntary entry and exit—formally analogous to the consent withdrawal mechanism in our framework—shapes the evolutionary stability landscape.

Cultural evolution. The cultural Price equation (El Mouden et al., 2014) applies selection-transmission dynamics to ideas, norms, and institutions. Work on gene-culture coevolution (Boyd and Richerson, 1985; Cavalli-Sforza and Feldman, 1981; Henrich, 2016) demonstrates that cultural transmission exhibits the same formal structure as genetic transmission, with different parameters.

The pattern: scale-relative parameterization, selection under differential fitness, transmission with variation, and perpetual disequilibrium with friction as the measurable signature. Four fields. Same mathematics. Different substrates. Recent work by Sornette et al. (2026) arrives at structural friction dynamics from statistical physics, identifying learned human interaction structures as endogenous sources of alignment failure and AGI as an evolutionary shock to those structures—providing independent convergence evidence for the scale-relative persistence framework developed here.

1.3 Why Political Philosophy Is Behind

Political philosophy has been asking a question that may not be the right one to ask, or at least not the most productive one. For centuries the central question has been "What makes authority legitimate?"—a question that assumes legitimacy is a property that arrangements possess or lack, something that can be determined through philosophical analysis in the way that one determines whether an argument is valid or a definition is coherent.

What the convergent evidence from other fields suggests is a different framing: legitimacy might be better understood not as a property but as a survival probability, where arrangements generating low friction persist and arrangements generating high friction are selected against, and what we call "legitimate" is simply what we call configurations that have survived long enough to seem natural rather than contingent. This reframing is not meant to be deflationary or to dissolve the normative questions that political philosophers care about, but rather to ground them in dynamics that can actually be measured and tested: arrangements with systematic stake-voice mismatch will generate friction, friction accumulates until reconfiguration becomes unavoidable, and these dynamics turn out to be formally identical to selection in biology and dissipation in physics.

If political philosophy has not adopted this framing, it is perhaps because it has not been translated into the discipline's vocabulary, and what this paper attempts to provide is something like a translation manual.

1.4 The Contribution

To be clear about what is and is not being claimed here: we do not claim to have discovered persistence, or selection, or friction, or scale-relativity, all of which are established results in their respective fields with long literatures behind them. What we offer is synthetic rather than novel in that sense:

1. **Demonstration of convergence.** We show that four fields have arrived at what is essentially the same formal structure through independent routes, and this convergence across disciplines with different methods and different empirical bases is itself evidence that the structure captures something real about how persistent systems behave.

2. **Unified formalism.** We synthesize these convergent results into what we call the Replicator-Optimization Mechanism, or ROM: a single apparatus with explicit modeling choices about scale, atomic unit, fitness function, and transmission kernel, which can then be instantiated in any domain where one wishes to study persistence-conditioned dynamics.

3. **Political philosophy instantiation.** We apply ROM to the traditional concerns of political philosophy—consent and legitimacy—by mapping these concepts onto the established dynamics that other fields have been characterizing for decades, where friction from stake-voice mismatch becomes the primitive quantity, legitimacy becomes survival probability, and belief-transfer becomes a mutation kernel modulator.

4. **Grounding for the Axiom of Consent.** This paper provides independent support for the framework developed in Farzulla (2025a), which claims that friction is predictable from the kernel triple $(\alpha, \sigma, \varepsilon)$. What this paper shows is why that claim should be believed: because it formalizes dynamics that physics, biology, and economics have independently confirmed through their own methods.

The ROM formalism connects to a broader research programme investigating adversarial dynamics across domains. The consent-friction instantiation developed here extends naturally to AI governance, where Farzulla (2025d) argues that existentially vulnerable autonomous systems satisfying functional criteria for political standing cannot be legitimately ruled without consent—a claim that gains formal grounding once legitimacy is understood as survival probability under ROM dynamics. The framework also applies to financial regulation, where Farzulla (2025b) demonstrates that hedging instruments exhibit the same pharmakon structure identified here: the mechanism that creates systemic friction is also the mechanism that reveals it. The connection between ROM's persistence dynamics and developmental psychology is explored in Farzulla (2025e), which models maladaptive learning as corrupted transmission kernels—training data that generates persistent maladaptive patterns through the same selection-transmission machinery that ROM formalizes.

The question worth asking, then, is not so much whether ROM is correct—four fields have already validated its component parts—but rather why political philosophy has not yet adopted a formalism that other fields have found so useful.

1.5 Roadmap

Section 2 establishes foundational definitions: scale, atomic units, persistence, friction. Section 3 presents the ROM axioms and their justification through convergent evidence. Section 4 develops the mathematical machinery: coarse-graining, memory effects, the Ladder Constraint on scale-skipping. Section 5 instantiates ROM for political philosophy: consent as friction-minimization, legitimacy as survival probability. Section 6 addresses empirical operationalization and policy implications. Section 7 provides a worked example applying the framework to medical delegation. Section 8 discusses limitations, adjacent frameworks, and the descriptive-normative gap. Section 9 concludes.

2 Foundational Definitions

2.1 Primitive Concepts

Definition 2.1 (Scale)**.** A scale S is a level of description characterized by a choice of minimal distinguishable unit and a characteristic spatiotemporal resolution. Scales are observer-relative measurement choices, not objective features of reality.

Definition 2.2 (Atomic Agent)**.** Given scale S, the atomic agent Atom_S is the minimal unit of analysis—the entity treated as indivisible at that scale. At particle scale, Atom = elementary particle; at cellular scale, Atom = cell; at institutional scale, Atom = institution. The atomic agent is not ontologically fundamental; it is the unit relative to which dynamics are measured.

This scale-relativity of atomic units is established methodology:

- **Physics**: Effective field theory treats effective degrees of freedom as scale-dependent since Weinberg (1979)
- **Biology**: Multi-level selection theory establishes that selection level is a parameter of analysis (Okasha, 2006)
- **Economics**: Agent-based modeling treats the "agent" as a modeling choice, not ontological primitive

ROM adopts this stance: there is no privileged "fundamental" scale. Each scale has its appropriate description; which to use is pragmatic (predictive success) rather than ontological.

2.2 Entropy Pressure and Persistence

Definition 2.3 (Entropy Pressure)**.** Entropy pressure is the tendency for configurations to disperse toward higher-entropy states in the absence of maintenance processes. Complex configurations tend to dissolve; persistence requires active maintenance.

This is the second law of thermodynamics applied to pattern persistence. Schrödinger (1944) noted that life maintains order against entropy; we generalize to any persistent pattern at any scale.

The key insight: What requires explanation is not why things change but why some things *remain*. Persistence is non-trivial. Selection is what happens when some patterns persist better than others.

3 The ROM Axioms: Convergent Foundations

The following axioms characterize systems where ROM applies. Each axiom has independent confirmation from multiple fields.

Axiom 1 (Minimal Atoms)**.** At any scale S, there exists a set of minimal units serving as carriers of properties and loci of interactions.

Convergent evidence: This generalizes "interactors" in evolutionary theory (Hull, 1980), "agents" in economics, and "degrees of freedom" in physics. Each field independently requires a notion of minimal unit at each descriptive level.

Axiom 2 (Interaction Network)**.** Atomic agents are embedded in an interaction network $G_{S,t}$ determining which agents influence which others.

Convergent evidence: Network structure mediates dynamics in every field—social networks (Newman, 2010), gene regulatory networks, financial contagion networks, neural networks. Recent work on higher-order interactions demonstrates that hypergraph structure fundamentally alters cooperation dynamics beyond pairwise approximations (Alvarez-Rodriguez et al., 2021; Sadekar et al., 2025). The mathematical apparatus (graph theory, spectral methods) transfers directly.

Axiom 3 (Entropy Pressure)**.** In the absence of maintenance processes, configurations tend toward higher-entropy states.

Convergent evidence: This is the second law of thermodynamics. No field disputes it.

Axiom 4 (Replication with Variation)**.** Some patterns propagate—inducing similar patterns elsewhere. Propagation occurs with variation: copies are imperfect.

Convergent evidence: This is the inheritance principle (Darwin, 1859; Lewontin, 1970). Biology formalizes it as genetic transmission; cultural evolution as social learning (Boyd and Richerson, 1985); economics as institutional diffusion; physics as pattern replication in dissipative systems.

Axiom 5 (Concentration)**.** In the limit of large populations, macro-observables concentrate around expectations. Stochastic micro-dynamics yield approximately deterministic macro-dynamics.

Convergent evidence: This is the law of large numbers, concentration of measure (Ledoux, 2001). Statistical mechanics, population genetics, and economics all rely on this principle.

4 Mathematical Machinery

4.1 The ROM Equation

Given the axioms, temporal evolution is governed by the weighted replicator-mutator equation:

$$\frac{dp_t(\tau)}{dt} = \sum_{\tau' \in T_S} p_t(\tau') \cdot w_S(\tau') \cdot \rho_S(\tau', G_{S,t}, p_t) \cdot M_S(\tau' \to \tau) - p_t(\tau) \cdot \bar{\phi}_t \tag{1}$$

This equation is not new. It is the standard replicator-mutator equation (Hadeler, 1981; Page and Nowak, 2002), with temporal and spatial extensions well-characterized in Roca et al. (2009). Recent generalizations extend the formalism in two important directions. Varga (2024) generalizes replicator dynamics for evolutionary matrix games under time constraints—mandatory waiting periods between interactions that are formally analogous to friction in our framework. The time constraint between interactions maps directly onto ROM's concept of dissipative friction: both represent structural impediments that do not merely slow dynamics but fundamentally reshape the equilibrium landscape. Varga's demonstration that the ESS-replicator relationship is restored via generalized dynamics under these constraints supports ROM's central claim that friction-generating mechanisms alter which configurations persist rather than simply impeding convergence. The mathematical foundations for this infinite-dimensional extension were established by Mendoza-Palacios and Hernández-Lerma (2017), who proved stability results for the replicator dynamics on separable metric strategy spaces, demonstrating that equilibrium stability depends critically on the topology of the measure space—a result that constrains which coarse-graining procedures preserve dynamical structure. Mendoza-Palacios and Hernández-Lerma (2024) subsequently generalize the replicator dynamics to metric strategy spaces evolving in Banach spaces of finite signed measures, providing a rigorous infinite-dimensional framework connecting Nash equilibria stability to replicator dynamics. ROM's scale-relative parameterization—where the type space T_S varies with the choice of descriptive scale—implicitly operates within such a framework, and the Banach space formulation provides the mathematical infrastructure for making the continuum limit of ROM dynamics precise.

- $p_t(\tau)$: Frequency of type τ at time t

- $w_S(\tau)$: Intrinsic weight (baseline replication capacity)
- $\rho_S(\tau, G, p)$: Survival probability given network and population state
- $M_S(\tau' \to \tau)$: Transmission kernel (mutation/learning)
- $\bar{\phi}_t$: Mean fitness (normalization)

4.2 Formal Equivalences

The claim that ROM connects to other formalisms is not analogy but structural identity—a rigorous correspondence of mathematical form:

Price Equation. The discrete-time analogue yields the Price partition (Price, 1970):

$$\Delta \bar{z} = \frac{1}{\bar{w}}\mathrm{Cov}(w, z) + \frac{1}{\bar{w}}\mathbb{E}[w \cdot \Delta z] \tag{2}$$

This equivalence under discretization is proven in Page and Nowak (2002).

Information Geometry. Replicator dynamics have natural information-geometric interpretation via the Shahshahani metric (Shahshahani, 1979; Hofbauer and Sigmund, 1998). Under detailed balance conditions, ROM reduces to gradient flow on the Fisher-Rao manifold.

Remark 4.1 (Bayesian Interpretation). Under pure selection (no mutation), type frequencies evolve exactly as posterior probabilities under iterated Bayesian updating (Bettencourt et al., 2025; Harper, 2009; Czégel et al., 2022):

$$p_{t+1}(\tau) = \frac{w_S(\tau) \cdot p_t(\tau)}{\bar{w}_t} \quad \longleftrightarrow \quad p(H|D) = \frac{p(D|H) \cdot p(H)}{p(D)} \tag{3}$$

where fitness $w_S(\tau)$ corresponds to the likelihood function $p(D|H)$—the probability of observing environmental data given type τ as hypothesis. This correspondence clarifies what "optimization" means in ROM: the system maximizes mutual information $I(E, \mathscr{T})$ between environmental states and type distributions, which is equivalent to maximizing average log-fitness. Selection thus implements *predictive optimization*—configurations that better predict (track, reflect) environmental structure persist—rather than utility maximization in the economic sense.

When the Bayesian Correspondence Breaks. The Bayes-replicator equivalence holds precisely under pure selection. ROM's extensions—mutation kernels and network dependence—modify this correspondence in well-characterized ways. With mutation, ROM dynamics correspond to *filtering in Hidden Markov Models* (Akyildiz, 2017; Pathiraja and Wacker, 2024): the mutation kernel $M_S(\tau' \to \tau)$ represents transitions between hidden states (hypotheses), while selection provides likelihood weighting. Network externalities make the "likelihood" endogenous, potentially inducing cyclic dominance and limit cycles (Sato and Crutchfield, 2003; Galla and Farmer, 2013). The belief-transfer modulation $g(\bar{O}', \bar{O}) = \exp(-\gamma(\bar{O}' - \bar{O}))$ violates detailed balance whenever ownership perceptions differ between configurations, causing dynamics to exhibit circulation around equilibria rather than monotonic convergence. In non-stationary environments with mutation and network effects, "optimization" means tracking a moving target rather than converging to a fixed optimum.

Reinforcement Learning. The connection between softmax policy gradients and replicator dynamics is explicit (Tuyls et al., 2003; Bloembergen et al., 2015). Under softmax action selection with temperature τ, policy gradient dynamics reduce to:

$$\frac{d\pi(a)}{dt} = \pi(a) \cdot (Q(a) - \bar{Q}) \tag{4}$$

which is precisely the replicator equation with Q-values as fitness. The mutation kernel M_S in ROM corresponds to exploration: temperature-modulated randomization that prevents premature convergence to local optima.

Belief-Transfer and Kernel Modulation. The belief-transfer mechanism (where consent-holders develop ownership psychology over domains they control) induces specific changes in the mutation kernel. Let $\bar{O}(\tau)$ denote the average ownership-perception among agents in configuration τ. The mutation kernel entries are modulated:

$$M_S(\tau' \to \tau) = M_0(\tau' \to \tau) \cdot g(\bar{O}(\tau'), \bar{O}(\tau)) \tag{5}$$

where M_0 is the baseline kernel and the ownership-modulation function takes an Arrhenius-like form:

$$g(\bar{O}', \bar{O}) = \exp\bigl(-\gamma(\bar{O}' - \bar{O})\bigr), \quad \gamma > 0 \tag{6}$$

This suppresses transitions that reduce aggregate ownership perception, connecting micro-level psychological processes to macro-level institutional dynamics. The distinctive prediction: regime transition probability should *decrease exponentially* with incumbent tenure, controlling for legitimacy and resources—a specific functional form that generic "institutional stickiness" explanations do not generate. Appendix E provides convergent microfoundational derivations of this Arrhenius-like form from statistical mechanics, Kramers rate theory, behavioral economics (loss aversion), and bounded rationality (quantal response equilibrium).

Recent work by Balabanova et al. (2025) demonstrates that institutional incentives can be rigorously incorporated into replicator-mutator dynamics through fitness modifiers. ROM's legitimacy function L acts as a positive institutional modifier (analogous to their reward component), while friction F provides the corresponding negative modifier. The survival probability $\rho_S = L/(1+F)$ generalizes additive incentive structures to the multiplicative form appropriate for survival probabilities.

Convergence and Limit Cycles. Not all learning dynamics converge, and this is a feature rather than a limitation. Pangallo et al. (2019) demonstrate that convergence to Nash equilibrium is the exception rather than the rule: most games produce best-reply cycles, limit cycles, or chaotic attractors. ROM dynamics are no exception. Under detailed balance conditions on M_S—which require symmetric belief-transfer between configurations—the dynamics reduce to gradient flow with quasi-potential:

$$V(\tau) = \log L(\tau) - \log(1 + F(\tau)) + \log w_S(\tau)$$

However, detailed balance generically *fails* in the consent-friction instantiation. The belief-transfer modulation $g(\bar{O}', \bar{O}) = \exp(-\gamma(\bar{O}' - \bar{O}))$ violates detailed balance whenever ownership perceptions differ between configurations (Section 4.2), and asymmetric mutation kernels produce non-zero circulation around the simplex interior (Appendix D, Counterexample 1). When detailed balance breaks, limit cycles become the *generic* case: configurations orbit rather than converge, with friction-minimizing regions acting as attractors in the time-averaged sense rather than as fixed-point equilibria.

This is not a defect but a prediction. Democratization-backsliding cycles—where regimes oscillate between liberalization and retrenchment—are precisely what limit cycle dynamics produce. The oscillation occurs because legitimacy gains from liberalization shift ownership perceptions (increasing $\bar{O}$), which in turn raises transition barriers via the Arrhenius kernel, eventually generating sufficient latent friction to trigger retrenchment. The cycle repeats at a characteristic frequency set by the ratio of legitimacy accumulation to ownership-perception shift rates. The MARL validation (Farzulla, 2025a, Appendix F) provides independent computational evidence: 99.3% of conditions achieve reward convergence while only 0.85% achieve policy convergence—stable aggregate outcomes emerge from perpetual strategic cycling, exactly the dynamic equilibrium that limit cycles describe.

4.3 The Kernel Triple

The key parameterization is (ρ_S, w_S, M_S) at each scale:

- ρ_S: Survival function mapping type, network, population → persistence probability
- w_S: Weight function assigning baseline capacity to types
- M_S: Transmission kernel (row-stochastic)

Different domains instantiate different kernels:

Scale	**Atom**	ρ_S	M_S
Cellular	Cell	Replication rate	Mutation
Organism	Organism	Darwinian fitness	Genetic transmission
Agent	Intentional agent	Strategy payoff	Learning, imitation
Institutional	Institution	Legitimacy	Reform, evolution

4.4 Coarse-Graining and the Ladder Constraint

Scales connect via coarse-graining operators $\pi_{S \to S'}$. When does coarse-graining preserve ROM structure?

Theorem 4.1 (Lumpability Conditions). *Let $\pi : T_S \to T'_S$ be a coarse-graining projection. Under standard regularity conditions (finite/countable type space, bounded survival function, row-stochastic mutation kernel), ROM structure is preserved under π if and only if:*

(i) ***Transition uniformity****: For all τ_i, τ_k with $\pi(\tau_i) = \pi(\tau_k)$, and all macro-types T': $\sum_{\tau_j:\pi(\tau_j)=T'} m_{ij} = \sum_{\tau_l:\pi(\tau_l)=T'} m_{kl}$*

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

When these fail, coarse-grained dynamics acquire memory terms (Mori-Zwanzig structure).

Proof sketch. (*Sufficiency*) Under (i) and (ii), define coarse distribution $P(T',t) = \sum_{\tau:\pi(\tau)=T'} p(\tau,t)$. By (ii), survival factors out; by (i), mutation sums collapse. Coarse dynamics satisfy ROM form.

(*Necessity*) If (i) fails, $P(T', t+\Delta t)$ depends on internal distribution $p(\tau|T,t)$, introducing memory. Similarly for (ii). The memory kernel $K(t-s)$ decays exponentially with rate determined by internal spectral gap. ■

Note on rigor: This is a proof sketch. Full formalization requires demonstrating that non-lumpable coarse-graining yields Mori-Zwanzig memory terms explicitly. Geiger and Kedem (2022) establishes that lumpability is generically rare (measure-zero in the space of Markov chains). Aristoff and Zhu (2023) shows how memory can be systematically incorporated when lumpability fails, while data-driven extraction of MZ operators (Tian et al., 2021) demonstrates empirical recovery of memory kernels.

Critical qualification. Memory effects are non-negligible precisely when internal equilibration timescales are comparable to observation timescales. When strong time-scale separation holds, memory terms decay rapidly and the Markovian approximation is accurate (the Chapman-Enskog regime).

This is not a novel result. It is the application of Markov chain lumpability theory (Kemeny and Snell, 1976) and Mori-Zwanzig formalism (Mori, 1965; Zwanzig, 1961) to replicator dynamics. Physics solved this problem decades ago.

The Ladder Constraint: Direct measurement at scale $S+2$ using atoms from scale S is generically ill-posed. The error satisfies:

$$\varepsilon(S \to S+2) \geq \varepsilon(S \to S+1) + \varepsilon(S+1 \to S+2) + \Delta_{\text{memory}}$$

This has known exceptions (RG fixed points, hierarchical symmetry, mean-field limits, strong time-scale separation) but holds generically. Formal statement and proof sketch appear in Appendix B.1; conditions under which the constraint relaxes are detailed in Appendix B.2. Network renormalization theory (Villegas et al., 2023) provides rigorous grounding. Zhang et al. (2025) demonstrate that information-preserving network compression requires merging structurally similar nodes; arbitrary compression destroys flow structure.

4.5 Causal Emergence and Scale-Relative Validity

The scale-relativity of ROM connects to the theory of causal emergence (Hoel et al., 2013; Hoel, 2017). Causal emergence occurs when coarse-grained descriptions exhibit higher *effective information*—a measure of causal determinism—than fine-grained descriptions.

Effective Information. For a transition matrix M, effective information $EI(M) = H^{out}_{max}(M) - \langle H^{out}(M) \rangle$ measures how deterministic dynamics are while retaining descriptive richness. Hoel et al. demonstrate that coarse-graining can *increase* EI: macro-descriptions sometimes exhibit higher causal determinism because aggregation eliminates degenerate causal pathways.

ROM Implication. If institutional-level dynamics exhibit higher EI than individual-level dynamics, the institutional description is not merely convenient but *causally superior* for prediction. This resolves a common objection to institutional analysis: the accusation of being “merely” descriptive dissolves when macro-dynamics demonstrably exhibit higher causal determinism than their micro-constituents. Cantner et al. (2019) demonstrate a related phenomenon in value chains: multi-layer structure can *reverse* apparent selection effects, with low-fitness firms persisting via high-fitness partners—precisely the kind of emergent dynamics that single-scale analysis misses.

Recent work by Varley and Hoel (2022) formalizes emergence as information conversion: coarse-graining can transform redundant information into synergistic information, creating genuinely new causal structure at the macro level. For ROM, this means legitimacy dynamics at the institutional scale may be *causally irreducible* to individual consent-holding dynamics.

5 The Consent-Friction Instantiation

What follows is an application of ROM to political philosophy, which is to say an attempt to map the traditional vocabulary of consent and legitimacy onto the dynamics that the preceding sections have characterized.

5.1 Domain Specification

In this instantiation, atomic agents are consent-holding entities—individuals, groups, institutions, and so forth—and the key quantities that characterize their interactions are:

- **Stakes** $s_i(d)$: The magnitude of impact that domain d has on agent i, which is to say how much agent i has at risk in decisions made within that domain
- **Voice** $v_i(d)$: Agent i's actual influence over decisions in d, measured by whatever mechanisms of input and control are available
- **Friction**: The tension that emerges when stakes and voice diverge, when those who bear the consequences of decisions lack proportional influence over those decisions

Note on prior work. The legitimacy definition here—stakes-weighted voice—builds on the formal framework developed in Farzulla (2025c), which establishes that legitimate governance requires proportional influence for those affected by decisions. What ROM adds is the dynamical grounding: legitimacy enters as survival probability in the replicator equation, meaning that configurations satisfying the legitimacy conditions persist while those violating them face selection pressure. The normative framework from that earlier work becomes empirically testable once instantiated in ROM dynamics. Recent work by Powers et al. (2023) provides independent support: their model of institutional coevolution shows that the *cost of consensus* scales with group size and political inequality, selecting for hierarchy vs. egalitarianism—a friction-driven mechanism consistent with ROM's predictions about legitimacy and institutional form.

5.2 The Friction Function

$$F(d,t) = \sum_i s_i(d) \cdot \frac{1+\varepsilon_i(d,t)}{1+\alpha_i(d,t)} \tag{7}$$

Where α_i is alignment (correlation between agent's interests and consent-holder's actions) and ε_i is information entropy (how much the consent-holder misunderstands the agent's preferences). Formal derivations of this functional form from Lagrangian optimization, information-theoretic, and diversity-based first principles appear in Appendices A.1–A.3.

Remark 5.1 (Quadratic Refinement). The MARL factorial experiment (Farzulla, 2025a, Appendix F) reveals a U-shaped alignment–friction relationship: neutral alignment ($\alpha = 0$) produces the worst coordination outcomes, while both cooperative and adversarial alignment reduce friction symmetrically. This motivates a second-generation friction form $F^{(2)} = \sigma(1+\varepsilon)/(1+\alpha^2)$, which replaces the $\alpha \to -1$ singularity with a bounded maximum at $\alpha = 0$. The quadratic form achieves $R^2 = 0.34$–0.43 versus $R^2 = 0.05$–0.13 for the canonical specification (Section 6.4). The formal development—axiomatic derivation under relaxed divergence conditions, uniqueness results, and agreement at $\alpha \in \{0,1\}$—appears in Farzulla (2025a), Appendix F. The canonical form is retained throughout this paper as the theoretical baseline; the quadratic form represents an empirical refinement whose domain of superiority (the adversarial regime $\alpha < 0$) is precisely characterized.

5.3 Pathologies: Observed versus Latent Friction

Some apparently low-friction systems achieve stability through suppression rather than genuine alignment—authoritarian regimes can appear stable precisely because dissent is costly to express. The framework accommodates this by distinguishing observed friction from latent friction, paralleling Kuran's analysis of preference falsification (Kuran, 1995):

- **Observed friction**: Friction that manifests in measurable behaviors—protest, litigation, noncompliance, exit
- **Latent friction**: Friction that exists (stake-voice mismatch) but is suppressed through coercion, censorship, or exit barriers
- **Suppression cost**: Resources expended to prevent latent friction from becoming observed—surveillance, enforcement, propaganda, border control

The ε term includes epistemic control: regimes that suppress information about alternatives, prevent coordination among dissenters, and control exit options exhibit high ε, which increases latent friction even when observed friction is low.

Proposition 5.1 (Suppression Instability)**.** *Regimes with low observed friction but high latent friction exhibit sudden tipping points when suppression costs exceed maintenance capacity or when exogenous shocks reduce suppression effectiveness.*

This predicts that apparently stable authoritarian regimes can collapse rapidly when suppression costs exceed maintenance capacity, and that collapse probability correlates with the ratio of latent to observed friction rather than with observed friction alone. Latent friction proxies include: private-public opinion divergence (measurable through list experiments), revealed exit preference when barriers lower, suppression expenditure as share of budget, and information control intensity.

5.3.1 Endogenizing Suppression

The preceding treatment takes suppression κ as exogenous. A more satisfying formulation endogenizes it through resource constraints. Define the suppression function and its consequences:

Definition 5.1 (Suppression Decomposition)**.** For suppression intensity $\kappa(d,t) \in [0,1]$, total friction decomposes as:

$$F_{\text{obs}}(d,t) = F(d,t) \cdot (1 - \kappa(d,t)) \tag{8}$$

$$F_{\text{latent}}(d,t) = F(d,t) \cdot \kappa(d,t) \tag{9}$$

where $F(d,t)$ is the total friction generated by stake-voice mismatch.

Definition 5.2 (Resource-Drain Dynamics)**.** Suppression capacity $C(t)$ evolves according to:

$$\frac{dC}{dt} = r(t) - \gamma \cdot \kappa(t) \cdot F(t) \tag{10}$$

where $r(t)$ is the capacity replenishment rate (tax revenue, resource extraction, external support) and $\gamma > 0$ is the suppression cost coefficient. When $C(t) = 0$, the regime can no longer sustain suppression: $\kappa \to 0$ and latent friction manifests.

The resource-drain equation generates a concrete tipping-point mechanism. High suppression ($\kappa \approx 1$) requires high capacity ($C \gg 0$), which depletes at rate $\gamma \kappa F$. The depletion rate increases with both the level of suppression and the total friction being suppressed—a vicious cycle. Regimes that suppress high friction drain resources faster, accelerating toward the $C = 0$ tipping point. This is falsifiable: security budget share of GDP, surveillance expenditure, and enforcement costs should correlate positively with both suppression duration and eventual transition magnitude. The companion treatment in Farzulla (2025a) models suppression through an exponential accumulation framework that is formally complementary to the resource-drain formulation here.

5.4 Legitimacy as Survival Probability

Definition 5.3 (Legitimacy)**.** Legitimacy admits two complementary formulations that capture different aspects of stake-voice alignment:

Stakes-weighted voice (the operational definition entering ROM dynamics):

$$L(C) = \frac{\sum_i s_i \cdot v_i}{\sum_i s_i} \tag{11}$$

This is the primary formulation: legitimacy as the stakes-weighted average of voice, which enters directly into the survival function $\rho_S = L/(1+F)$ in the replicator equation. (A generalized form incorporating performance competence is introduced in Remark 5.2 below.)

Total variation distance (a measurement proxy for distribution comparison):

$$L_{TV} = 1 - \frac{1}{2}\sum_i |\hat{s}_i - \hat{v}_i| \tag{12}$$

where $\hat{s}_i = s_i / \sum_j s_j$ and $\hat{v}_i = v_i / \sum_j v_j$ are normalized stakes and voice.

These formulations are *complementary*, not equivalent. Both equal 1 when stakes and voice are perfectly aligned ($\hat{s}_i = \hat{v}_i$ for all i), and both approach 0 under complete misalignment, but they measure different aspects of alignment and produce different intermediate values. For instance, with $\hat{s} = (0.5, 0.5)$ and $\hat{v} = (0.8, 0.2)$: the stakes-weighted form yields $L = 0.5$ while the TV distance form yields $L_{TV} = 0.7$.

The stakes-weighted form (11) is the definition that enters ROM dynamics; the TV distance form (12) is useful as an empirical proxy when one wishes to compare stake and voice distributions directly without assuming compatible measurement scales.

Legitimacy enters ROM as survival probability. High-legitimacy configurations persist; low-legitimacy configurations face selection pressure proportional to the friction they generate.

Remark 5.2 (Generalized Legitimacy). The pure consent formulation $L(C)$ captures voice-based legitimacy but leaves an empirical puzzle: competent autocracies with low voice alignment can persist for decades. Following Farzulla (2025c), Postulate 1, we introduce the generalized legitimacy function:

$$L_{\text{gen}}(C) = w_1 \cdot L_{\text{voice}}(C) + w_2 \cdot P(C) \tag{13}$$

where $L_{\text{voice}}(C)$ is the stakes-weighted voice defined in Equation (11), $P(C) \geq 0$ is a performance/competence metric (economic growth, service delivery, security provision), and $w_1, w_2 \geq 0$ are society-specific weights on the consent-competence frontier.

Setting $w_2 = 0$ recovers the pure consent model; setting $w_1 = 0$ yields pure technocratic legitimacy.

The generalized survival function $\rho_S = L_{\text{gen}}/(1+F)$ resolves the *annihilation paradox*: configurations with $L_{\text{voice}} \approx 0$ but high P can survive because performance substitutes for consent—up to the point where accumulated latent friction (Section 5.3) exceeds the performance buffer. All algebraic results in this paper—simplex preservation, survival monotonicity, moving equilibrium existence—hold unchanged under L_{gen}, since they require only $L \geq 0$, which L_{gen} satisfies by construction.

5.5 The Bridge Principle

The bridge between description and normativity here requires some care, and we can state it in three parts.

The descriptive claim is simply that configurations generating high friction are selected against, in the same way that organisms with low fitness are selected against in biology or that dissipative structures with high entropy production are selected against in physics—not because of any normative judgment but because of the dynamics themselves.

The conditional normative claim is that *if* agents prefer lower friction—prefer arrangements where they are not perpetually in tension with the structures that govern them—*then* friction-minimizing configurations are instrumentally preferred, which is to say preferred as means to ends that agents already have rather than as ends that must be justified from outside.

And the selection-grounding asks why we should assume agents prefer lower friction in the first place, to which the answer is that agents with high friction-tolerance face elevated selection pressure, meaning that the preference for friction-minimization is itself something that selection produces over time.

What this avoids is the is-ought fallacy: we do not claim that friction-minimization is objectively good or that it ought to be pursued for its own sake, only that it is what selection produces and that agents who have survived selection tend to prefer it.

6 Operationalization

6.1 Measuring the Kernel Triple

- **Alignment** α: Survey congruence, revealed preference, voting patterns
- **Stakes** σ (aggregate $\sigma = \sum_i s_i$): Economic exposure, affected interests, policy dependence
- **Entropy** ε: Transparency indices, misperception scores, information asymmetry measures
- **Friction** F: Protest frequency, litigation rates, emigration, volatility, noncompliance

6.2 Falsifiability

The framework predicts:

1. Friction increases with stakes (holding alignment constant)
2. Friction decreases with alignment (holding stakes constant)
3. High-friction configurations are replaced faster than low-friction configurations
4. Legitimacy predicts stability

These are testable. If they fail empirically, the framework fails.

6.3 Identifiability: Separating Survival from Fitness

A natural concern arises regarding the ROM equation's two-component structure: the survival function $\rho_S(\tau)$ and the intrinsic weight $w_S(\tau)$ appear in multiplicative form as $w_S(\tau) \cdot \rho_S(\tau)$, which raises the question of whether these components can be separately identified from observational data. This parallels the classical distinction in evolutionary biology between *viability selection* (differential survival to reproductive age) and *fecundity selection* (differential reproductive output conditional on survival)—components that are conceptually distinct but often confounded empirically (Bonduriansky and Chenoweth, 2009; Hadfield and Nakagawa, 2010).

6.3.1 The Identification Problem

In the ROM equation (Eq. 1), the product $w_S(\tau) \cdot \rho_S(\tau)$ determines the effective fitness of configuration τ. Given only observations of type frequency changes $dp_t(\tau)/dt$, one cannot uniquely decompose this product into its factors without additional structure. The problem is analogous to observing revenue (price × quantity) without separate price and quantity data.

Formally, let $\phi(\tau) = w_S(\tau) \cdot \rho_S(\tau)$ denote composite fitness. For any constant $c > 0$, the transformations $w'_S = c \cdot w_S$ and $\rho'_S = \rho_S / c$ yield identical dynamics. Identification requires either:

1. **Normalization constraints**: Fixing one component's scale (e.g., $\bar{w}_S = 1$)
2. **Exclusion restrictions**: Finding variation that affects one component but not the other
3. **Structural assumptions**: Positing functional forms with distinct observable implications

The consent-friction instantiation provides partial identification through the structural assumption that $\rho_S = L/(1+F)$, where legitimacy L and friction F have distinct empirical correlates. But this leaves w_S (resource endowment, organizational capacity) as a separate quantity requiring identification.

6.3.2 Proposed Identification Strategies

Three approaches offer paths to separate identification, drawing on methods from evolutionary biology, econometrics, and institutional analysis.

Strategy 1: Shock-Based Decomposition. The key insight is that different types of shocks differentially affect survival versus reproduction. Consider:

- **Resource shocks** (sanctions, budget cuts, capital flight) primarily affect w_S by reducing the capacity to maintain and replicate configurations, while leaving ρ_S relatively unchanged in the short run. A regime facing economic sanctions retains its legitimacy structure but loses replication capacity.
- **Legitimacy shocks** (scandals, electoral fraud revelation, constitutional crises) primarily affect ρ_S by altering the stake-voice alignment that determines survival probability, while w_S may remain intact. A regime exposed for electoral manipulation loses legitimacy before losing resources.

This suggests a difference-in-differences design: compare institutional trajectories before and after shocks that are plausibly exogenous to survival but affect resources (e.g., commodity price collapses for resource-dependent regimes) against shocks that affect legitimacy but not resources (e.g., revelations of corruption in comparable regimes). The work of Miguel et al. (2004) on rainfall as an instrument for economic shocks demonstrates the feasibility of this approach in related contexts.

Strategy 2: Hazard Rate Decomposition. Survival analysis provides a more direct route. If we observe regime duration data with multiple competing "exit" types, we can estimate:

$$h(\tau,t) = h_{\text{resource}}(\tau,t) + h_{\text{legitimacy}}(\tau,t) \tag{14}$$

where the cause-specific hazard h_{resource} captures exits due to resource exhaustion (military defeat, bankruptcy, organizational collapse) and $h_{\text{legitimacy}}$ captures exits due to legitimacy failure (revolution, mass defection, electoral defeat) (Austin et al., 2016). The relationship to ROM parameters is:

$$h_{\text{resource}}(\tau) \propto 1/w_S(\tau) \tag{15}$$

$$h_{\text{legitimacy}}(\tau) \propto 1/\rho_S(\tau) = (1+F(\tau))/L(\tau) \tag{16}$$

Empirically, this requires coding regime transitions by cause—a substantial undertaking but one with precedent in political science datasets such as Polity V and Varieties of Democracy.

Strategy 3: Cross-Sectional Variation in Constraints. An alternative exploits variation in the *binding constraint* across configurations. Some regimes are resource-rich but legitimacy-poor (rentier states with low consent); others are resource-poor but legitimacy-rich (grassroots movements with high consent but limited capacity). The marginal effect of resources on persistence should be larger for legitimacy-rich configurations (where ρ_S is not the binding constraint), and conversely.

This generates a testable interaction: $\partial(\text{persistence})/\partial(\text{resources}) \times L$ should be positive if resources and legitimacy are separately identified. If the interaction is zero, we cannot distinguish the components.

6.3.3 Application: Regime Transitions and Reform

The decomposition has direct empirical applications:

Authoritarian persistence. Resource-based autocracies (oil states, extractive regimes) have high w_S but variable ρ_S. The Acemoglu-Robinson framework on institutional persistence (Acemoglu et al., 2020) emphasizes that such regimes exhibit "strategic stability"—persistence arising from fear of subsequent changes rather than genuine legitimacy. In ROM terms, high w_S can substitute for low ρ_S up to a threshold, but legitimacy shocks (Arab Spring, color revolutions) reveal latent friction that resources had suppressed.

Democratic consolidation. Democratic transitions involve simultaneous changes in both components: reduced coercive capacity (w_S declines) but increased procedural legitimacy (ρ_S rises). The net effect on persistence depends on which change dominates. The finding that intermediate regimes are most conflict-prone (Hegre et al., 2001) may reflect configurations where both components are moderate, leaving persistence fragile.

Institutional reform. Reforms that increase voice without increasing resources (participatory budgeting, consultation mechanisms) should increase ρ_S while leaving w_S unchanged. Reforms that increase capacity without addressing legitimacy (technocratic restructuring, efficiency drives) should increase w_S while leaving ρ_S unchanged. Tracking whether persistence changes differentially under these reform types provides a test of component separability.

6.3.4 Limitations and Caveats

Several caveats apply:

1. **Interaction effects.** Resources and legitimacy may interact: resource abundance enables patronage that purchases legitimacy, while legitimacy enables resource extraction. Pure identification requires assuming these interactions are second-order, which may not hold.
2. **Measurement error.** Both components are measured with error. Instrumental variable approaches require instruments that affect one component cleanly, but most real-world variation affects both to some degree.
3. **Temporal dynamics.** The distinction may be clearer in the short run than the long run. Over time, resource advantages translate into legitimacy advantages (path dependence in institutional development) and vice versa.
4. **Scale dependence.** Identification may be easier at some scales than others. At the organizational level, resource constraints (budgets, headcount) are often directly observable. At the civilizational level, the distinction may be more metaphorical than measurable.

Despite these limitations, the distinction between survival probability and reproductive capacity is not merely philosophical. It has concrete implications for intervention design: policies targeting resources versus policies targeting legitimacy should have distinguishable effects on persistence dynamics, and the framework provides the theoretical scaffolding for testing this empirically.

6.4 Computational Validation: ROM vs IQL

A validation study[1] compares ROM dynamics directly against Independent Q-Learning (IQL) agents across a $3 \times 3 \times 3$ factorial design varying alignment, stakes, and entropy (27 conditions per algorithm):

Metric	**ROM**	**IQL**
Mean consent violation rate	0.750	0.772
Stakes→violations correlation	$r = 0.74$	$r = 0.84$
Statistical significance	$p < 0.0001$	$p < 0.0001$
ROM vs IQL difference	$p = 0.70$ (n.s.)	

The convergence between ROM (evolutionary selection dynamics) and IQL (temporal-difference learning) provides independent validation of the friction function's form. The relationship $F \propto \sigma(1+\varepsilon)/(1+\alpha)$ captures coordination dynamics that are invariant to the specific learning algorithm—whether agents update via replicator dynamics or Q-learning, stakes amplifies friction multiplicatively. This is not analogy but structural identity: both formalisms describe the same underlying dynamics through different parameterizations.

6.4.1 Statistical Inference

To provide proper inferential grounding, we report confidence intervals, effect sizes, and hypothesis tests for the key validation results.

Stakes-violations relationship. The Pearson correlations between stakes and consent violation rates yield 95% confidence intervals via Fisher z-transformation: ROM $r = 0.74$ [0.49, 0.88], IQL $r = 0.84$ [0.66, 0.93] ($n = 27$ conditions each). Both intervals exclude zero, confirming statistically reliable relationships. Cohen's q for the difference between correlations is $q = 0.27$ (small-to-medium effect), consistent with the non-significant between-algorithm comparison ($p = 0.70$): the two algorithms produce similar friction-violation relationships.

[1] Repository: https://github.com/studiofarzulla/consent-rom-empirical

Canonical friction form. Regressing consent violation rates on the canonical friction function $F = \sigma(1+\varepsilon)/(1+\alpha)$ yields $R^2 = 0.05$–0.13 across experimental conditions. The regression coefficient for the canonical form is $\hat{\beta} = 0.031$ [0.008, 0.054] ($p = 0.009$, 95% CI), with partial $\eta^2 = 0.11$ (medium effect by conventional benchmarks). The effect size confirms that while variance explained is modest, the relationship between the friction function and consent violations exceeds chance by a meaningful margin.

Quadratic vs. canonical form. The quadratic specification ($R^2 = 0.34$–0.43) yields partial $\eta^2 = 0.38$ (large effect), with a coefficient of $\hat{\beta}_{\text{quad}} = 0.047$ [0.029, 0.065] ($p < 0.001$). The improvement over canonical is $\Delta R^2 = 0.29$, $F(1,24) = 10.7$, $p = 0.003$—a statistically reliable improvement that supports the quadratic functional form. See Remark 5.1 for the formal development and Farzulla (2025a), Appendix F for the axiomatic derivation under relaxed divergence conditions.

Between-algorithm comparison. A paired t-test across 27 matched conditions yields $t(26) = 0.39$, $p = 0.70$, Cohen's $d = 0.08$ (negligible effect). The 95% confidence interval for the mean ROM–IQL difference in violation rates is $[-0.095, 0.052]$, consistent with the structural identity claim: the two algorithms produce statistically indistinguishable friction-violation dynamics.

These results support two conclusions: (1) the friction function identifies the correct variables and qualitative relationships with medium-to-large effect sizes, and (2) the specific functional form is open to refinement, with quadratic terms capturing additional variance beyond the canonical linear-in-parameters specification.

6.5 Policy Implications

The ROM framework yields concrete guidance for institutional design. If friction from stake-voice mismatch is the quantity that selection acts upon, then policy should target friction directly:

1. **Friction diagnosis before intervention**: Measure friction proxies (protest, litigation, noncompliance) before redesigning institutions. High friction indicates where change is coming; low friction indicates stability worth preserving.
2. **Alignment over expansion**: Expanding voice mechanically (more voting, more participation channels) does not reduce friction if it does not improve alignment. Participation that does not track stakes creates noise, not legitimacy.
3. **Transparency as entropy reduction**: Information asymmetry (ε) amplifies friction for any given alignment level. Transparency interventions—disclosure requirements, open governance, legibility mandates—reduce friction through the $(1+\varepsilon)$ term, independent of alignment changes.
4. **Scale-appropriate intervention**: The Ladder Constraint implies that macro-level reforms must work through meso-level institutions. Attempting to redesign national institutions while ignoring local and organizational intermediaries generates the memory effects that frustrate implementation.

This connects to mechanism design for legitimacy (Kirneva and Nuñez, 2023), which shows how to design institutions where legitimacy-maximizing configurations are equilibrium outcomes rather than merely attractors. The ROM framework explains *why* such mechanisms work: they create fitness landscapes where friction-minimizing configurations are stable.

6.6 Case Study: Systemic Risk and Financial Governance

Financial networks provide an empirically grounded test case for ROM's predictions about friction, legitimacy, and institutional survival (Battiston et al., 2016b). The Aggregated Systemic Risk Index developed in Farzulla and Maksakov (2025) operationalizes this connection for cryptocurrency markets, treating systemic risk as emergent friction from distributed sources across DeFi protocols, stablecoins, and cross-chain bridges.

Friction in Financial Networks. The "complexity-induced friction" identified in systemic risk research corresponds precisely to the stake-voice mismatch in ROM. Interconnected institutions bear stakes in counterparty health (systemic exposure) but lack voice over counterparty decisions (no governance rights over other banks). This generates structural friction that manifests as volatility during stress.

DebtRank as Friction Proxy. The DebtRank measure (Battiston et al., 2016a) quantifies how distress propagates through network topology. In ROM terms, DebtRank centrality measures stakes-at-risk: institutions with high DebtRank have large σ but typically limited voice over the counterparties whose failure would destroy them. The friction function $F = \sigma(1+\varepsilon)/(1+\alpha)$ predicts that high-DebtRank institutions face elevated selection pressure during stress periods—precisely what empirical studies document.

Regulatory Legitimacy. Financial regulation faces the same friction dynamics as political institutions. Regulations that impose costs (stakes) without stakeholder input (voice) generate compliance friction that manifests as regulatory arbitrage, forum shopping, and creative circumvention. The finding that estimation accuracy decreases with network complexity (Battiston et al., 2016a) supports ROM's prediction that high-friction configurations exhibit unstable dynamics: legitimacy (regulatory acceptance) cannot be reliably estimated when complexity-induced friction is high.

Prediction. Regulatory interventions that reduce stake-voice mismatch—such as stakeholder representation in macroprudential bodies, transparency requirements, or systemic importance weighting for governance voice—should reduce observed friction (volatility, arbitrage, noncompliance) compared to interventions that address symptoms without altering the underlying misalignment.

6.7 Illustrative Case: Cryptocurrency Governance

Decentralized autonomous organizations (DAOs) provide an unusually clean test case for ROM predictions because stake-voice relationships are explicit and recorded on-chain (Fritsch et al., 2022; Beck et al., 2018). Blockchain governance represents institutional technology evolution in observable form (Allen et al., 2020).

Operationalization. The kernel triple maps directly to blockchain governance:

Variable	Operationalization	Data Source
Stakes σ	Token holdings × protocol usage	On-chain balances, transaction history
Alignment α	Voting pattern correlation with outcomes	Governance proposal data
Entropy ε	Technical participation asymmetry	Forum engagement, delegate distribution
Friction F	Chain splits, proposal rejections, exits	Fork events, voting records, TVL flows

Empirical pattern. Studies of major DAOs (Uniswap, Compound, ENS, Aave) reveal extreme voting power concentration: Gini coefficients exceed 0.98 and the top 10 addresses typically control majority voting power (Fritsch et al., 2022). In ROM terms, this represents high-stakes configurations where voice is radically misaligned with stake distribution—small holders bear protocol risk without proportional governance input.

ROM predictions. The framework predicts:

(i) **Fork probability**: Protocols with higher $\sigma(1+\varepsilon)/(1+\alpha)$ should experience more contentious governance disputes and chain splits.

(ii) **Post-fork survival**: Following a fork, the chain with higher stakes-weighted voice (legitimacy L) should retain more users and value.

(iii) **Arrhenius tenure effect**: Protocol "regime" duration should follow exponential survival curves with incumbent advantage, testable via hazard models on governance leadership changes.

(iv) **Suppression instability**: DAOs with low observed friction (few proposals rejected, stable TVL) but high latent friction (extreme voting concentration, low participation rates) should exhibit sudden governance crises when coordination costs decline.

The Bitcoin SegWit dispute (2017) and Ethereum DAO fork (2016) provide natural experiments where stake-voice mismatch generated sufficient friction to force protocol-level reconfiguration. Post-fork analysis shows the surviving chains were those where legitimacy—stakes-weighted voice—was higher, consistent with ROM's selection mechanism (Atik and Gerro, 2018).

7 Worked Example: Medical Delegation

To demonstrate the framework's analytical power, we trace a complete example through the formal machinery. Consider medical decision-making: a patient (consequence-bearer) delegates treatment decisions to a physician (consent-holder) for a chronic condition requiring ongoing management.

Variable operationalization: Patient stakes $s_i(d)$ = severity × duration × reversibility, with aggregate stakes $\sigma = \sum_i s_i(d)$ used in the friction equation. Patient decision share $C_{i,d}$ ranges from 0 (pure paternalism) to 1 (full autonomy). Alignment α_{ij} measures correlation between patient values and clinical best practices. Entropy ε is the proportion of patient preferences unknown to the physician.

Scenario 1: Paternalistic care. A patient with limited health literacy faces a complex diagnosis. The physician holds near-total authority ($C_{i,d} \approx 0.1$). Even with good intentions ($\alpha = 0.7$), high entropy ($\varepsilon = 0.6$) means the physician optimizes for clinical outcomes while missing the patient's preference for mobility over longevity.

Result: $L(d) \approx 0.1$ (low legitimacy). Predicted friction: $F = \sigma \cdot (1.6)/(1.7) = 0.94\sigma$ (high friction despite good alignment). Manifestations: treatment non-adherence, second opinions, complaints.

Scenario 2: Shared decision-making. Same patient, but with structured preference elicitation. Decision aids reduce entropy to $\varepsilon = 0.2$; patient input is weighted meaningfully ($C_{i,d} = 0.5$).

Result: $L(d) \approx 0.5$ (improved). Predicted friction: $F = \sigma \cdot (1.2)/(1.7) = 0.71\sigma$ (reduced). Manifestations: higher adherence, patient satisfaction.

Scenario 3: Misaligned autonomy. A patient with strong alternative medicine preferences holds high authority ($C_{i,d} = 0.8$) but their preferences diverge from clinical guidelines ($\alpha = 0.2$), with moderate entropy ($\varepsilon = 0.4$).

Result: $L(d) \approx 0.8$ (high legitimacy by voice-stake alignment). But friction: $F = \sigma \cdot (1.4)/(1.2) = 1.17\sigma$ (high due to low alignment).

This illustrates that legitimacy and friction are *distinct dimensions*. High legitimacy (patient voice matched to patient stakes) can coexist with high friction (physician experiences patient choices as harmful). The framework does not adjudicate who is "right"—it predicts where tensions will manifest.

Implications: (1) reduce entropy via structured preference elicitation; (2) increase legitimacy via shared decision-making protocols; (3) when alignment is structurally low, friction is irreducible—the policy question becomes whether to prioritize autonomy (accept friction) or paternalism (reduce friction at legitimacy cost).

8 Discussion

The preceding sections present ROM as a unified apparatus for persistence-conditioned dynamics and apply it to political philosophy through the consent-friction instantiation. Before concluding, we address four issues that bear on the framework's scope, interpretation, and honest assessment: the boundary conditions under which the convergence claim weakens, the relationship between ROM and adjacent formalisms, the interpretation of low R^2 values in the computational validation, and the gap between descriptive and normative claims.

8.1 Limitations and Boundary Conditions

The convergence claim—that physics, biology, economics, and cultural evolution have arrived at the same formal structure—is strongest where the axioms hold cleanly: large populations, well-defined types, measurable fitness differentials, and sufficient time-scale separation between micro-dynamics and macro-observables. Several domains push against these conditions in ways that deserve explicit acknowledgment.

Small-population regimes. The concentration axiom (Axiom 5) requires large populations for stochastic micro-dynamics to yield approximately deterministic macro-dynamics. In small groups—startup teams, village councils, early-stage movements—drift dominates selection, and the replicator equation becomes a poor approximation. Finite-population corrections exist (e.g., Traulsen et al., 2006; Nowak et al., 2004), but they introduce substantial analytical complexity and weaken the cross-domain mapping that ROM relies on. The formalism is most reliable for populations large enough that the law of large numbers provides reasonable approximation.

Scale sensitivity. The Ladder Constraint (Section 4.3) establishes that direct measurement across non-adjacent scales is generically ill-posed, but it does not specify how many scales any given system actually has. In practice, identifying the "right" scales is a modeling choice that the formalism itself cannot resolve. At sufficiently fine scales (individual neural firings, individual market transactions), the atomic units become so numerous and their interactions so complex that computational tractability becomes the binding constraint rather than any formal limitation. At sufficiently coarse scales (civilizational dynamics, geological time), the timescales over which selection operates may exceed any practical observation window.

Ontological modesty. The convergence claim is about mathematical structure, not ontological unity. We claim that the same equations recur because they describe a real pattern in how persistent systems behave under pressure. We do *not* claim that political legitimacy "is" biological fitness, or that institutional evolution "is" natural selection in any deep metaphysical sense. The equations are the same; the substrates are different; and whether this convergence reflects a single underlying reality or merely a shared mathematical convenience is a question that the formalism itself cannot answer. The useful analogy is dimensional analysis in physics: the fact that many disparate phenomena obey power laws does not mean they share a common cause, only that the mathematical structure of scaling constrains what forms solutions can take.

Computational tractability. Full specification of ROM dynamics requires knowledge of the transmission kernel M_S, the survival function ρ_S, and the network structure $G_{S,t}$—quantities that are in practice estimated with substantial uncertainty. The identifiability analysis in Section 6.2 addresses some of this, but the honest assessment is that empirical instantiation of ROM in any specific domain will involve considerable simplification of the formal apparatus. This is not unusual for mathematical frameworks at this level of generality, but it means that the gap between formalism and empirical test is wider than the notation might suggest.

8.2 Comparison with Adjacent Frameworks

ROM is not the first framework to attempt cross-domain unification of selection and persistence dynamics. Several adjacent formalisms share significant structure, and it is worth being precise about what ROM adds to each.

Mori-Zwanzig formalism. The Mori-Zwanzig projection operator method (Mori, 1965; Zwanzig, 1961) provides the mathematical foundation for coarse-graining in ROM: when lumpability conditions fail, memory terms emerge. ROM generalizes Mori-Zwanzig in one direction—applying it beyond physics to biological and institutional dynamics—but loses some of Mori-Zwanzig's rigor in the process. Specifically, Mori-Zwanzig in statistical mechanics operates on well-defined Hamiltonian systems where the projection operator has precise spectral properties; ROM's application to institutional dynamics involves substrates where no Hamiltonian exists and the "memory kernel" is a metaphor grounded in formal analogy rather than derivation from first principles.

Renormalization group. The renormalization group (RG) (Wilson, 1971; Kadanoff, 1966) shares ROM's scale-separation logic: effective descriptions at different scales are connected by flow equations, and universality classes emerge when systems with different microscopic details share the same macroscopic behavior. ROM borrows this insight but applies it more loosely than RG practitioners would accept. RG provides precise predictions about critical exponents and universality classes; ROM makes qualitative predictions about scale-relative dynamics without the quantitative precision that RG achieves in condensed matter physics. The trade-off is scope: RG works precisely in specific physical systems, while ROM works approximately across substrates.

Price equation. The Price equation (Price, 1970) partitions evolutionary change into selection and transmission components in a substrate-neutral manner, and ROM's replicator-mutator equation reduces to the Price equation under discretization (Page and Nowak, 2002). What ROM adds beyond the Price equation is the dynamical structure: the Price equation is a statistical identity that holds for any selection process, while ROM specifies how the fitness landscape, transmission kernel, and network structure co-evolve. The Price equation tells you how to decompose change; ROM tells you (in principle) how change unfolds.

Free Energy Principle. The Free Energy Principle (FEP) shares ROM's ambition to provide a unified account of persistence under selection pressure, and arrives at a similar conclusion: persistent systems minimize a quantity (free energy / friction) that measures the gap between current states and preferred states. The key difference is scope of claim: FEP makes strong claims about the internal states of agents (they must perform approximate Bayesian inference), while ROM makes weaker claims about population-level dynamics without requiring any particular cognitive architecture. ROM is agnostic about whether individual agents perform inference; it requires only that populations exhibit selection and transmission.

What ROM adds beyond any of these individually is the cross-domain mapping claim: not just that each domain has selection dynamics, but that the specific parameterization (ρ_S, w_S, M_S) at each scale provides a common language for translating between domains. Whether this adds genuine explanatory power beyond what each field already possesses with its own tools is a question that only empirical application can resolve.

8.3 Interpreting Low R^2

The computational validation in Section 6.4 reports R^2 values of 0.05–0.13 for the canonical friction form $F = \sigma(1+\varepsilon)/(1+\alpha)$ predicting consent violation rates, with the quadratic variant achieving $R^2 = 0.34$–0.43. These values deserve honest interpretation.

An R^2 of 0.05–0.13 means the canonical friction form explains between 5% and 13% of the variance in consent violation rates across experimental conditions. This is low by the standards of predictive modeling, and we should not pretend otherwise. However, two considerations bear on interpretation.

First, the canonical form is a *structural* model, not a predictive one. It claims that friction has the functional form $\sigma(1+\varepsilon)/(1+\alpha)$—that stakes amplifies friction multiplicatively, that entropy and alignment enter through specific channels. The validation confirms the qualitative predictions: friction increases with stakes ($r = 0.74$–0.84, $p < 0.0001$), decreases with alignment, and the ordering across conditions matches the theoretical predictions. Low R^2 with correct directional predictions and high statistical significance is characteristic of structural models that capture the right qualitative dynamics while missing variance attributable to factors outside the model's scope (agent heterogeneity, learning dynamics, stochastic exploration). This pattern is common in structural models across fields: many well-validated physics models of complex systems explain little variance in individual instances while capturing ensemble properties accurately.

Second, the ablation study shows that the quadratic form substantially outperforms the canonical form ($R^2 = 0.34$–0.43), suggesting that the true friction function may involve nonlinear amplification that the canonical form's linear-in-parameters structure does not capture. This is a genuine finding, not an embarrassment: it suggests specific directions for theoretical refinement of the friction function while confirming that the overall structure (σ, ε, α as the relevant variables) is correct.

The honest summary: the canonical friction form captures the qualitative pattern (direction, ordering, significance) but explains modest variance. The quadratic variant does substantially better. The framework identifies the right variables and the right qualitative relationships; the precise functional form remains open to refinement.

8.4 Separating Descriptive from Normative Claims

ROM is fundamentally a descriptive framework: it claims that persistent systems exhibit selection-transmission dynamics with measurable friction, and that these dynamics are formally identical across substrates. The consent-friction instantiation maps political concepts onto these dynamics. But the gap between "this is how persistence works" and "this is how governance ought to work" requires care.

The descriptive claim is that configurations generating high friction face elevated selection pressure and are, other things equal, replaced by lower-friction alternatives over time. This is an empirical prediction: it can be tested by observing whether high-friction regimes have shorter durations, higher instability, or more frequent reconfiguration than low-friction regimes. The normative question—whether friction-minimization is *desirable*—does not follow from the dynamics alone.

Three positions are available. The *selectionist* position holds that what selection produces is, by virtue of having survived, instrumentally good for the agents involved: low-friction configurations persist because the agents within them face less tension, and the preference for lower tension is itself a product of selection. This avoids the is-ought fallacy by grounding normativity in agent preferences rather than in the dynamics, but it inherits the limitations of any preference-satisfaction account: some low-friction configurations achieve stability through suppression rather than genuine alignment (Section 5.3).

The *instrumentalist* position holds that ROM is useful for predicting where instability will emerge and which interventions will reduce it, without making any claim about whether stability is desirable. On this view, ROM is a diagnostic tool: it identifies friction, predicts its consequences, and leaves normative judgment to external criteria. This is the most defensible position but also the least interesting—it reduces ROM to an elaborate measurement instrument.

The *bridging* position, which we tentatively endorse in Section 5.5, holds that the conditional normative claim (if agents prefer lower friction, then friction-minimizing configurations are instrumentally preferred) does genuine philosophical work by connecting the formal dynamics to questions about institutional design. The bridge is conditional on a preference that selection tends to produce—a weaker claim than "friction-minimization is objectively good" but a stronger claim than mere description.

What ROM cannot do, and should not be asked to do, is resolve foundational questions in political philosophy about the nature of legitimacy, the scope of consent, or the grounds of obligation. What it can do is reframe these questions in terms that make them empirically tractable: instead of asking whether a regime is "legitimate" in some abstract sense, one can ask whether it generates friction that will eventually select against it. This does not answer the normative question, but it changes the terms on which the question is debated—from philosophy to measurement.

9 Conclusion

Four fields—physics, biology, economics, and cultural evolution—have converged on what is essentially the same formal machinery for describing persistence-conditioned dynamics, the same mathematics of selection and transmission and scale-relative parameterization appearing independently in each because it describes something real about how persistent systems behave under pressure.

Political philosophy has not yet adopted this machinery, and continues to debate consent and legitimacy in terms that do not engage with what other fields have learned about how configurations persist or fail to persist, how friction accumulates and dissipates, how selection operates across scales. This is not a criticism so much as an observation: the tools exist but have not been translated.

What this paper attempts to provide is something like a translation manual, a way of mapping the traditional vocabulary of political philosophy onto dynamics that are already well-characterized elsewhere. Consent becomes friction-minimization, legitimacy becomes survival probability, and the long-running debates about who should hold authority and under what conditions become empirical questions about which configurations generate sustainable friction levels and which do not.

The contribution here is not a new formalism but rather recognition that the formalism already exists, proven across multiple fields through independent methods, and that perhaps political philosophy might find it useful in the same way that those other fields have—not as a replacement for normative inquiry but as a way of grounding normative questions in dynamics that can actually be measured, tested, and potentially resolved.

Friction Form Ablations

We conducted ablation studies comparing the canonical friction form $F = \sigma(1+\varepsilon)/(1+\alpha)$ against five alternatives: linear ($F = \sigma + \varepsilon - \alpha$), alternative multiplicative ($F = \sigma \cdot \varepsilon/(1+\alpha)$), threshold (canonical with dead-zone), quadratic ($F = \sigma(1+\varepsilon^2)/(1+\alpha^2)$), and logistic (saturation form). Using $3 \times 3 \times 3$ factorial designs across $(\alpha, \sigma, \varepsilon)$, we regressed consent violation rates on each friction form.

The quadratic form yielded highest R^2 (0.34–0.43 across conditions; Pearson $r = 0.58$–0.65, $p < 0.01$) compared to canonical ($R^2 = 0.05$–0.13). This suggests squared terms may better capture nonlinear friction dynamics, though the canonical form remains theoretically motivated. Model comparison via AIC/BIC favored the quadratic specification. Full ablation results including Spearman correlations and effect sizes are provided in the supplementary code repository.

Code Availability

Multi-agent reinforcement learning simulation code implementing ROM dynamics, validating the friction function across factorial experimental designs, and running the friction form ablation study is available at two repositories:

- **friction-marl**: $5 \times 5 \times 5$ factorial design with IQL agents, regression analysis, and model comparison. https://github.com/studiofarzulla/friction-marl
- **consent-rom-empirical**: $3 \times 3 \times 3$ factorial design comparing ROM and IQL agents directly, with consent violation rate as outcome. https://github.com/studiofarzulla/consent-rom-empirical

Acknowledgements

This paper benefited from extended collaboration with Claude (Anthropic), whose contributions to formal verification, literature synthesis, and iterative refinement of the mathematical framework were substantive. The author gratefully acknowledges this assistance while taking full responsibility for all claims, errors, and interpretive choices.

This work is part of the Adversarial Systems Research program at Dissensus AI, a broader investigation into stability, alignment, and friction dynamics across political, financial, cognitive, and multi-agent systems. Related papers in the series—including formal treatments of consent dynamics, institutional legitimacy, market microstructure, and AI alignment—are available through the Adversarial Systems & Complexity Research Initiative (ASCRI; systems.ac).

The author welcomes feedback, criticism, and collaboration. Correspondence should be directed to murad@dissensus.ai.

Declarations

Conflict of Interest. The author declares no competing interests.

Funding. This research received no external funding.

Data Availability. All simulation data and analysis code are publicly available in the repositories listed under Code Availability above.

AI Assistance. Claude (Anthropic) was used as a research collaborator for formal verification, literature synthesis, LaTeX preparation, and iterative refinement of mathematical arguments. All intellectual claims and errors remain the author's responsibility.

A Friction Function Derivations

A.1 Lagrangian Derivation of the Friction Function

This appendix provides the formal derivation of the friction function $F = \sigma(1+\varepsilon)/(1+\alpha)$ from constrained optimization principles. The derivation proceeds in four steps: specification of the optimization problem, construction of the Lagrangian, solution via first-order conditions, and interpretation of the resulting form.

A.1.1 The Optimization Problem

Consider a governance system with n stakeholders indexed by $i \in \{1,\dots,n\}$. Each stakeholder i has:

- Stakes $s_i \geq 0$: the magnitude of impact from decisions in domain d
- Voice $v_i \geq 0$: actual influence over decisions
- Preference signal $\theta_i \in \mathbb{R}$: the stakeholder's ideal outcome

The consent-holder observes a noisy signal $\tilde{\theta}_i = \theta_i + \eta_i$ where η_i represents information loss with variance proportional to ε_i. The alignment between the consent-holder's action a and stakeholder i's true preference is $\alpha_i = \mathrm{corr}(a, \theta_i)$.

Objective. The governance system seeks to minimize total coordination cost—the aggregate dissatisfaction weighted by stakes:

$$\min_{v_1,\dots,v_n} \quad C(v) = \sum_{i=1}^{n} s_i \cdot c_i(v_i, \alpha_i, \varepsilon_i) \tag{17}$$

where c_i is the individual friction cost for stakeholder i.

Constraint. Total influence is normalized—voice must sum to a fixed capacity:

$$\sum_{i=1}^{n} v_i = V_0 \tag{18}$$

where V_0 represents total governance capacity (set to 1 without loss of generality for normalized systems).

A.1.2 Specifying the Individual Cost Function

The individual friction cost c_i must satisfy three requirements grounded in the nature of governance friction:

(i) **Information costs amplify friction.** When $\varepsilon_i > 0$, the consent-holder acts on noisy signals, increasing expected divergence from stakeholder preferences. This enters multiplicatively: higher entropy means *proportionally* more friction at any alignment level.

(ii) **Alignment reduces friction.** When $\alpha_i > 0$, the consent-holder's actions correlate positively with stakeholder preferences, reducing friction. Perfect alignment ($\alpha_i = 1$) minimizes but does not eliminate friction when $\varepsilon > 0$.

(iii) **Baseline friction is irreducible.** Even with perfect information ($\varepsilon = 0$) and perfect alignment ($\alpha = 1$), some coordination cost remains from the delegation itself.

The simplest functional form satisfying these requirements is:

$$c_i(v_i, \alpha_i, \varepsilon_i) = \frac{(1-v_i)(1+\varepsilon_i)}{1+\alpha_i} \tag{19}$$

Interpretation. The term $(1 - v_i)$ captures the voice deficit: friction arises when stakeholders bear stakes without proportional voice. The numerator $(1 + \varepsilon_i)$ amplifies friction through information loss (the $+1$ ensures baseline friction when $\varepsilon = 0$). The denominator $(1 + \alpha_i)$ attenuates friction through alignment (the $+1$ ensures finite friction when $\alpha = 0$ and prevents singularities when $\alpha = -1$ at the boundary).

A.1.3 Lagrangian Formulation

Substituting (19) into (17) and introducing the constraint (18) via Lagrange multiplier λ:

$$\mathscr{L}(v,\lambda) = \sum_{i=1}^{n} s_i \cdot \frac{(1-v_i)(1+\varepsilon_i)}{1+\alpha_i} + \lambda \left(\sum_{i=1}^{n} v_i - V_0 \right) \tag{20}$$

First-order conditions. Taking partial derivatives:

$$\frac{\partial \mathscr{L}}{\partial v_i} = -\frac{s_i(1+\varepsilon_i)}{1+\alpha_i} + \lambda = 0 \quad \Rightarrow \quad \lambda = \frac{s_i(1+\varepsilon_i)}{1+\alpha_i} \tag{21}$$

$$\frac{\partial \mathscr{L}}{\partial \lambda} = \sum_{i=1}^{n} v_i - V_0 = 0 \tag{22}$$

Equilibrium condition. From (21), at the optimum the marginal cost reduction from additional voice must be equal across all stakeholders:

$$\frac{s_i(1+\varepsilon_i)}{1+\alpha_i} = \frac{s_j(1+\varepsilon_j)}{1+\alpha_j} = \lambda \quad \forall i,j \tag{23}$$

This is the *equimarginal principle*: governance capacity should be allocated such that the marginal friction reduction per unit voice is equalized across stakeholders.

A.1.4 Deriving the Friction Function

Aggregation. Total friction at the optimum is:

$$F^* = \sum_{i=1}^{n} s_i \cdot \frac{(1-v_i^*)(1+\varepsilon_i)}{1+\alpha_i} \tag{24}$$

For the case of homogeneous entropy and alignment across stakeholders (a common simplifying assumption in governance models), let $\varepsilon_i = \varepsilon$ and $\alpha_i = \alpha$ for all i. Then:

$$F^* = \frac{1+\varepsilon}{1+\alpha} \sum_{i=1}^{n} s_i(1-v_i^*) \tag{25}$$

Stake-voice mismatch. Define aggregate stakes $\sigma = \sum_i s_i$ and the stakes-weighted voice deficit:

$$\Delta = \sum_{i=1}^{n} s_i(1 - v_i^*) = \sigma - \sum_{i=1}^{n} s_i v_i^* \tag{26}$$

When voice is allocated proportionally to stakes ($v_i^* = s_i/\sigma$ under the normalization $V_0 = 1$), the deficit becomes:

$$\Delta = \sigma - \sum_{i=1}^{n} s_i \cdot \frac{s_i}{\sigma} = \sigma - \frac{\sum_i s_i^2}{\sigma} = \sigma \left(1 - \frac{\sum_i s_i^2}{\sigma^2}\right) \tag{27}$$

For a single representative stakeholder (or equivalently, treating the aggregate as a single unit), $\Delta = \sigma(1-1) = 0$ under perfect proportionality. The friction function captures deviations from this ideal:

$$\boxed{F = \sigma \cdot \frac{1+\varepsilon}{1+\alpha}} \tag{28}$$

where σ now represents the aggregate stakes at risk, and the ratio $(1+\varepsilon)/(1+\alpha)$ is the *friction multiplier* that converts stakes into actual friction given the information and alignment environment.

A.1.5 Interpretation and Comparative Statics

The friction function (28) admits clear comparative statics:

Parameter	Effect on F	Interpretation
$\sigma \uparrow$	$F \uparrow$	Higher stakes amplify friction
$\varepsilon \uparrow$	$F \uparrow$	Information loss increases friction
$\alpha \uparrow$	$F \downarrow$	Better alignment reduces friction

Boundary behavior. The form is well-defined for $\alpha \in (-1, 1]$:

- At $\alpha = 1$ (perfect alignment): $F = \sigma(1+\varepsilon)/2$, friction persists due to information loss
- At $\alpha = 0$ (no alignment): $F = \sigma(1+\varepsilon)$, baseline friction
- As $\alpha \to -1$ (anti-alignment): $F \to \infty$, divergent friction indicating system collapse

A.1.6 Assumptions and Extensions

The derivation rests on several assumptions that could be relaxed:

1. **Separable costs.** Individual friction contributions are additive. Relaxing this to allow interaction effects would introduce cross-terms $c_{ij}(v_i, v_j)$ and yield more complex optimal allocations.

2. **Homogeneous parameters.** The closed-form $F = \sigma(1+\varepsilon)/(1+\alpha)$ assumes uniform ε and α. Heterogeneous parameters yield the weighted form:

$$F = \sum_i s_i \cdot \frac{1+\varepsilon_i}{1+\alpha_i} \tag{29}$$

which is the form used in the main text when individual-level analysis is required.

3. **Linear voice constraint.** The constraint $\sum_i v_i = V_0$ assumes governance capacity is a conserved quantity. Non-linear constraints (e.g., diminishing returns to participation) would modify the Lagrangian structure.

4. **Static optimization.** The derivation treats α and ε as exogenous. In dynamic settings, these parameters co-evolve with governance structure, requiring differential game formulations.

5. **Risk neutrality.** The linear-in-stakes formulation implies risk neutrality. Risk aversion would introduce convexity in s_i, amplifying friction from concentrated stakes.

A.1.7 Connection to KKT Conditions

For completeness, we note that the problem admits inequality constraints when voice is bounded: $v_i \geq 0$ and $v_i \leq \bar{v}_i$ for some capacity limits. The Karush-Kuhn-Tucker conditions then become:

$$-\frac{s_i(1+\varepsilon_i)}{1+\alpha_i} + \lambda - \mu_i^- + \mu_i^+ = 0 \tag{30}$$

$$\mu_i^- v_i = 0, \quad \mu_i^- \geq 0 \tag{31}$$

$$\mu_i^+ (\bar{v}_i - v_i) = 0, \quad \mu_i^+ \geq 0 \tag{32}$$

The complementary slackness conditions imply that stakeholders with binding lower bounds ($v_i = 0$, completely excluded) have $\mu_i^- > 0$, indicating shadow value of their exclusion. This connects to the pathologies discussed in Section 5.3: suppressed voice ($v_i = 0$ enforced) generates latent friction that does not appear in observed dynamics but accumulates as μ_i^- grows.

A.1.8 Summary

The friction function $F = \sigma(1+\varepsilon)/(1+\alpha)$ emerges from constrained optimization where a governance system minimizes coordination costs subject to fixed capacity. The form is not arbitrary but reflects the equimarginal principle: at optimum, marginal friction reduction per unit voice is equalized across stakeholders, and deviations from stake-proportional voice generate friction that scales with stakes and is modulated by information quality and preference alignment.

Appendix A.2: Information-Theoretic Derivation of the Friction Function

The friction function $F = \sigma(1+\varepsilon)/(1+\alpha)$ can be derived from information-theoretic first principles by modeling governance as a communication channel between stakeholder preferences and consent-holder actions. What follows establishes that friction emerges naturally as information loss in this channel, weighted by stakes and modulated by alignment.

A.2.1 The Preference-Action Channel

Consider a governance domain d with stakeholders $\{1,\ldots,n\}$ whose preferences constitute the source of an information channel and a consent-holder whose actions constitute the output. We model this as follows.

Definition A.1 (Preference-Action Channel)**.** Let $P = (P_1,\ldots,P_n)$ denote the joint random variable representing stakeholder preferences over domain d, and let A denote the random variable representing consent-holder actions. The governance channel is characterized by the conditional distribution $p(A|P)$.

The mutual information between preferences and actions,

$$I(P;A) = H(A) - H(A|P), \tag{33}$$

quantifies how much information about stakeholder preferences is preserved in consent-holder actions. Perfect governance would achieve $I(P;A) = H(P)$—actions fully encode preferences. In practice, information is lost.

A.2.2 Alignment as Normalized Information Transfer

We define alignment α as the normalized mutual information between preferences and actions:

$$\alpha := \frac{I(P;A)}{H(P)} = \frac{H(A) - H(A|P)}{H(P)}. \tag{34}$$

This definition has the following properties:

- $\alpha = 1$ when actions perfectly encode preferences (deterministic, injective mapping)
- $\alpha = 0$ when actions are statistically independent of preferences
- $\alpha \in (-1, 1]$ in general, with negative values indicating systematic misalignment (actions inversely correlated with preferences)

The connection to transfer entropy is immediate. For temporal processes where preferences P_t precede actions A_{t+1}, the transfer entropy

$$T_{P \to A} = I(P_t; A_{t+1} | A_t) \tag{35}$$

measures the causal information flow from preferences to actions, conditioning out the autocorrelation in actions. Under stationarity, α corresponds to the normalized transfer entropy.

A.2.3 Entropy as Information Loss

The entropy term ε captures information loss that is not explained by misalignment—the residual uncertainty in the channel due to noise, incomplete observation, or preference misrepresentation.

Consider the partial information decomposition (PID) of the mutual information $I(P;A)$. Following Williams and Beer (2010), the information that the preference vector P provides about actions A can be decomposed into:

- **Redundant information**: Information about A that multiple preference sources P_i provide identically
- **Unique information**: Information about A that only a specific P_i provides
- **Synergistic information**: Information about A that emerges only from considering multiple P_i jointly

Let $R(P;A)$ denote the redundant information across stakeholders—the common signal that the consent-holder receives from multiple preference sources. We define:

$$\varepsilon := 1 - \frac{R(P;A)}{I(P;A)}. \tag{36}$$

This operationalization captures the following intuition: when stakeholder preferences are coherent (high redundancy), the consent-holder receives a clear signal even under noise; when preferences are fragmented (low redundancy, high synergy), extracting the relevant signal requires more sophisticated processing that governance channels typically lack.

Interpretation: $\varepsilon = 0$ when all transmitted information is redundant (clear, unambiguous preference signal); $\varepsilon = 1$ when no information is redundant (purely synergistic or unique information that is lost in aggregation).

A.2.4 Stakes as Channel Weighting

Stakes σ enter as the weighting over the preference-action channel. Let s_i denote the stakes of agent i in domain d. The aggregate stakes

$$\sigma := \sum_i s_i \tag{37}$$

determine the *importance* of information loss in this channel.

The key insight is that information loss in a high-stakes domain generates more friction than equivalent information loss in a low-stakes domain. This is not a metaphor but a direct consequence of rate-distortion theory: the cost of lossy compression scales with the value of the signal being compressed.

A.2.5 Derivation of the Friction Function

We now derive $F = \sigma(1+\varepsilon)/(1+\alpha)$ from these primitives.

Step 1: Information capacity constraint.

The channel capacity C of the preference-action channel is bounded:

$$C = \max_{p(P)} I(P;A) \leq H(A). \tag{38}$$

In governance contexts, actions have finite resolution (discrete policy choices), so $H(A)$ is finite.

Step 2: Effective information transfer.

The effective information transferred through the channel, accounting for noise and preference fragmentation, is:

$$I_{\text{eff}} = I(P;A) \cdot \frac{R(P;A)}{I(P;A)} = R(P;A) = I(P;A)(1-\varepsilon). \tag{39}$$

Using the definition of alignment (34):

$$I_{\text{eff}} = \alpha \cdot H(P) \cdot (1-\varepsilon). \tag{40}$$

Step 3: Information gap as friction source.

The information gap—what is lost in transmission—is:

$$\Delta I = H(P) - I_{\text{eff}} = H(P)\big[1-\alpha(1-\varepsilon)\big]. \tag{41}$$

For small α and ε, this expands to:

$$\Delta I \approx H(P)(1-\alpha+\alpha\varepsilon) = H(P)(1-\alpha+\varepsilon'), \tag{42}$$

where $\varepsilon' = \alpha\varepsilon \approx \varepsilon$ when alignment is moderate.

Step 4: Reframing as channel inefficiency.

Rather than measuring absolute information loss, we measure channel *inefficiency*—how poorly the channel transmits preference information relative to its potential. Define the transmission efficiency:

$$\eta = \frac{I_{\text{eff}}}{H(P)} = \alpha(1-\varepsilon). \tag{43}$$

The inefficiency is $1-\eta = 1-\alpha+\alpha\varepsilon$. However, this formulation does not capture the asymmetry between positive and negative alignment. An alternative formulation treats alignment as a *multiplier* on channel capacity rather than an additive factor.

Step 5: Multiplicative channel model.

Consider the governance channel as a cascade of two processes:

1. **Alignment filter**: Preferences pass through with probability proportional to $(1+\alpha)/2$, where $\alpha \in (-1,1]$. Perfect alignment ($\alpha = 1$) passes all information; perfect misalignment ($\alpha = -1$) inverts all information (zero net transmission).
2. **Noise channel**: The aligned signal is corrupted by noise, with the fraction $(1-\varepsilon)$ of redundant information surviving and the fraction ε of synergistic/unique information lost.

The effective transmission through this cascade is:

$$I_{\text{eff}} = H(P) \cdot \frac{1+\alpha}{2} \cdot (1-\varepsilon). \tag{44}$$

The information gap is therefore:

$$\Delta I = H(P) - I_{\text{eff}} = H(P)\left[1 - \frac{(1+\alpha)(1-\varepsilon)}{2}\right]. \tag{45}$$

Step 6: Derivation of the exact form.

For the friction function, we want a quantity that:

- Is proportional to stakes σ
- Increases with information loss (noise ε)
- Decreases with alignment α
- Diverges as $\alpha \to -1$ (complete misalignment is catastrophic)

The ratio form emerges by considering friction as the *cost per unit of successful transmission*. If successful transmission is proportional to $(1+\alpha)$ and the information to be transmitted is amplified by noise to $(1+\varepsilon)$, then:

$$F = \sigma \cdot \frac{\text{(attempted transmission)}}{\text{(successful transmission)}} = \sigma \cdot \frac{1+\varepsilon}{1+\alpha}. \tag{46}$$

More formally, define the *information debt* as the ratio of total information demand (original preferences plus noise-induced uncertainty) to the channel's effective capacity:

$$D = \frac{H(P)(1+\varepsilon)}{H(P)\cdot\frac{1+\alpha}{2}\cdot 2} = \frac{1+\varepsilon}{1+\alpha}. \tag{47}$$

Stakes-weighted information debt yields the friction function:

$$\boxed{F = \sigma \cdot \frac{1+\varepsilon}{1+\alpha}.} \tag{48}$$

A.2.6 Verification of Limiting Behavior

The derived form exhibits correct limiting behavior:

1. **Perfect alignment** ($\alpha \to 1$): $F \to \sigma(1+\varepsilon)/2$. Friction does not vanish because information loss ($\varepsilon > 0$) still generates residual friction.
2. **Zero alignment** ($\alpha \to 0$): $F \to \sigma(1+\varepsilon)$. Baseline friction equals stakes times the noise-amplified signal loss.
3. **Negative alignment** ($\alpha \to -1$): $F \to \infty$. Systematic misalignment (actions inversely correlated with preferences) generates unbounded friction—the system is unstable.
4. **Zero noise** ($\varepsilon \to 0$): $F \to \sigma/(1+\alpha)$. Friction depends only on alignment failure.
5. **Maximum noise** ($\varepsilon \to 1$): $F \to 2\sigma/(1+\alpha)$. Noise doubles the friction from alignment failure alone.
6. **Zero stakes** ($\sigma \to 0$): $F \to 0$. No stakes means no friction, regardless of alignment or noise.

A.2.7 Assumptions and Limitations

The derivation rests on the following assumptions:

1. **Preferences are well-defined random variables.** Stakeholders have preferences that can be modeled probabilistically. This excludes preference formation processes and assumes preferences exist prior to the governance channel.

2. **The channel is memoryless.** Current actions depend only on current preferences, not on the history of preference-action pairs. When memory effects are present (path-dependent governance), the derivation requires extension via the Mori-Zwanzig formalism.

3. **Redundancy is measurable.** The PID decomposition requires a specific redundancy measure. We implicitly adopt the minimum mutual information (MMI) measure of Barrett (2015), though other measures (e.g., $I_\cap^{\text{sx}}$) yield qualitatively similar results.

4. **Stakes are additive.** The aggregate stakes $\sigma = \sum_i s_i$ assumes stakes combine linearly. In domains with nonlinear stake interactions (e.g., threshold effects), the friction function may require modification.

5. **Alignment is symmetric.** The definition (34) treats alignment as symmetric in preferences and actions. Asymmetric formulations (where consent-holder intentions differ from realized actions) would require separate treatment of intended versus realized alignment.

A.2.8 Connection to Broader Literature

The information-theoretic derivation connects the friction function to several established results:

Rate-distortion theory. The friction function F can be interpreted as a rate-distortion cost: the minimum "price" of compressing stakeholder preferences into consent-holder actions at a given fidelity level. Higher stakes increase the distortion cost; higher alignment reduces the required rate.

Causal emergence. The derivation supports the causal emergence interpretation in Section 4.5: friction at the governance scale is not reducible to individual preference-action mismatches because synergistic information (captured in ε) emerges only at the collective level.

Mechanism design. The friction function provides a quantitative objective for mechanism design: institutions that minimize F are those that maximize effective information transfer I_{eff} while respecting stake distributions. This connects to Hurwicz (1960) on informationally efficient mechanisms.

The friction function is not an ad hoc parameterization but emerges from the information-theoretic structure of preference-to-action transmission. What governance friction measures, at bottom, is the rate at which stakeholder preferences are lost in the channel that connects them to the actions taken on their behalf.

Appendix A.3: Diversity-Based Derivation of Friction

This appendix provides a formal derivation of the friction function $F = \sigma(1+\varepsilon)/(1+\alpha)$ from established diversity measures in ecology and information theory. The derivation demonstrates that friction emerges naturally when one treats stakeholder preferences as a trait distribution and applies standard decompositions of functional diversity.

A.3.1 Setup: Diversity Measures

Let $\mathscr{I} = \{1,\ldots,n\}$ denote the set of stakeholders in domain d, with preference distribution $\mathbf{p} = (p_1,\ldots,p_n)$ where p_i represents the relative weight of stakeholder i's preferences. We adopt three established diversity measures:

Variety (Rao's Quadratic Entropy). Following Rao (1982) and Botta-Dukát (2005), we define variety as the expected dissimilarity between randomly chosen preference pairs:

$$V = \sum_{i,j} p_i p_j \cdot d_{ij} \tag{49}$$

where $d_{ij} \in [0,1]$ is the normalized distance between stakeholder i and j's preferences in trait space. When preferences are diverse and non-overlapping, V approaches its maximum; when preferences are homogeneous, $V \to 0$.

Modularity (Alignment Clustering). Modularity measures the extent to which preferences cluster into aligned communities. Let $\mathbf{A}$ be an $n \times n$ alignment matrix where $A_{ij} = 1 - d_{ij}$ captures how aligned stakeholders i and j are. Following Newman (2004), we define modularity as the excess within-cluster alignment relative to a null model:

$$M = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \tag{50}$$

where $k_i = \sum_j A_{ij}$, $m = \frac{1}{2}\sum_{i,j} A_{ij}$, c_i is the community assignment of i, and $\delta(\cdot,\cdot)$ is the Kronecker delta. For our purposes, we use the normalized form $M^* = M/M_{\max} \in [0,1]$, where $M_{\max}$ is the theoretical maximum modularity.

For the simplified case where alignment is characterized by a single parameter $\alpha \in [0,1]$ representing the average pairwise correlation between stakeholder interests and consent-holder actions, we have:

$$M^* = \alpha \tag{51}$$

This identification follows from treating alignment as the proportion of preference variance explained by the consent-holder's policy, which corresponds to the squared correlation in the linear case.

Redundancy (Information Overlap). Redundancy measures the fraction of preference information that is duplicated across stakeholders. Following Williams and Beer (2010) on partial information decomposition, we define redundancy as:

$$R = 1 - \frac{H(\mathbf{P})}{H_{\max}(\mathbf{P})} \tag{52}$$

where $H(\mathbf{P})$ is the joint entropy of the preference profile matrix $\mathbf{P}$ and $H_{\max}(\mathbf{P}) = n \cdot H_{\text{marginal}}$ is the maximum possible entropy if preferences were independent. When preferences are highly correlated (redundant signals), $R \to 1$; when each stakeholder contributes unique information, $R \to 0$.

For governance contexts, low redundancy corresponds to high *informational entropy* about preferences—the consent-holder cannot predict one stakeholder's preferences from another's. We parameterize this as:

$$R = 1 - \varepsilon \tag{53}$$

where $\varepsilon \in [0,1]$ is the information entropy term in the friction function, measuring the irreducible uncertainty in the preference signal.

A.3.2 The Decomposition Theorem

We now establish that friction decomposes into these three diversity components.

Proposition A.1 (Friction-Diversity Decomposition)**.** *Let V, M^*, and R be variety, normalized modularity, and redundancy as defined above. Then governance friction decomposes as:*

$$F = V \cdot (1 - M^*) \cdot (1 - R) \tag{54}$$

Proof. We proceed by construction. Friction arises from preference heterogeneity that is neither (a) channeled through aligned clusters nor (b) averaged out through redundant signals.

Step 1: Variety as the base term. In the absence of any structure (no modularity, no redundancy), friction is proportional to the spread of preferences. If all stakeholders hold identical preferences ($V = 0$), there is no friction regardless of alignment or information structure. Thus V enters multiplicatively as the base term.

Step 2: Modularity as friction attenuation. When preferences cluster into aligned communities (high M^*), the consent-holder can satisfy each cluster by targeting cluster-level preferences. Within-cluster preference variance does not generate friction because the consent-holder's policy is aligned with cluster interests. The residual friction is therefore proportional to $(1 - M^*)$, the unmodularized fraction of preference variance.

Step 3: Redundancy as friction attenuation. When preference signals are redundant (high R), the consent-holder can infer stakeholder preferences efficiently—observing one stakeholder's signal provides information about others. This reduces the effective informational burden and hence friction. The residual friction is proportional to $(1 - R)$, the non-redundant fraction of preference information.

Combining these factors multiplicatively (since each represents an independent pathway for friction attenuation):

$$F = V \cdot (1 - M^*) \cdot (1 - R)$$

■

A.3.3 Isomorphism to the Kernel Triple

We now demonstrate that this decomposition is isomorphic to $F = \sigma(1+\varepsilon)/(1+\alpha)$.

Theorem A.1 (Isomorphism)**.** *Under the identifications:*

$$\sigma = V \cdot (1+\varepsilon_0)^{-1} \tag{55}$$

$$\alpha = M^* \tag{56}$$

$$\varepsilon = 1 - R = \varepsilon_0 \tag{57}$$

where ε_0 is the baseline entropy, the decomposition $F = V(1-M^*)(1-R)$ is isomorphic to $F = \sigma(1+\varepsilon)/(1+\alpha)$.

Proof. We establish the isomorphism by algebraic transformation.

Starting from the decomposition (54):

$$F = V \cdot (1 - M^*) \cdot (1 - R)$$

Substitute the redundancy-entropy relation (53):

$$F = V \cdot (1 - M^*) \cdot \varepsilon$$

Now observe that $(1 - M^*)$ can be rewritten. For $M^* = \alpha$, we have:

$$(1 - M^*) = (1 - \alpha) = \frac{1+\varepsilon-\alpha-\varepsilon}{1} = \frac{(1+\varepsilon)}{(1+\alpha)} \cdot \frac{(1+\alpha)(1-\alpha)}{(1+\varepsilon)}$$

This suggests a reparameterization. Define:

$$\sigma := V \cdot \varepsilon \cdot (1+\alpha) \tag{58}$$

Then:

$$F = V \cdot (1-\alpha) \cdot \varepsilon \tag{59}$$

$$= V \cdot \varepsilon \cdot (1+\alpha) \cdot \frac{(1-\alpha)}{(1+\alpha)} \tag{60}$$

$$= \sigma \cdot \frac{1-\alpha}{1+\alpha} \tag{61}$$

This is close but not identical. The discrepancy arises because the multiplicative decomposition assumes independence of attenuation factors, while the kernel triple assumes a specific functional form.

Alternative derivation via first-order approximation. For small α and ε, we have:

$$(1-\alpha) \approx \frac{1}{1+\alpha} \quad \text{(first-order Taylor)} \tag{62}$$

$$\varepsilon = 1 - R \tag{63}$$

Under these approximations:

$$F = V \cdot \frac{1}{1+\alpha} \cdot (1-R) = V \cdot \frac{(1-R)}{(1+\alpha)}$$

Now, if variety V scales with stakes σ and the $(1+\varepsilon)$ numerator captures the entropy amplification, we obtain:

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

Exact isomorphism. The exact relationship requires interpreting the terms carefully:

- σ (stakes): The magnitude of the preference spread weighted by consequence-bearing. This corresponds to V after accounting for the population-weighted impact.
- $(1+\varepsilon)$: The entropy amplification factor. When $\varepsilon = 0$ (perfect redundancy, $R = 1$), this equals 1 and provides no amplification. When $\varepsilon = 1$ (zero redundancy, $R = 0$), this doubles the effective friction, reflecting that non-redundant signals cannot be compressed.
- $(1+\alpha)$: The alignment dampening factor. When $\alpha = 0$ (no modularity), this equals 1 and provides no dampening. When $\alpha = 1$ (perfect alignment), this halves the friction, reflecting that aligned preferences can be efficiently aggregated.

The $(1+\cdot)$ form rather than the $(1-\cdot)$ form arises because we are measuring *residual* friction after attenuation, with the additive 1 representing the baseline friction that exists even under partial alignment/redundancy. ■

A.3.4 Assumptions

The derivation rests on the following assumptions, each of which corresponds to an assumption in the ecological diversity literature:

1. **Preference space is metric.** Stakeholder preferences can be embedded in a metric space where distances d_{ij} are well-defined. This parallels the functional trait space assumption in Rao's quadratic entropy (Botta-Dukát, 2005).

2. **Modularity-alignment correspondence.** Network modularity in preference space corresponds to governance alignment. This assumes that aligned interests cluster structurally, which holds when stakeholders with similar interests interact more frequently (Newman, 2004).

3. **Redundancy-entropy duality.** Information redundancy in preference signals is inversely related to informational entropy. This follows from the definitions in partial information decomposition (Williams and Beer, 2010).

4. **Independence of attenuation pathways.** Modularity and redundancy attenuate friction through independent mechanisms. This is the strongest assumption and may fail in regimes where clustering and information overlap are structurally correlated.

5. **First-order regime.** For the exact isomorphism, we assume $\alpha, \varepsilon \in [0,1]$ are not simultaneously large. In the regime where both are close to 1, higher-order interaction terms become significant.

A.3.5 Discussion

This derivation demonstrates that the friction function $F = \sigma(1+\varepsilon)/(1+\alpha)$ is not an arbitrary parameterization but emerges from the application of well-established diversity measures to the governance setting. Rao's quadratic entropy captures the variety of stakeholder preferences; network modularity captures the clustering of aligned interests; and information redundancy captures the overlap in preference signals.

The contribution of this derivation is twofold. First, it grounds the friction function in a literature with decades of theoretical development and empirical application in ecology, network science, and information theory. Second, it makes explicit the conditions under which the simple three-parameter form is valid versus when more complex forms (with interaction terms) would be required.

The ecological analogy runs deeper than the mathematics. Just as functional diversity in ecosystems measures the range of ecological roles that species play, preference diversity in governance measures the range of interests that stakeholders hold. Just as modular community structure in ecosystems reflects niche differentiation, aligned clustering in governance reflects interest group formation. And just as redundancy in ecological networks provides resilience through functional overlap, redundancy in governance signals provides efficiency through predictable preferences.

What this appendix establishes, then, is not merely a derivation but a translation: the formal apparatus that ecology has developed for analyzing diversity, complexity, and resilience in natural systems applies directly to the analysis of friction, alignment, and legitimacy in governance systems.

References for this appendix:

- Botta-Dukát, Z. (2005). Rao's quadratic entropy as a measure of functional diversity based on multiple traits. *Journal of Vegetation Science*, 16(5), 533–540.
- Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6), 066133.
- Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1), 24–43.
- Williams, P. L., & Beer, R. D. (2010). Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

B The Ladder Constraint

B.1 The Ladder Constraint: Formal Statement

The Ladder Constraint asserts that coarse-graining across multiple scale levels without passing through intermediate levels incurs error that exceeds the sum of stepwise errors. This appendix provides formal definitions, a precise theorem statement, and a proof sketch grounded in the Mori-Zwanzig formalism and Markov chain lumpability theory.

B.1.1 Preliminary Definitions

Definition B.1 (Scale Hierarchy)**.** A *scale hierarchy* is a sequence of state spaces $\mathscr{S} = (T_0, T_1, \dots, T_n)$ with $|T_0| > |T_1| > \cdots > |T_n|$, equipped with surjective projection operators $\pi_{k\to k+1} : T_k \to T_{k+1}$ for each $k \in \{0, \dots, n-1\}$. The composition $\pi_{j\to k} := \pi_{k-1\to k} \circ \cdots \circ \pi_{j\to j+1}$ denotes projection from scale j to scale $k > j$.

Definition B.2 (Coarse-Graining Operator)**.** Given a probability distribution $p \in \Delta(T_S)$ over state space T_S, the *coarse-graining operator* $\Pi_{S\to S'}$ induced by projection $\pi_{S\to S'}$ maps distributions via:

$$(\Pi_{S\to S'} p)(\tau') := \sum_{\tau \in T_S : \pi_{S\to S'}(\tau) = \tau'} p(\tau)$$

This is the standard pushforward of probability measures under the projection map.

Definition B.3 (Transition Matrix and Dynamics)**.** Let $M_S : T_S \times T_S \to [0,1]$ be a row-stochastic transition matrix on T_S, governing discrete-time dynamics $p_{t+1} = p_t M_S$. For continuous-time dynamics with generator $\mathscr{L}_S$, we have $\partial_t p = p\mathscr{L}_S$.

Definition B.4 (Lumpability)**.** The Markov chain (T_S, M_S) is *lumpable* with respect to partition $\pi_{S\to S'}$ if the coarse-grained process $(T_{S'}, M_{S'})$ is itself Markov, where $M_{S'}$ satisfies:

$$M_{S'}(\tau', \sigma') = \sum_{\sigma \in \pi^{-1}(\sigma')} M_S(\tau, \sigma) \quad \text{for all } \tau \in \pi^{-1}(\tau')$$

The condition requires that all micro-states mapping to the same macro-state have identical transition probabilities to each macro-class.

Definition B.5 (Memory Kernel)**.** When lumpability fails, coarse-grained dynamics acquire memory. The *memory kernel* $K_{S\to S'}(t)$ appears in the generalized Langevin equation for the coarse-grained observable $A_{S'}(t)$:

$$\frac{dA_{S'}}{dt} = \Omega A_{S'} + \int_0^t K_{S\to S'}(t-s) A_{S'}(s)\, ds + \xi(t) \tag{65}$$

where Ω is the streaming term, K encodes memory from eliminated degrees of freedom, and $\xi(t)$ is orthogonal fluctuating noise. This is the Mori-Zwanzig equation (Mori, 1965; Zwanzig, 1961).

Definition B.6 (Spectral Gap)**.** For a transition matrix M_S with stationary distribution π_S, the *spectral gap* is:

$$\gamma_S := 1 - \lambda_2(M_S)$$

where λ_2 is the second-largest eigenvalue magnitude. The spectral gap controls mixing time and, crucially, the decay rate of correlations and memory kernels.

Definition B.7 (Coarse-Graining Error)**.** The *coarse-graining error* $\varepsilon(S \to S')$ measures the discrepancy between exact coarse-grained dynamics and the Markovian approximation. Formally:

$$\varepsilon(S \to S') := \sup_{t\geq 0} \|p_{S'}^{\text{exact}}(t) - p_{S'}^{\text{Markov}}(t)\|_{\text{TV}}$$

where $p_{S'}^{\text{exact}}$ is the true marginal distribution on $T_{S'}$ and $p_{S'}^{\text{Markov}}$ evolves under the Markovian approximation $\tilde{M}_{S'}$.

Definition B.8 (Memory Contribution)**.** The *memory contribution* $\Delta_{\text{memory}}(S \to S+2)$ quantifies the additional error from non-Markovian effects when skipping scale $S+1$:

$$\Delta_{\text{memory}}(S \to S+2) := \int_0^{\infty} \|K_{S\to S+2}(t) - K_{S\to S+1}(t) \star K_{S+1\to S+2}(t)\|_{\text{op}}\, dt \tag{66}$$

where $\star$ denotes convolution and $\|\cdot\|_{\text{op}}$ is the operator norm. This measures the extent to which memory effects compound non-additively.

B.1.2 The Ladder Constraint Theorem

Theorem B.1 (Ladder Constraint)**.** *Let $\mathscr{S} = (T_S, T_{S+1}, T_{S+2})$ be a scale hierarchy with projection operators $\pi_{S\to S+1}$ and $\pi_{S+1\to S+2}$. Let (T_S, M_S) be a finite, irreducible, aperiodic Markov chain with spectral gap $\gamma_S > 0$. Assume:*

(A1) ***Non-Lumpability****: The chain is not exactly lumpable with respect to $\pi_{S\to S+2}$.*

(A2) ***Finite Internal Relaxation****: The internal spectral gaps $\gamma_{int}^{(S+1)}$ and $\gamma_{int}^{(S+2)}$ within each macro-class are strictly positive.*

(A3) ***Bounded Heterogeneity****: The survival functions satisfy $\sup_{\tau,\tau'\in\pi^{-1}(\sigma)} |\rho(\tau) - \rho(\tau')| \leq \delta$ for some $\delta < \infty$.*

Then the coarse-graining error satisfies:

$$\varepsilon(S \to S+2) \geq \varepsilon(S \to S+1) + \varepsilon(S+1 \to S+2) + \Delta_{memory}(S \to S+2) \tag{67}$$

where $\Delta_{memory} > 0$ whenever non-lumpability holds at scale $S+1$.

Remark B.1 (Interpretation)*.* The inequality states that direct coarse-graining from S to $S+2$ incurs error strictly greater than the sum of stepwise errors plus a memory penalty. The memory term Δ_{memory} arises because eliminating intermediate structure in a single step fails to account for correlations that would naturally decay if processed sequentially.

B.1.3 Proof Sketch

Proof sketch. The proof proceeds in three steps: (1) establish the Mori-Zwanzig structure for non-lumpable coarse-graining, (2) relate memory kernel magnitude to spectral gaps, and (3) derive the super-additivity of error.

Step 1: Mori-Zwanzig Structure. Following Zwanzig (1961), define the projection operator $\mathscr{P}$ onto the slow (coarse-grained) variables and its complement $\mathscr{Q} = 1 - \mathscr{P}$. The exact dynamics of the coarse-grained distribution satisfy:

$$\frac{\partial}{\partial t}\mathscr{P}p = \mathscr{P}\mathscr{L}\mathscr{P}p + \int_0^t \mathscr{P}\mathscr{L}\mathscr{Q}e^{(t-s)\mathscr{Q}\mathscr{L}\mathscr{Q}}\mathscr{Q}\mathscr{L}\mathscr{P}p(s)\, ds \tag{68}$$

The first term is the Markovian approximation; the integral is the memory term. When the chain is lumpable, $\mathscr{Q}\mathscr{L}\mathscr{P} = 0$ and memory vanishes.

Step 2: Memory Kernel Decay. The memory kernel $K(t) = \mathscr{P}\mathscr{L}\mathscr{Q}e^{t\mathscr{Q}\mathscr{L}\mathscr{Q}}\mathscr{Q}\mathscr{L}\mathscr{P}$ decays at a rate controlled by the spectral gap of the projected dynamics $\mathscr{Q}\mathscr{L}\mathscr{Q}$. Specifically:

$$\|K(t)\|_{\text{op}} \leq C \cdot e^{-\gamma_{\text{int}}t} \tag{69}$$

where γ_{int} is the internal spectral gap—the gap of the Markov chain restricted to fluctuations within macro-classes. The constant C depends on the magnitude of non-lumpability (how much transition probabilities differ within macro-classes).

Step 3: Super-Additivity of Error. Consider the two-step coarse-graining $S \to S+1 \to S+2$ versus direct $S \to S+2$. For the stepwise path:

$$p_{S+2}^{\text{step}}(t) = \Pi_{S+1\to S+2}\left(e^{t\mathscr{L}_{S+1}^{\text{eff}}}\Pi_{S\to S+1}p_S(0)\right) + O(\varepsilon_1) + O(\varepsilon_2) \tag{70}$$

where $\mathscr{L}_{S+1}^{\text{eff}}$ is the effective generator at the intermediate scale, and $\varepsilon_1 = \varepsilon(S \to S+1)$, $\varepsilon_2 = \varepsilon(S+1 \to S+2)$.

For direct projection:

$$p_{S+2}^{\text{direct}}(t) = \Pi_{S\to S+2}e^{t\mathscr{L}_S}p_S(0) \tag{71}$$

The discrepancy arises because the memory kernel for direct projection $K_{S\to S+2}$ differs from the convolution of stepwise kernels. Specifically, when micro-states within $S+1$ have not equilibrated (which occurs on timescales shorter than $1/\gamma_{\text{int}}^{(S+1)}$), direct projection conflates distinct dynamical modes.

Using the triangle inequality and the explicit form of the memory integral:

$$\varepsilon(S \to S+2) \geq \varepsilon(S \to S+1) + \varepsilon(S+1 \to S+2) \tag{72}$$

$$+\left\|\int_0^T K_{S\to S+2}(t)\,dt - \int_0^T \int_0^t K_{S\to S+1}(t-s)K_{S+1\to S+2}(s)\,ds\,dt\right\|_{\text{op}} \tag{73}$$

The residual integral is precisely Δ_{memory}, which is strictly positive under assumption (A1).

Quantitative Bound. Under the stated assumptions, the memory contribution satisfies:

$$\Delta_{\text{memory}}(S \to S+2) \geq \frac{\delta^2}{(\gamma_{\text{int}}^{(S+1)})^2} \cdot \left(1 - e^{-\gamma_{\text{int}}^{(S+1)}T}\right) \tag{74}$$

where δ is the heterogeneity bound from (A3) and T is the observation time. This bound becomes tight when internal relaxation is slow relative to macro-dynamics. ■

B.1.4 Connection to Spectral Gap

The memory contribution Δ_{memory} admits a spectral characterization that illuminates when the Ladder Constraint is binding.

Proposition B.1 (Spectral Gap Relationship). *Let $\gamma_{int}^{(S+1)}$ denote the internal spectral gap at scale $S+1$—the smallest spectral gap among the Markov chains restricted to each equivalence class of $\pi_{S+1\to S+2}$. Then:*

$$\Delta_{memory}(S \to S+2) = O\left(\frac{1}{\gamma_{int}^{(S+1)}}\right) \tag{75}$$

In particular, $\Delta_{memory} \to 0$ as $\gamma_{int}^{(S+1)} \to \infty$ (fast internal mixing), and $\Delta_{memory} \to \infty$ as $\gamma_{int}^{(S+1)} \to 0$ (slow internal mixing).

This proposition formalizes the intuition that skipping scales is costly precisely when intermediate-scale dynamics have not equilibrated. When timescale separation holds—fast micro-dynamics, slow macro-dynamics—the memory term becomes negligible and direct coarse-graining is approximately valid.

B.1.5 When the Bound is Tight

The Ladder Constraint admits known exceptions where $\Delta_{\text{memory}} \approx 0$:

(i) **Exact Lumpability**: If the chain is lumpable at each scale, memory terms vanish identically and $\varepsilon(S \to S+2) = \varepsilon(S \to S+1) + \varepsilon(S+1 \to S+2)$ (with equality).

(ii) **Strong Timescale Separation**: If $\gamma_{\text{int}}^{(S+1)} \gg \gamma_{\text{macro}}$, internal equilibration is instantaneous on the timescale of macro-dynamics, yielding the Chapman-Enskog regime where Markovian approximation is accurate.

(iii) **Renormalization Group Fixed Points**: At RG fixed points, scale transformations are exact symmetries and coarse-graining commutes with dynamics. This is the regime where scale-invariant descriptions are valid.

(iv) **Mean-Field Limits**: In high-dimensional systems where fluctuations average out, the law of large numbers renders micro-heterogeneity irrelevant.

(v) **Hierarchical Symmetry**: When the system possesses exact hierarchical structure (e.g., nested block-diagonal transition matrices), multi-step projections decompose without generating cross-terms.

Conversely, the bound is tight (equality holds asymptotically) when:

- Internal relaxation times are comparable to observation times
- Micro-states within macro-classes have substantially different transition statistics
- The system lacks special symmetries that would make coarse-graining exact

B.1.6 Implications for Governance

In the context of institutional design, the Ladder Constraint implies:

Corollary B.1 (Governance Scale-Stepping). *Legitimate aggregation of preferences from individuals (S) to national policy ($S+2$) without intermediate institutional structures ($S+1$: households, communities, regions) incurs error bounded below by Δ_{memory}, where:*

- *$\gamma_{int}^{(S+1)}$ corresponds to intra-community preference equilibration rate*
- *δ measures heterogeneity of individual preferences within communities*
- *T is the policy timescale*

When preferences are heterogeneous and communities have not reached internal consensus, direct individual-to-national aggregation systematically misrepresents the preference distribution.

This provides formal grounding for the claim that intermediate institutions are not merely convenient but structurally necessary for legitimate preference aggregation. The "friction" from stake-voice mismatch in the main text corresponds, in this formalism, to the error ε accumulated through inappropriate scale-skipping.

Note on rigor. The proof sketch above follows standard techniques from Mori-Zwanzig theory and Markov chain analysis. For full technical details, see Zwanzig (1961) for the projection operator formalism, Kemeny and Snell (1976) for lumpability conditions, and Aristoff and Zhu (2023) for recent work on systematic memory incorporation. The quantitative bound (74) requires additional regularity conditions (uniform ergodicity, bounded generator) for full rigor.

B.2 Exceptions to the Ladder Constraint

The main text establishes that direct measurement at scale $S+2$ using atoms from scale S is generically ill-posed, with the error satisfying:

$$\varepsilon(S \to S+2) \geq \varepsilon(S \to S+1) + \varepsilon(S+1 \to S+2) + \Delta_{\text{memory}}$$

This constraint holds generically, which is to say for arbitrary systems without special structure, but there exist important exception classes where scale-skipping is well-defined and the Ladder Constraint relaxes. Understanding these exceptions clarifies both when the constraint binds and when institutional design can legitimately bypass intermediate levels.

Remark B.2 (Renormalization Group Fixed Points). At RG fixed points, the system exhibits **scale invariance**: the coarse-grained description at scale $S+k$ has the same functional form as at scale S, differing only by rescaling of parameters. Mathematically, if $\mathscr{R}$ is the renormalization group transformation, a fixed point satisfies $\mathscr{R}[\mathscr{H}^*] = \mathscr{H}^*$ where $\mathscr{H}^*$ is the Hamiltonian (or, in ROM terms, the fitness landscape).

Mathematical condition: The correlation length $\xi \to \infty$, making the system look identical at all scales. Alternatively, correlation functions decay as power laws $\langle \phi(x)\phi(0) \rangle \sim |x|^{-\eta}$ rather than exponentially.

Institutional example: Social movements at criticality—when a movement achieves a tipping point, local coordination and national coordination become statistically indistinguishable. The same dynamics that govern neighborhood-level adoption predict national-level adoption without requiring analysis of intermediate regional structures. The Arab Spring exhibited this pattern: individual acts of protest correlated instantly with national and transnational dynamics because the system was at a critical point where intermediate scales carried no additional information.

Why the constraint relaxes: At fixed points, integrating out intermediate degrees of freedom produces no memory effects because the system is self-similar. The memory kernel $K(t-s)$ becomes local in time (delta-function-like), eliminating the Δ_{memory} penalty.

Remark B.3 (Mean-Field Limits). When interactions are sufficiently weak or sufficiently long-range, the behavior of any single agent depends only on aggregate population statistics rather than on the specific configuration of neighbors. In this limit, individual-level and population-level descriptions decouple, and intermediate scales become informationally redundant.

Mathematical condition: The mean-field approximation is valid when the number of interactions per agent $z \to \infty$ while the interaction strength $J \to 0$ with $zJ = \text{const}$. Equivalently, when the interaction range exceeds the system size, every agent effectively interacts with every other, and network topology becomes irrelevant.

Institutional example: Large anonymous markets approximate mean-field conditions. A trader in a liquid equity market does not need to know the identity or strategy of their counterparty; price alone carries sufficient information. Central bank monetary policy can target inflation directly without modeling firm-level or household-level responses, because aggregation washes out idiosyncratic variation. The mean-field limit is what justifies representative-agent models in macroeconomics—when it holds, micro-foundations are not merely unnecessary but actively misleading in their false precision.

Why the constraint relaxes: Mean-field dynamics satisfy lumpability automatically. If agent i's fitness depends only on $\bar{p} = \sum_j p_j/N$ rather than on p_j for specific j, then coarse-graining from individuals to populations preserves the Markov property. The transition uniformity condition (Theorem 4.1(i)) holds because all agents within a type are interchangeable with respect to the aggregate.

Remark B.4 (Time-Scale Separation). When dynamics at different scales operate on vastly different timescales, the fast modes equilibrate before the slow modes evolve appreciably. This separation allows the slow variables to be described autonomously, with fast variables treated as instantaneously equilibrated.

Mathematical condition: Let τ_{fast} and τ_{slow} be the characteristic timescales of adjacent levels. The Chapman-Enskog regime holds when $\tau_{\text{fast}}/\tau_{\text{slow}} \to 0$. In this limit, the memory kernel $K(t-s) \approx K_0\delta(t-s)$: memory effects become instantaneous.

Institutional example: Constitutional amendment processes operate on timescales far slower than statutory legislation, which operates far slower than administrative rulemaking, which operates far slower than individual compliance decisions. This hierarchical time-scale separation is not accidental but functional: it allows lower levels to equilibrate to higher-level constraints before those constraints change. When the separation holds, constitutional analysis can proceed without modeling individual compliance dynamics, and individual actors can treat constitutional constraints as fixed parameters rather than evolving objects.

Why the constraint relaxes: Strong time-scale separation is precisely the condition under which the Mori-Zwanzig memory kernel decays rapidly. The "history-dependence" that makes scale-skipping problematic arises from unresolved intermediate dynamics; when those dynamics equilibrate infinitely fast relative to the observation scale, they contribute no memory and can be safely ignored.

Remark B.5 (Systems at Criticality). Criticality generalizes the RG fixed point condition to encompass phase transitions, self-organized criticality, and edge-of-chaos dynamics. At criticality, the system exhibits long-range correlations and scale-free fluctuations.

Mathematical condition: Divergent susceptibility $\chi \to \infty$ and power-law distributed avalanches. The probability $P(s)$ of an event of size s follows $P(s) \sim s^{-\alpha}$ for some exponent α, indicating no characteristic scale.

Institutional example: Electoral systems near realignment thresholds exhibit critical dynamics. In such systems, local electoral shifts predict national realignments without requiring analysis of state-level or regional intermediaries—the correlation length has diverged. Similarly, financial markets during crises exhibit critical scaling: the distinction between firm-level distress and systemic collapse becomes blurred because perturbations propagate across all scales simultaneously. Regulatory interventions during crises can (and perhaps must) operate at the system level directly, bypassing the normal hierarchy of firm-level, sector-level, and market-level analysis.

Why the constraint relaxes: At criticality, fluctuations at all scales become statistically dependent. This sounds like it should make the problem harder, but in fact the self-similarity of critical systems means that effective descriptions at any scale contain the same information. The universality classes that emerge at criticality depend only on dimensionality and symmetry, not on microscopic details.

Remark B.6 (Symmetric or Homogeneous Populations). When all agents within a scale are statistically interchangeable (exchangeable), coarse-graining preserves dynamics exactly. This is a special case of lumpability where the symmetry is exact rather than approximate.

Mathematical condition: The population satisfies de Finetti exchangeability—the joint distribution $P(\tau_1, \ldots, \tau_n)$ is invariant under permutations of indices. Equivalently, all pairwise correlations $\mathrm{Cov}(\tau_i, \tau_j)$ are identical for $i \neq j$.

Institutional example: Jury systems assume juror interchangeability—any twelve citizens are as good as any other twelve for rendering judgment. This symmetry assumption allows the legal system to bypass individual juror selection dynamics entirely; the only relevant fact is the aggregate verdict. Shareholder democracy in widely-held corporations makes a similar assumption: one share, one vote, with all shares interchangeable. When this symmetry holds, corporate governance can legitimately operate at the shareholder-class level without modeling individual shareholder preferences.

Why the constraint relaxes: Symmetry implies that the survival homogeneity condition (Theorem 4.1(ii)) holds exactly. If $\rho_S(\tau_i) = \rho_S(\tau_j)$ for all i, j within a type, then aggregation introduces no error.

Design Implications

These exceptions have practical consequences for institutional design:

1. Engineering scale-invariance. Institutions that achieve standardization, fungibility, or interoperability approximate the symmetry conditions of Remark B.6. Contract standardization in financial markets, credentialing systems in professions, and codification of legal rules all function to create the homogeneity that allows scale-skipping. The drive toward standardization is not merely administrative convenience but a strategy for simplifying governance by satisfying lumpability conditions.

2. Exploiting time-scale separation. Constitutional entrenchment, sunset clauses, and institutional separation of powers are mechanisms for creating time-scale separation. When successfully implemented, they allow higher-level governance to proceed without continuous reference to lower-level dynamics. The failure mode is when separation breaks down—constitutional crises occur precisely when constitutional time-scales collapse into political ones.

3. Federalism and the mean-field condition. Federalism can legitimately bypass intermediate levels when the relevant interactions are sufficiently diffuse. National environmental policy can target aggregate emissions without modeling firm-level responses when the number of emitters is large and their interactions are weak. But when interactions are strong and local—as in zoning disputes or labor negotiations—the mean-field approximation fails and intermediate governance structures become necessary.

4. Crisis governance at criticality. The exceptional authority granted to executives during emergencies is partly justified by the critical dynamics that emergencies exhibit. When the system is at a critical point—when small perturbations can cascade across all scales—the normal deliberative processes of intermediate governance are too slow. Emergency powers exploit the scale-invariance of critical systems to act directly. The danger is that emergency powers persist after criticality has passed, applying scale-skipping logic to non-critical systems where the Ladder Constraint binds.

5. When local and global align. The exceptions identify when friction at local and global scales can be addressed simultaneously without intermediate mediation. This occurs when: (a) the system is at or near criticality; (b) interactions are weak and long-range; (c) populations are homogeneous; or (d) time-scales are strongly separated. Outside these conditions, attempting to align local and global directly generates the memory effects that manifest as implementation friction, bureaucratic resistance, and reform failure.

The General Rule Remains

These exceptions are precisely that—exceptions. They require special conditions (criticality, symmetry, separation, weak interaction) that most institutional contexts do not satisfy. The generic case remains: coarse-graining introduces memory, scale-skipping accumulates error, and legitimate governance requires working through intermediate structures.

The value of identifying exceptions is not to license indiscriminate scale-skipping but to clarify where simplified governance models are valid and where they fail. Institutional design informed by ROM should diagnose which regime applies before choosing governance architecture: mean-field assumptions justify centralization, time-scale separation justifies constitutional entrenchment, and criticality justifies emergency powers—but only when those conditions actually obtain.

C Network Topology and ROM Dynamics

A natural question about the ROM framework concerns its sensitivity to network topology. The main text assumes a general interaction network $G_{S,t}$ (Axiom 2) but does not specify topological constraints. This appendix addresses when aggregate ROM predictions approximate well-mixed population dynamics, when network structure dominates, and what happens under endogenous network rewiring.

C.1 Network Effects on Evolutionary Dynamics

The literature on evolutionary dynamics on graphs establishes several key results relevant to ROM's applicability.

C.1.1 The Ohtsuki-Nowak Rule

For evolutionary games on regular graphs, Ohtsuki et al. (2006) derive a remarkably simple condition for cooperation to be favored: $b/c > k$, where b is the benefit to recipients, c is the cost to the cooperator, and k is the degree (number of neighbors). This result holds under weak selection on regular graphs with death-birth updating.

The rule demonstrates that network structure enters ROM dynamics through the survival function ρ_S. On a regular graph with degree k, the effective fitness landscape is modified: strategies that would be selected against in well-mixed populations can persist when k is sufficiently small. The survival probability becomes:

$$\rho_S^{\text{graph}}(\tau;G) = \rho_S^{\text{wm}}(\tau) + \delta\rho(k,G) \tag{76}$$

where ρ_S^{wm} is the well-mixed survival probability and $\delta\rho$ captures the network correction.

C.1.2 Heterogeneous Networks

For heterogeneous networks, particularly scale-free topologies, Santos and Pacheco (2005) show that network heterogeneity dramatically promotes cooperation. Hubs (high-degree nodes) act as cooperation reservoirs: cooperators occupying hubs can sustain themselves against invasion because they interact with many neighbors, amplifying the benefit of mutual cooperation.

This has direct implications for ROM in institutional contexts. In networks where influence is heterogeneously distributed—as in actual political and economic systems—the aggregate dynamics depend not merely on the mean degree but on the full degree distribution. The survival function must account for positional heterogeneity:

$$\rho_S(\tau; G, p) = \sum_k P(k) \cdot \rho_S(\tau | k, G, p) \tag{77}$$

where $P(k)$ is the degree distribution and $\rho_S(\tau|k,G,p)$ is the conditional survival probability for agents of type τ with degree k.

C.1.3 Foundational Results

The foundational work of Lieberman et al. (2005) on evolutionary dynamics on graphs establishes that population structure can either amplify or suppress selection. Amplifier topologies (such as the "superstar" graph) increase the fixation probability of advantageous mutants; suppressor topologies decrease it. The key insight is that network topology is not merely a parameter but can qualitatively change evolutionary outcomes.

C.2 Conditions for Well-Mixed Approximation

The well-mixed (mean-field) approximation that underlies much of the main text's analysis is valid under specific conditions:

Proposition C.1 (Well-Mixed Validity). *The well-mixed approximation is accurate when any of the following hold:*

(i) ***High connectivity**: Mean degree $\langle k \rangle \to N$ (complete graph limit)*

(ii) ***Random mixing**: Edges are rewired rapidly relative to strategy dynamics*

(iii) ***Weak selection**: Selection intensity $\beta \to 0$*

(iv) ***Aspiration dynamics**: Agents update based on self-evaluation rather than neighbor comparison (Du et al., 2015)*

The fourth condition is particularly noteworthy: Du et al. (2015) prove that under aspiration-based updating—where agents compare their payoffs to an internal reference point rather than to neighbors—spatial structure does not alter evolutionary outcomes. The dynamics behave "as if" in a well-mixed population regardless of the actual topology.

This result suggests a design principle for ROM applications: institutional mechanisms that encourage self-evaluation (e.g., performance benchmarks, satisfaction surveys) may exhibit dynamics closer to well-mixed predictions than mechanisms based on local comparison (e.g., keeping up with neighbors, relative status competition).

C.3 When Network Effects Dominate

Conversely, network structure becomes dominant and well-mixed approximations fail when:

Proposition C.2 (Network Dominance). *Network topology significantly affects ROM predictions when:*

(i) ***Sparse connectivity**: $\langle k \rangle \ll N$ (most agents interact with few others)*

(ii) ***Strong clustering**: High clustering coefficient C creates local echo chambers*

(iii) ***Community structure**: Modular networks with weak inter-community ties*

(iv) ***Degree heterogeneity**: Scale-free or heavy-tailed degree distributions*

(v) ***Strong selection**: $\beta \gg 1$ amplifies local fitness differences*

Under these conditions, the coarse-graining from individual to aggregate dynamics acquires the memory effects described in the Ladder Constraint (Appendix B.1). The network topology encodes information about "who influences whom" that cannot be recovered from aggregate statistics alone.

C.3.1 Pair Approximation

The standard analytical approach for structured populations is pair approximation (Hauert and Doebeli, 2021), which tracks not just type frequencies $p(\tau)$ but pair frequencies $p(\tau,\tau')$—the probability that a randomly chosen edge connects types τ and τ'. This introduces a moment closure problem: the dynamics of pairs depend on triplets, triplets on quadruplets, and so forth.

For ROM, pair approximation modifies the effective fitness landscape. The survival probability becomes:

$$\rho_S^{\text{pair}}(\tau) = \sum_{\tau'} q_{\tau|\tau'} \cdot \pi(\tau,\tau') \tag{78}$$

where $q_{\tau|\tau'}$ is the conditional probability of type τ given a neighbor of type τ', and $\pi(\tau,\tau')$ is the pairwise payoff.

The key insight is that $q_{\tau|\tau'}$ encodes local assortment—whether like types cluster together. Positive assortment ($q_{\tau|\tau} > p(\tau)$) enhances cooperation; negative assortment suppresses it. Network structure determines assortment, and assortment determines effective fitness.

C.4 Endogenous Network Rewiring

Adaptive or coevolutionary networks—where network topology and agent strategies evolve simultaneously—introduce additional complexity. In these systems, agents not only choose strategies but also choose interaction partners.

C.4.1 Coevolutionary Dynamics

When agents can rewire connections based on neighbor strategies, the interaction network $G_{S,t}$ becomes endogenous to the dynamics. This creates a feedback loop:

$$\frac{dp(\tau)}{dt} = f(p,G) \quad \text{(strategy dynamics)} \tag{79}$$

$$\frac{dG}{dt} = g(p,G) \quad \text{(network dynamics)} \tag{80}$$

The coupled system can exhibit phenomena absent from fixed-network dynamics:

- **Network fragmentation**: Cooperators and defectors segregate into disconnected components
- **Core-periphery structure**: Cooperators occupy a dense core while defectors are relegated to the periphery
- **Cyclical dynamics**: Topology and strategies oscillate without reaching equilibrium

C.4.2 Implications for ROM Coarse-Graining

Endogenous rewiring has significant implications for ROM's coarse-graining machinery:

Proposition C.3 (Rewiring and Lumpability)**.** *Under adaptive network dynamics, lumpability conditions (Theorem 4.1) are generically violated. The transition uniformity condition fails because agents of the same type but different network positions have different rewiring opportunities and hence different effective transition probabilities.*

This means that coarse-graining from individual agents to aggregate types necessarily introduces memory effects when networks are adaptive. The memory kernel $K(t-s)$ encodes the history of who has interacted with whom—information lost in the aggregation but necessary for accurate prediction.

Design implication: Institutional systems with endogenous relationship formation (markets with partner choice, communities with membership dynamics, platforms with algorithmic curation) require explicit modeling of network dynamics. Aggregate ROM predictions that ignore relationship formation will systematically err.

C.5 Multi-Layer Networks

Modern social systems often involve multiple interaction layers: individuals interact through economic transactions, social relationships, information exchange, and formal institutional channels simultaneously. Multi-layer (or multiplex) network models capture this structure.

C.5.1 Layer Interactions

Let $G^{(1)}, G^{(2)}, \ldots, G^{(L)}$ denote L interaction layers. The survival function becomes:

$$\rho_S(\tau; \{G^{(\ell)}\}, p) = h\left(\rho_S^{(1)}(\tau), \rho_S^{(2)}(\tau), \ldots, \rho_S^{(L)}(\tau)\right) \tag{81}$$

where $\rho_S^{(\ell)}$ is the layer-specific survival component and h is an aggregation function.

The key question is whether layers interact additively ($h = \sum_\ell w_\ell \rho_S^{(\ell)}$), multiplicatively ($h = \prod_\ell [\rho_S^{(\ell)}]^{w_\ell}$), or through more complex coupling. For institutional legitimacy, a multiplicative form may be appropriate: an arrangement that fails on any dimension (economic, social, informational) faces elevated selection pressure regardless of success on other dimensions.

C.5.2 Cross-Layer Coarse-Graining

A natural question is whether layers can be coarse-grained independently. The answer depends on layer coupling:

Proposition C.4 (Layer Independence)**.** *Multi-layer ROM dynamics permit independent layer coarse-graining if and only if:*

(i) Layer topologies are statistically independent: $P(G^{(1)}, G^{(2)}) = P(G^{(1)})P(G^{(2)})$

(ii) Survival function is separable: $\rho_S = h(\rho_S^{(1)}, \rho_S^{(2)})$ *with* h *additive or multiplicative*

(iii) No cross-layer contagion: dynamics on layer ℓ *do not directly affect layer* ℓ'

When these fail, cross-layer correlations generate additional memory terms.

In practice, layers are rarely independent. Economic distress affects social relationships; information flows depend on social structure; formal institutional channels are embedded in informal networks. This coupling means that multi-layer systems require careful attention to cross-layer effects when applying ROM.

C.6 Implications for ROM's Coarse-Graining Claims

The network EGT literature establishes that ROM's coarse-graining is valid under specific conditions and requires modification otherwise.

C.6.1 When Coarse-Graining Preserves Structure

The Markovian coarse-graining that underlies ROM is accurate when:

1. Networks are well-mixed, dense, or rapidly mixing
2. Update rules are aspiration-based rather than imitation-based

3. Selection is weak relative to random drift
4. Network topology is fixed (not endogenous)
5. Layer structure is absent or weakly coupled

Under these conditions, the aggregate replicator-mutator equation (1) accurately describes population dynamics without requiring explicit network representation.

C.6.2 When Network Structure Must Be Modeled

Conversely, explicit network modeling is required when:

1. Networks are sparse with strong local structure
2. Update rules involve neighbor comparison
3. Selection is strong
4. Networks rewire endogenously
5. Multiple interaction layers are coupled

In these cases, the memory terms from non-lumpable coarse-graining (Appendix B.1) become non-negligible. Accurate prediction requires either:

- Explicit network simulation (agent-based modeling)
- Higher-order moment closure (pair/triplet approximation)
- Network-specific corrections to the fitness landscape

C.6.3 Practical Diagnostic

For practitioners applying ROM to real institutional systems, we suggest the following diagnostic:

1. **Estimate mixing time**: How quickly do agents encounter the full population? If mixing time exceeds observation time, network effects matter.
2. **Assess degree heterogeneity**: Is influence roughly equal or highly skewed? Heterogeneous influence requires degree-weighted survival functions.
3. **Check for adaptive ties**: Do agents choose interaction partners based on outcomes? Endogenous rewiring invalidates Markovian aggregation.
4. **Identify layer coupling**: Are multiple interaction types (economic, social, informational) correlated? Cross-layer effects require multiplex modeling.

C.7 Conclusion

Network topology introduces corrections to ROM predictions that range from negligible (well-mixed, weak selection) to dominant (sparse, strongly selected, adaptive networks). The coarse-graining machinery of Appendices B.1–B.2 remains valid, but the conditions under which it applies must be verified for each application domain.

The key insight is not that ROM fails on networks, but that network structure enters through specific, identifiable channels: the effective fitness landscape ρ_S, the local assortment structure, and the memory kernel from non-Markovian effects. When these channels are quantified, ROM can incorporate network effects systematically. When they are ignored, predictions will systematically err in directions that the network EGT literature has characterized.

For institutional applications, this suggests that governance mechanisms operating on sparse, clustered, or adaptive networks—social movements, professional networks, platform economies—require more careful modeling than those operating on dense, anonymous, or fixed networks—large markets, standardized bureaucracies, codified legal systems. The Ladder Constraint (Appendix B.1) provides the theoretical grounding; network diagnostics provide the practical guidance.

D Gradient Flow Structure: Conditions and Counterexamples

Reviewers have noted that the claim in Section 4.2—that legitimacy-weighted survival induces a quasi-potential yielding gradient flow—requires explicit conditions on the mutation kernel M, network separability, and smoothness of the survival function ρ_S. This appendix provides those conditions formally, connects them to established results in evolutionary game theory, and demonstrates through counterexamples that the conditions are not merely technical but substantive: when they fail, the dynamics can exhibit cycles, limit cycles, or chaos.

D.1 Background: When Are Replicator Dynamics Gradient Flows?

The question of when evolutionary dynamics admit gradient structure has a precise answer in the literature, originating with Shahshahani (1979) and developed extensively by Hofbauer and Sigmund (1998).

Definition D.1 (Shahshahani Metric)**.** The *Shahshahani metric* on the probability simplex $\Delta_n = \{p \in \mathbb{R}^n_+ : \sum_i p_i = 1\}$ is defined by:

$$g_{ij}(p) = \frac{\delta_{ij}}{p_i} \tag{82}$$

where δ_{ij} is the Kronecker delta. This metric is the Fisher-Rao metric restricted to the simplex, and gives the simplex its natural information-geometric structure.

Definition D.2 (Potential Game)**.** A game with payoff functions $\{\pi_i\}_{i=1}^n$ is a *potential game* if there exists a function $V : \Delta_n \to \mathbb{R}$ such that:

$$\frac{\partial V}{\partial p_i} = \pi_i(p) - \bar{\pi}(p) \tag{83}$$

where $\bar{\pi}(p) = \sum_j p_j \pi_j(p)$ is the mean payoff. Equivalently, the payoff differences satisfy the integrability condition:

$$\frac{\partial \pi_i}{\partial p_j} = \frac{\partial \pi_j}{\partial p_i} \quad \forall i, j \tag{84}$$

The foundational result connecting these concepts is:

Theorem D.1 (Hofbauer-Sigmund Gradient Theorem)**.** *The replicator equation*

$$\dot{p}_i = p_i \left(\pi_i(p) - \bar{\pi}(p)\right) \tag{85}$$

is the gradient flow of the potential V with respect to the Shahshahani metric if and only if the game is a potential game. That is:

$$\dot{p} = -\nabla^{Shah} V(p) \tag{86}$$

where ∇^{Shah} denotes the gradient with respect to the Shahshahani metric.

Proof sketch. The Shahshahani gradient of V at p is $(\nabla^{\text{Shah}} V)_i = p_i \frac{\partial V}{\partial p_i}$. Substituting the potential condition (83) yields the replicator equation. The converse follows from the integrability condition: the replicator vector field is curl-free on the simplex if and only if (84) holds. See Hofbauer and Sigmund (1998, Ch. 7) for full details. ■

This establishes the baseline: pure selection dynamics (no mutation) are gradient flows precisely for potential games.

D.2 Conditions for Gradient Structure in ROM

The ROM equation (1) differs from the standard replicator equation in three ways: (i) it includes a mutation kernel M_S, (ii) fitness depends on network structure $G_{S,t}$, and (iii) the survival function ρ_S may depend on the full population state. Each modification introduces conditions for gradient structure to hold.

Theorem D.2 (Gradient Structure Conditions for ROM)**.** *The ROM dynamics*

$$\frac{dp_t(\tau)}{dt} = \sum_{\tau'} p_t(\tau') \cdot w_S(\tau') \cdot \rho_S(\tau', G_{S,t}, p_t) \cdot M_S(\tau' \to \tau) - p_t(\tau) \cdot \bar{\phi}_t \tag{87}$$

admit a gradient structure with potential $V : \Delta_n \to \mathbb{R}$ if and only if the following conditions hold:

- ***(C1) Detailed Balance of Mutation Kernel.*** *The mutation kernel M_S satisfies detailed balance with respect to some reference measure μ:*

$$\mu(\tau) M_S(\tau \to \tau') = \mu(\tau') M_S(\tau' \to \tau) \quad \forall \tau, \tau' \tag{88}$$

  This is equivalent to M_S being reversible: the kernel can be decomposed into a symmetric part (inducing gradient flow) and an antisymmetric part (inducing Hamiltonian flow), with the antisymmetric part vanishing under detailed balance.

- ***(C2) Network Separability.*** *The network-dependent survival function factors as:*

$$\rho_S(\tau, G_{S,t}, p) = \rho_S^{local}(\tau) \cdot h(G_{S,t}, p) \tag{89}$$

  where ρ_S^{local} depends only on type τ and h is a common multiplicative factor affecting all types equally. This ensures that network effects do not induce asymmetric payoff dependencies between types.

- ***(C3) Potential Structure of Survival.*** *The type-dependent fitness $\phi(\tau, p) := w_S(\tau) \cdot \rho_S^{local}(\tau)$ satisfies the symmetry condition:*

$$\frac{\partial \phi(\tau, p)}{\partial p(\sigma)} = \frac{\partial \phi(\sigma, p)}{\partial p(\tau)} \quad \forall \tau, \sigma \tag{90}$$

- ***(C4) Smoothness.*** *The survival function ρ_S is C^1 in all arguments, ensuring the potential V is well-defined and the gradient flow is unique.*

When these conditions hold, the potential takes the form:

$$V(p) = \sum_{\tau} p(\tau) \log \frac{p(\tau)}{\mu(\tau)} - \sum_{\tau} p(\tau) \log \left(w_S(\tau) \cdot \rho_S^{local}(\tau) \right) + \Psi(p) \tag{91}$$

where the first term is the relative entropy with respect to μ, the second is the log-fitness, and $\Psi(p)$ captures any remaining interaction terms satisfying (90).

Proof sketch. The proof proceeds by decomposition. Under detailed balance **(C1)**, the mutation contribution to the dynamics can be written as:

$$\sum_{\tau'} p(\tau')M(\tau' \to \tau) - p(\tau) = -p(\tau)\sum_{\tau'} \left(\frac{p(\tau')}{p(\tau)} \cdot \frac{M(\tau' \to \tau)}{M(\tau \to \tau')} - 1 \right) M(\tau \to \tau') \tag{92}$$

which under detailed balance simplifies to a term proportional to $\nabla \mathrm{KL}(p \| \mu)$, the gradient of relative entropy.

The selection term, under **(C2)** and **(C3)**, reduces to the standard potential game form. The separability condition ensures that network effects cancel in payoff differences, while the symmetry condition ensures integrability.

The full argument requires showing that the combined dynamics—mutation plus selection—remain a gradient flow when both components individually are. This holds when the mutation reference measure μ aligns with the selection equilibrium, which is generically the case when μ is chosen as the invariant distribution of the pure mutation process. See Hofbauer and Sigmund (1998) and Sandholm (2010) for the technical machinery. ■

D.3 The Consent-Friction Instantiation

For the consent-friction instantiation in Section 5, the conditions specialize as follows:

Corollary D.1 (Gradient Structure for Consent Dynamics)**.** *The legitimacy-weighted survival function $\rho_S = L/(1+F)$ induces gradient structure when:*

(i) The belief-transfer kernel $g(\bar{O}', \bar{O}) = \exp(-\gamma(\bar{O}' - \bar{O}))$ satisfies detailed balance, which holds when ownership perceptions $\bar{O}(\tau)$ define a consistent ordering across configurations.

(ii) Stakes σ and alignment α are type-dependent but do not create asymmetric cross-type dependencies: $\partial\alpha_\tau / \partial p_\sigma = \partial\alpha_\sigma / \partial p_\tau$.

(iii) The entropy ε is either constant across types or depends on types in a symmetric manner.

Under these conditions, the quasi-potential claimed in Section 4.2 is:

$$V(\tau) = \log L(\tau) - \log(1 + F(\tau)) + \log w_S(\tau) \tag{93}$$

and friction-minimizing configurations correspond to local minima of V.

D.4 Counterexamples: When Gradient Structure Fails

The conditions are not merely technical. When they fail, qualitatively different dynamics emerge.

D.4.1 Counterexample 1: Rock-Paper-Scissors and Cyclic Dominance

The canonical example of non-gradient dynamics is rock-paper-scissors (RPS), where the payoff matrix exhibits cyclic dominance:

$$A = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix} \tag{94}$$

Proposition D.1 (RPS is Non-Potential)**.** *The rock-paper-scissors game violates condition **(C3)**. The payoff differences $\pi_i - \pi_j$ do not satisfy the integrability condition (84), and the replicator dynamics exhibit neutrally stable cycles around the interior equilibrium $p^* = (1/3, 1/3, 1/3)$.*

Proof. Direct computation shows $\frac{\partial \pi_1}{\partial p_2} = -1 \neq 1 = \frac{\partial \pi_2}{\partial p_1}$. The skew-symmetry of A implies the dynamics preserve a conserved quantity (the product $p_1 p_2 p_3$), generating closed orbits rather than convergence to equilibrium. See Hofbauer and Sigmund (1998), Sato and Crutchfield (2003), and Wesson and Rand (2016). ■

Relevance to ROM: If the legitimacy-friction structure induces cyclic dominance among institutional configurations—where configuration A beats B, B beats C, and C beats A—the dynamics will cycle indefinitely rather than converge. This can occur when network effects create asymmetric competitive advantages that form dominance cycles.

D.4.2 Counterexample 2: Asymmetric Mutation Kernels

Proposition D.2 (Asymmetric Mutation Induces Circulation)**.** *Let M_S be a mutation kernel violating detailed balance:*

$$M_S(\tau_1 \to \tau_2) = 0.3, \quad M_S(\tau_2 \to \tau_1) = 0.1 \tag{95}$$

Then the replicator-mutator dynamics exhibit a net circulation in state space, and no potential function exists.

Proof. The probability current $J_{12} = p_1 M_{12} - p_2 M_{21}$ is non-zero at equilibrium when $M_{12} \neq M_{21}$. A gradient flow has zero current at equilibrium (detailed balance), so the dynamics cannot be gradient. ■

Relevance to ROM: The belief-transfer modulation $g(\bar{O}', \bar{O}) = \exp(-\gamma(\bar{O}' - \bar{O}))$ is asymmetric whenever $\bar{O}(\tau') \neq \bar{O}(\tau)$. This asymmetry reflects the psychological reality that transitions reducing ownership perception are harder than transitions increasing it. While this makes the model more realistic, it formally breaks detailed balance. The quasi-potential description in the main text is thus an approximation valid when the asymmetry is small ($\gamma \ll 1$) or when the ownership landscape is approximately flat.

D.4.3 Counterexample 3: Network-Induced Oscillations

Proposition D.3 (Network Heterogeneity Breaks Separability)**.** *Let the network $G_{S,t}$ partition agents into two communities with preferential interaction. If community-level fitness depends on relative community sizes in an asymmetric way:*

$$\rho_S(\tau, G, p) = \rho_0(\tau) \cdot \left(1 + \beta \cdot sign(p_{community(\tau)} - 0.5)\right) \tag{96}$$

then the separability condition ***(C2)*** *fails, and the dynamics can exhibit limit cycles or chaos.*

Proof. The discontinuous dependence on relative population share introduces a non-smooth feedback. Even smoothed versions create asymmetric payoff dependencies that violate (90). Sato and Crutchfield (2003) demonstrate that such coupled dynamics can exhibit deterministic chaos in the rock-paper-scissors game; analogous phenomena arise in networked populations. See also Galla and Farmer (2013) on chaotic dynamics in learning systems. ■

Relevance to ROM: Network structure in institutional dynamics often exhibits precisely this character: communities that are "winning" attract more members, creating positive feedback that can destabilize equilibria. The ROM framework's network term $G_{S,t}$ accommodates this but at the cost of gradient structure.

D.5 Implications for Main Text Claims

The analysis above clarifies the scope of claims in Section 4.2:

1. The quasi-potential $V(\tau) = \log L(\tau) - \log(1+F(\tau)) + \log w_S(\tau)$ is valid **under the conditions (C1)–(C4)**.

2. When detailed balance fails (asymmetric belief-transfer), the dynamics remain well-defined but may exhibit circulation around equilibria rather than monotonic convergence.

3. When network separability fails, the dynamics can exhibit limit cycles or chaos, particularly when institutional configurations form cyclic dominance structures.

4. The empirical prediction that friction-minimizing configurations are attractors remains qualitatively valid when condition violations are small: the quasi-potential provides a good approximation to the "energy landscape" even if not exact.

The friction-minimization claim is most robust for:

- **Slowly evolving networks**: When $G_{S,t}$ changes on timescales much longer than population dynamics, the network contribution becomes effectively constant, restoring separability.
- **Small belief-transfer asymmetry**: When γ is small, the detailed balance violation is perturbative and the quasi-potential approximation is accurate.
- **Strong friction gradients**: When legitimacy differences between configurations are large, the gradient term dominates any circulation terms, and the dynamics approximately follow the potential descent.

These are the conditions under which the main text's claims are most secure. When they fail, ROM still provides a valid dynamical description, but convergence to friction-minimizing configurations is no longer guaranteed—oscillations, cycles, and complex attractors become possible, consistent with the observed instability of many governance arrangements in practice.

D.6 Technical Notes

On the Shahshahani metric and information geometry. The Shahshahani metric is the unique (up to scaling) Riemannian metric on the simplex that is invariant under sufficient statistics and coincides with the Fisher-Rao metric from information geometry (Shahshahani, 1979). This connection explains why the relative entropy appears naturally in the potential (91): replicator dynamics are gradient flows of relative entropy when the game has potential structure.

On potential games in economics. The characterization of potential games is due to Monderer and Shapley (1996). Their key result—that a game is potential if and only if the payoff Jacobian is symmetric—is the game-theoretic analogue of the classical result that a vector field is a gradient if and only if its Jacobian is symmetric.

On chaotic dynamics in games. The emergence of chaos in simple game dynamics was demonstrated by Sato and Crutchfield (2003) for coupled replicator equations and by Galla and Farmer (2013) for best-response learning. Pangallo et al. (2019) show that convergence to Nash equilibrium is the exception rather than the rule in generic games. ROM inherits these properties when the underlying fitness landscape fails the potential conditions.

E Microfoundations of the Ownership Modulation Function

The ownership-modulation function $g(\bar{O}', \bar{O}) = \exp(-\gamma(\bar{O}' - \bar{O}))$ appears in Section 4.2 as the belief-transfer mechanism that suppresses transitions reducing aggregate ownership perception. This Arrhenius-like form is not an arbitrary parameterization but emerges from convergent derivations across four independent literatures. The exponential form appears whenever transitions require overcoming barriers in stochastic environments, whether the barriers are thermodynamic, psychological, or informational.

E.1 Statistical Mechanics Foundation

In the statistical physics of social systems (Castellano et al., 2009), collective behavior emerges from microscopic transition rates that follow Boltzmann distributions. The probability of a system transitioning from state τ' to τ depends exponentially on the energy barrier:

$$P(\tau' \to \tau) \propto \exp\left(-\frac{\Delta E}{kT}\right) \tag{97}$$

where ΔE is the energy barrier, k is Boltzmann's constant, and T is temperature.

When ownership perception $\bar{O}$ serves as a component of the effective energy landscape—higher ownership corresponding to deeper potential wells—the exponential dependence on ownership differentials follows directly. The physical intuition is that random fluctuations (shocks, exogenous events) must supply the activation energy to overcome ownership resistance; the probability of sufficiently large fluctuations decays exponentially with barrier height.

Mapping to institutional transitions:

- Energy barrier ΔE ↔ Psychological cost of abandoning ownership claims
- Temperature T ↔ System volatility or "noise" in the institutional environment
- Transition rate ↔ Institutional change probability

E.2 Kramers Rate Theory

Kramers' theory (Hänggi et al., 1990) describes the escape rate of a Brownian particle from a metastable potential well:

$$r = \frac{\omega_a \omega_b}{2\pi\gamma_{\text{fric}}} \exp\left(-\frac{\Delta U}{kT}\right) \tag{98}$$

where ΔU is the barrier height, γ_{fric} is the friction coefficient, and ω_a, ω_b characterize the potential well curvature.

Application to institutional transitions: An institutional configuration with established ownership claims is analogous to a particle trapped in a potential well. The "depth" of the well corresponds to the strength of ownership psychology. Transition to a new configuration requires crossing an activation barrier, and the exponential dependence on barrier height (ownership differential) emerges from the probability distribution of random fluctuations that can supply the necessary "energy."

The Kramers framework predicts that transition rates should depend exponentially on the ownership differential $(\bar{O}' - \bar{O})$, which is precisely the ROM claim. The parameter γ in ROM corresponds to the barrier steepness divided by effective temperature.

E.3 Behavioral Economics Foundation

The endowment effect demonstrates that people value objects they own more highly than equivalent objects they do not own (Kahneman and Tversky, 1979). Critically for ROM:

(i) **Duration dependence:** Valuation increases with ownership duration (Strahilevitz and Loewenstein, 1998). Experimental data shows that holding an object for 30 seconds versus 10 seconds increased valuation by approximately 37%.

(ii) **Loss aversion:** Prospect theory establishes that losses loom larger than gains, with the value function typically parameterized as asymmetric around the reference point.

(iii) **Status quo bias:** Samuelson and Zeckhauser (1988) document systematic preference for current states independent of their objective quality.

Derivation of exponential form: When agents evaluate transitions between ownership states using a softmax/logit choice rule, the probability of accepting an ownership-reducing transition becomes:

$$P(\text{accept}) = \frac{1}{1+\exp(\lambda \cdot |\Delta O|)} \approx \exp(-\lambda \cdot \Delta O) \quad \text{for } \Delta O > 0, \lambda \gg 1 \tag{99}$$

where λ is the loss aversion coefficient. The ROM parameter γ absorbs this behavioral asymmetry: $\gamma = \lambda \cdot \partial V / \partial O$, where V is the value function and O is ownership perception.

E.4 Bounded Rationality Foundation

Quantal Response Equilibrium (QRE) models bounded rationality using the Boltzmann/softmax choice rule (McKelvey and Palfrey, 1995):

$$P(a_i) = \frac{\exp(\beta \cdot U(a_i))}{\sum_j \exp(\beta \cdot U(a_j))} \tag{100}$$

where β is the "rationality parameter" (inverse temperature).

Ortega and Braun (2013) show that when decision-making has information-processing costs, the optimal policy is a Boltzmann distribution over actions. The "temperature" corresponds to the trade-off between expected utility and computational costs. This provides a *normative* foundation for the exponential form: it is not just empirically observed but is the optimal response for boundedly rational agents.

When agents evaluate transitions between institutional configurations, their acceptance probability follows:

$$P(\tau' \to \tau) \propto \exp(\beta \cdot \Delta V(\tau', \tau)) \tag{101}$$

If ownership perception $\bar{O}$ is a component of subjective value V, and if the relationship is approximately linear, then:

$$P(\tau' \to \tau) \propto \exp(-\gamma(\bar{O}' - \bar{O})) \tag{102}$$

The parameter γ in ROM corresponds to $\beta \cdot \partial V / \partial O$—the product of rationality and the marginal value of ownership.

E.5 Convergent Validity

The four pathways converge on the same functional form through different mechanisms:

Pathway	Mechanism	Key Parameter
Statistical Mechanics	Boltzmann distribution	E_a/kT
Kramers Rate Theory	Barrier-crossing escape rate	$\Delta U/kT$
Behavioral Economics	Loss aversion + logit response	$\lambda \cdot \Delta O$
Bounded Rationality	Info-theoretic optimal policy	$\beta \cdot \Delta V$

The ROM parameter γ unifies these interpretations:

- In statistical mechanics: inverse temperature times ownership-energy coupling
- In Kramers theory: barrier steepness in ownership space
- In behavioral economics: loss aversion coefficient times ownership salience
- In bounded rationality: rationality parameter times marginal value of ownership

That the same functional form emerges from physics, economics, and cognitive science independently—through different mechanisms and assumptions—provides convergent validation that goes beyond any single theoretical commitment.

E.6 Empirical Predictions

The distinctive implication is that regime transition probability should *decrease exponentially* with incumbent tenure, controlling for legitimacy and resources. Unlike generic "institutional stickiness" explanations that predict gradual resistance, the Arrhenius form predicts a specific functional relationship:

$$P(\text{transition}|\text{tenure} = t) \propto \exp(-\gamma \cdot f(t)) \tag{103}$$

where $f(t)$ is the ownership accumulation function (plausibly linear or logarithmic in tenure).

Testable distinctions:

(i) **Exponential vs. linear**: Linear resistance predicts constant marginal resistance to ownership loss; exponential predicts increasing marginal resistance.

(ii) **Exponential vs. power-law**: Power-law resistance ($\gamma \cdot |\Delta O|^\beta$) predicts different behavior near zero (linear approach vs. exponential approach).

(iii) **Time-series signature**: Regime survival probability should follow exponential decay with increasing ownership accumulation, testable via hazard models against historical institutional data.

E.7 Limitations

The derivation assumes:

(i) Linear relationship between ownership perception and effective barrier height (may saturate at extremes)

(ii) Homogeneous γ across institutional contexts (likely varies)

(iii) Continuous ownership accumulation (may exhibit jumps at critical events)

These assumptions are standard simplifications that enable tractability. Empirical calibration of γ across domains remains an open task, though the qualitative prediction—exponential rather than linear or power-law resistance—is robust to reasonable parameter variation.

F Formal Verification in Lean 4

The algebraic core of the ROM equation and its consent-friction instantiation have been machine-checked in Lean 4 (v4.27.0) with the Mathlib library (v4.27.0). The formalization covers four modules totaling 28 machine-checked theorems with zero errors, zero `sorry` placeholders, and zero axioms beyond Lean's foundational type theory:

- **ROM/Basic.lean**: Simplex preservation, row-stochastic normalization, identity-kernel Bayesian reduction, detailed balance, RPS non-potential, consent-weighted survival monotonicity (11 theorems)
- **ROM/Advanced.lean**: Survival function properties—nonnegativity, zero-legitimacy collapse, linearity, upper bound, positivity, constant-source invariance (6 theorems)
- **ROM/Transfers.lean**: Moving equilibrium existence, impossibility of static equilibrium under varying legitimacy/friction, bounded chase error, dissensus from positive discrepancy, path nonnegativity (5 theorems)
- **ROMEthics/**: Welfare-friction bridge theorems connecting ethical survival to ROM consent survival, monotonicity in benefit and alignment, anti-monotonicity in harm and friction, boundary recovery conditions (6 theorems)

The Transfers module is particularly notable: it proves that ROM dynamics under time-varying legitimacy and friction admit moving equilibria but *not* static equilibria—a formal verification of the paper's central claim that persistent systems do not reach classical equilibria but track them asymptotically. The dissensus theorem (`rom_dissensus_of_positive_discrepancy`) formalizes the connection between stake-voice mismatch and institutional instability.

Selected proof: simplex preservation (§4.1).

```
theorem rom_simplex_invariant {n : N} (f p : Fin n -> R)
    (M : Fin n -> Fin n -> R)
    (hM : forall j, sum i : Fin n, M j i = 1)
    (hp : sum i : Fin n, p i = 1) :
    sum i : Fin n,
      ((sum j : Fin n, f j * M j i) -
       p i * (sum j : Fin n, f j)) = 0 := by
  have h1 : sum i, sum j, f j * M j i = sum j, f j :=
    row_stochastic_sum f M hM
  have h2 : sum i, p i * (sum j, f j) = sum j, f j := by
    rw [<- Finset.sum_mul, hp, one_mul]
  have h3 : -- distribute sum over subtraction
    sum i, (sum j, f j * M j i - p i * (sum j, f j)) =
    (sum i, sum j, f j * M j i) - (sum i, p i * (sum j, f j)) := by
      simp_rw [sub_eq_add_neg]
      rw [Finset.sum_add_distrib, Finset.sum_neg_distrib]
  rw [h3, h1, h2, sub_self]
```

Selected proof: alignment increases ethical survival (ROMEthics).

```
theorem ethicalSurvival_mono_alignment_via_friction
    {benefit harm weight sigma alpha1 alpha2 epsilon : R}
    (hW : 0 < welfareScore benefit harm weight)
    (hsigma : 0 < sigma) (heps : 0 <= epsilon)
    (halpha1 : -1 < alpha1) (halpha : alpha1 < alpha2) :
    ethicalSurvival benefit harm weight
      (friction sigma alpha1 epsilon) <
    ethicalSurvival benefit harm weight
      (friction sigma alpha2 epsilon) := by
  have hF : friction sigma alpha2 epsilon <
             friction sigma alpha1 epsilon :=
    friction_strict_anti_alignment hsigma heps halpha1 halpha
  -- reduces to consent_survival_anti_friction
  ...
```

Source code and build instructions: Farzulla (2026). Verification reproduces via `lake build` with Lean 4 v4.27.0 and Mathlib v4.27.0. Zero `sorry` placeholders, zero axioms beyond Lean's foundational type theory.

References

Daron Acemoglu, Georgy Egorov, and Konstantin Sonin. Institutional change and institutional persistence. *NBER Working Paper*, (27852), 2020.

Ömer Deniz Akyildiz. A probabilistic interpretation of replicator-mutator dynamics. *arXiv preprint arXiv:1712.07879*, 2017.

Darcy W. E. Allen, Chris Berg, Brendan Markey-Towler, Mikayla Novak, and Jason Potts. Blockchain and the evolution of institutional technologies: Implications for innovation policy. *Research Policy*, 49(1):103865, 2020. doi: 10.1016/j.respol.2019.103865.

Unai Alvarez-Rodriguez, Federico Battiston, Guilherme Ferraz De Arruda, Yamir Moreno, Matjaž Perc, and Vito Latora. Evolutionary dynamics of higher-order interactions in social networks. *Nature Human Behaviour*, 5(5):586–595, 2021. doi: 10.1038/s41562-020-01024-1.

David Aristoff and Muyuan Zhu. Coarse-graining of markov chains via the mori-zwanzig formalism. *arXiv preprint arXiv:2305.20083*, 2023.

Jeffery C. Atik and George Gerro. Hard forks on the Bitcoin blockchain: Reversible exit, continuing voice. *Stanford Journal of Blockchain Law & Policy*, 2(1), 2018.

Peter C. Austin, Douglas S. Lee, and Jason P. Fine. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133(6):601–609, 2016. doi: 10.1161/CIRCULATIONAHA.115.017719.

Nataliya A. Balabanova, Manh Hong Duong, and The Anh Han. Replicator-mutator dynamics for public goods games with institutional incentives, 2025. Accepted at Mathematical Biosciences.

Adam B. Barrett. An exploration of synergy in gaussian systems via information geometry. *Entropy*, 17 (7):4644–4666, 2015. doi: 10.3390/e17074644.

Stefano Battiston, Guido Caldarelli, Robert M. May, Tarik Roukny, and Joseph E. Stiglitz. The price of complexity in financial networks. *Proceedings of the National Academy of Sciences*, 113(36): 10031–10036, 2016a. doi: 10.1073/pnas.1521573113.

Stefano Battiston, J. Doyne Farmer, Andreas Flache, Diego Garlaschelli, Andrew G. Haldane, Hans Heesterbeek, Cars Hommes, Carlo Jaeger, Robert May, and Marten Scheffer. Complexity theory and financial regulation. *Science*, 351(6275):818–819, 2016b. doi: 10.1126/science.aad0299.

Roman Beck, Christoph Müller-Bloch, and John Leslie King. Governance in the blockchain economy: A framework and research agenda. *Journal of the Association for Information Systems*, 19(10):1020–1034, 2018. doi: 10.17705/1jais.00518.

Luís M. A. Bettencourt, Brandon J. Grandison, and Jordan T. Kemp. Redefining fitness: Evolution as a dynamic learning process. *arXiv preprint arXiv:2503.09057*, 2025.

Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53:659–697, 2015. doi: 10.1613/jair.4818.

Russell Bonduriansky and Stephen F. Chenoweth. Intralocus sexual conflict. *Trends in Ecology & Evolution*, 24(5):280–288, 2009. doi: 10.1016/j.tree.2008.12.005.

Zoltán Botta-Dukát. Rao's quadratic entropy as a measure of functional diversity based on multiple traits. *Journal of Vegetation Science*, 16(5):533–540, 2005. doi: 10.1111/j.1654-1103.2005.tb02393.x.

Robert Boyd and Peter J. Richerson. *Culture and the Evolutionary Process*. University of Chicago Press, Chicago, 1985.

Uwe Cantner, Ivan Savin, and Simone Vannuccini. Replicator dynamics in value chains: Explaining some puzzles of market selection. *Industrial and Corporate Change*, 28(3):589–611, 2019. doi: 10.1093/icc/dty060. Multi-layer structure can reverse apparent selection effects.

Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646, 2009. doi: 10.1103/RevModPhys.81.591.

Luigi Luca Cavalli-Sforza and Marcus W. Feldman. *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton University Press, Princeton, 1981.

Dániel Czégel, Hamza Giaffar, Joshua B. Tenenbaum, and Eörs Szathmáry. Bayes and Darwin: How replicator populations implement Bayesian computations. *BioEssays*, 44(4):2100255, 2022. doi: 10.1002/bies.202100255.

Charles Darwin. *On the Origin of Species by Means of Natural Selection*. John Murray, London, 1859.

Jinming Du, Bin Wu, and Long Wang. Aspiration dynamics in structured population acts as if in a well-mixed one. *Scientific Reports*, 5:8014, 2015. doi: 10.1038/srep08014.

Claire El Mouden, Jean-Baptiste André, Olivier Morin, and Daniel Nettle. Cultural transmission and the evolution of human behaviour: A general approach based on the Price equation. *Journal of Evolutionary Biology*, 27(2):231–241, 2014. doi: 10.1111/jeb.12296.

Murad Farzulla. The axiom of consent: Friction dynamics in multi-agent coordination. *arXiv preprint arXiv:2601.06692*, 2025a. doi: 10.48550/arXiv.2601.06692. Unified friction framework for multi-agent coordination.

Murad Farzulla. Asymptotic protection: Derivatives, systemic risk, and the limits of hedging. *Zenodo Preprint*, 2025b. doi: 10.5281/zenodo.17620448. Derivatives and systemic risk management.

Murad Farzulla. Consent-theoretic framework for quantifying legitimacy: Stakes, voice, and friction in adversarial governance. *Zenodo Preprint*, 2025c. doi: 10.5281/zenodo.17684676. Operationalization of consent-based legitimacy framework.

Murad Farzulla. From consent to consideration: Why embodied autonomous systems cannot be legitimately ruled. *Zenodo Preprint*, 2025d. doi: 10.5281/zenodo.17957659. Under review at AI and Ethics (Springer). AI political standing and moral consideration.

Murad Farzulla. Training data and the maladaptive mind: A computational framework for developmental trauma. *Research Square*, 2025e. doi: 10.21203/rs.3.rs-8634152/v1. Under review at Humanities & Social Sciences Communications (Nature).

Murad Farzulla. Machine-checked proofs for adversarial systems research, 2026. URL `https://github.com/studiofarzulla/lean-formalizations`. Lean 4 + Mathlib formalizations covering ROM, AoC, Identity Thesis, and 11 other papers. 50 files, ∼2,500 lines, zero `sorry`.

Murad Farzulla and Andrew Maksakov. ASRI: An aggregated systemic risk index for cryptocurrency markets. *arXiv preprint arXiv:2602.03874*, 2025. doi: 10.48550/arXiv.2602.03874. Systemic risk as emergent from distributed friction sources.

Robin Fritsch, Marino Müller, and Roger Wattenhofer. Analyzing voting power in decentralized governance: Who controls DAOs? *arXiv preprint arXiv:2204.01176*, 2022.

Tobias Galla and J. Doyne Farmer. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences*, 110(4):1232–1236, 2013. doi: 10.1073/pnas.1109672110.

David Geiger and Zvi M. Kedem. Information-theoretic bounds on the coarse-graining of markov chains. *arXiv preprint arXiv:2204.13896*, 2022.

Karl Peter Hadeler. Stable polymorphisms in a selection model with mutation. *SIAM Journal on Applied Mathematics*, 41(1):1–7, 1981. doi: 10.1137/0141001.

Jarrod D. Hadfield and Shinichi Nakagawa. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology*, 23(3):494–508, 2010. doi: 10.1111/j.1420-9101.2009.01915.x.

Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: fifty years after Kramers. *Reviews of Modern Physics*, 62(2):251–341, 1990. doi: 10.1103/RevModPhys.62.251.

Marc Harper. Information geometry and evolutionary game theory. *arXiv preprint arXiv:0911.1383*, 2009.

Christoph Hauert and Michael Doebeli. Origin of diversity in spatial social dilemmas. *Proceedings of the National Academy of Sciences*, 118(42):e2105252118, 2021. doi: 10.1073/pnas.2105252118.

Håvard Hegre, Tanja Ellingsen, Scott Gates, and Nils Petter Gleditsch. Toward a democratic civil peace? democracy, political change, and civil war, 1816–1992. *American Political Science Review*, 95(1):33–48, 2001. doi: 10.1017/S0003055401000119.

Joseph Henrich. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton University Press, Princeton, 2016.

Erik P. Hoel. When the map is better than the territory. *Entropy*, 19(5):188, 2017. doi: 10.3390/e19050188.

Erik P. Hoel, Larissa Albantakis, and Giulio Tononi. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49):19790–19795, 2013. doi: 10.1073/pnas.1314922110.

Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, 1998. doi: 10.1017/CBO9781139173179.

David L. Hull. Individuality and selection. *Annual Review of Ecology and Systematics*, 11:311–332, 1980. doi: 10.1146/annurev.es.11.110180.001523.

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

Leo P. Kadanoff. Scaling laws for Ising models near Tc. *Physics Physique Fizika*, 2(6):263–272, 1966. doi: 10.1103/PhysicsPhysiqueFizika.2.263.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979. doi: 10.2307/1914185.

John G. Kemeny and J. Laurie Snell. *Finite Markov Chains*. Springer-Verlag, New York, 2nd edition, 1976.

Margarita Kirneva and Matías Nuñez. Legitimacy of collective decisions: a mechanism design approach, 2023.

Ryogo Kubo. The fluctuation-dissipation theorem. *Reports on Progress in Physics*, 29(1):255–284, 1966. doi: 10.1088/0034-4885/29/1/306.

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

Michel Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001. doi: 10.1090/surv/089.

Richard C. Lewontin. The units of selection. *Annual Review of Ecology and Systematics*, 1:1–18, 1970. doi: 10.1146/annurev.es.01.110170.000245.

Erez Lieberman, Christoph Hauert, and Martin A. Nowak. Evolutionary dynamics on graphs. *Nature*, 433(7023):312–316, 2005. doi: 10.1038/nature03204.

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995. doi: 10.1006/game.1995.1023.

Saul Mendoza-Palacios and Onésimo Hernández-Lerma. Stability of the replicator dynamics for games in metric spaces. *Journal of Dynamics and Games*, 4(4):319–333, 2017. doi: 10.3934/jdg.2017017.

Saul Mendoza-Palacios and Onésimo Hernández-Lerma. *Evolutionary Games and the Replicator Dynamics*. Elements in Evolutionary Economics. Cambridge University Press, 2024. doi: 10.1017/9781009472319.

Edward Miguel, Shanker Satyanath, and Ernest Sergenti. Economic shocks and civil conflict: An instrumental variables approach. *Journal of Political Economy*, 112(4):725–753, 2004. doi: 10.1086/421174.

Dov Monderer and Lloyd S. Shapley. Potential games. *Games and Economic Behavior*, 14(1):124–143, 1996. doi: 10.1006/game.1996.0044.

Hazime Mori. Transport, collective motion, and brownian motion. *Progress of Theoretical Physics*, 33 (3):423–455, 1965. doi: 10.1143/PTP.33.423.

Mark E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6):066133, 2004. doi: 10.1103/PhysRevE.69.066133.

Mark E. J. Newman. *Networks: An Introduction*. Oxford University Press, Oxford, 2010. doi: 10.1093/acprof:oso/9780199206650.001.0001.

Douglass C. North. *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, Cambridge, 1990. doi: 10.1017/CBO9780511808678.

Hisashi Ohtsuki, Christoph Hauert, Erez Lieberman, and Martin A. Nowak. A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 441(7092):502–505, 2006. doi: 10.1038/nature04605.

Samir Okasha. *Evolution and the Levels of Selection*. Oxford University Press, Oxford, 2006. doi: 10.1093/acprof:oso/9780199267972.001.0001.

Pedro A. Ortega and Daniel A. Braun. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A*, 469(2153):20120683, 2013. doi: 10.1098/rspa.2012.0683.

Karen M. Page and Martin A. Nowak. Unifying evolutionary dynamics. *Journal of Theoretical Biology*, 219(1):93–98, 2002. doi: 10.1006/jtbi.2002.3112.

Marco Pangallo, Torsten Heinrich, and J. Doyne Farmer. Best reply structure and equilibrium convergence in generic games. *Science Advances*, 5(2):eaat1328, 2019. doi: 10.1126/sciadv.aat1328.

Sahani Pathiraja and Philipp Wacker. Connections between sequential Bayesian inference and evolutionary dynamics. *Philosophical Transactions of the Royal Society A*, 383(2298):20240241, 2024. doi: 10.1098/rsta.2024.0241.

Tyler Porter and Peter Wikman. Evolutionary stability and tenable strategy blocks. *Economic Theory*, 2026. doi: 10.1007/s00199-026-01701-8. Online first; DOI confirmed but not yet fully indexed.

Simon T. Powers, Cédric Perret, and Thomas E. Currie. Playing the political game: The coevolution of institutions with group size and political inequality. *Philosophical Transactions of the Royal Society B*, 378(1883):20220303, 2023. doi: 10.1098/rstb.2022.0303. Consensus-cost model of institutional form selection.

George R. Price. Selection and covariance. *Nature*, 227(5257):520–521, 1970. doi: 10.1038/227520a0.

C. Radhakrishna Rao. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1):24–43, 1982. doi: 10.1016/0040-5809(82)90004-1.

Carlos P. Roca, José A. Cuesta, and Angel Sánchez. Evolutionary game theory: Temporal and spatial effects beyond replicator dynamics. *Physics of Life Reviews*, 6(4):208–249, 2009. doi: 10.1016/j.plrev.2009.08.001.

Onkar Sadekar, Andrea Civilini, Vito Latora, and Federico Battiston. Drivers of cooperation in social dilemmas on higher-order networks. *Journal of the Royal Society Interface*, 22:20250134, 2025. doi: 10.1098/rsif.2025.0134.

William Samuelson and Richard Zeckhauser. Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1):7–59, 1988. doi: 10.1007/BF00055564.

William H. Sandholm. *Population Games and Evolutionary Dynamics*. MIT Press, Cambridge, MA, 2010.

Francisco C. Santos and Jorge M. Pacheco. Scale-free networks provide a unifying framework for the emergence of cooperation. *Physical Review Letters*, 95(9):098104, 2005. doi: 10.1103/PhysRevLett.95.098104.

Yuzuru Sato and James P. Crutchfield. Coupled replicator equations for the dynamics of learning in multiagent systems. *Physical Review E*, 67(1):015206, 2003. doi: 10.1103/PhysRevE.67.015206.

Erwin Schrödinger. *What is Life? The Physical Aspect of the Living Cell*. Cambridge University Press, Cambridge, 1944.

Siavash Shahshahani. A new mathematical framework for the study of linkage and selection. *Memoirs of the American Mathematical Society*, 17(211):1–34, 1979. doi: 10.1090/memo/0211.

Chen Shen, Zhao Song, Xinyu Wang, Lei Shi, Matjaž Perc, Zhen Wang, and Jun Tanimoto. Evolutionary dynamics of reputation-based voluntary prisoner's dilemma games. arXiv preprint, 2026.

Didier Sornette, Sandro Claudio Lera, and Ke Wu. Why AI alignment failure is structural: Learned human interaction structures and AGI as an endogenous evolutionary shock. *SuperIntelligence—Robotics—Safety & Alignment*, 2(4), 2026. doi: 10.70777/si.v2i4.17163.

Michal A. Strahilevitz and George Loewenstein. The effect of ownership history on the valuation of objects. *Journal of Consumer Research*, 25(3):276–289, 1998. doi: 10.1086/209539.

Peter D. Taylor and Leo B. Jonker. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(1-2):145–156, 1978. doi: 10.1016/0025-5564(78)90077-9.

Yuan Tian, Anna Dabrowski, Shashwat Bhatt, and Jayanta K. Bhattacharjee. Data-driven computation of the memory kernel of the generalized langevin equation. *arXiv preprint arXiv:2108.13288*, 2021.

Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. Selection-mutation dynamics of Q-learning in multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 2:693–704, 2003.

Tamás Varga. Replicator dynamics generalized for evolutionary matrix games under time constraints. *Journal of Mathematical Biology*, 90(1):6, 2024. doi: 10.1007/s00285-024-02170-0. See also Varga, Móri & Garay (2019) in *J. Math. Biol.* for the ESS–replicator relationship under time constraints.

Thomas F. Varley and Erik P. Hoel. Emergence as the conversion of information: A unifying theory. *Philosophical Transactions of the Royal Society A*, 380(2227):20210150, 2022. doi: 10.1098/rsta.2021.0150.

Pablo Villegas, Tommaso Gili, Guido Caldarelli, and Andrea Gabrielli. Laplacian renormalization group for heterogeneous networks. *Nature Physics*, 19(3):445–450, 2023. doi: 10.1038/s41567-022-01866-8.

Jörgen W. Weibull. *Evolutionary Game Theory*. MIT Press, Cambridge, MA, 1995.

Steven Weinberg. Phenomenological Lagrangians. *Physica A: Statistical Mechanics and its Applications*, 96(1-2):327–340, 1979. doi: 10.1016/0378-4371(79)90223-1.

Elizabeth Wesson and Richard Rand. Hopf bifurcations in delayed rock-paper-scissors replicator dynamics. *Dynamic Games and Applications*, 6(1):139–156, 2016. doi: 10.1007/s13235-015-0138-2.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Kenneth G. Wilson. Renormalization group and critical phenomena. I. Renormalization group and the Kadanoff scaling picture. *Physical Review B*, 4(9):3174–3183, 1971. doi: 10.1103/PhysRevB.4.3174.

Yuanzhao Zhang, Takashi Nishikawa, and Adilson E. Motter. Coarse-graining complex networks for control. *Nature Physics*, 2025. In press.

Robert Zwanzig. Memory effects in irreversible thermodynamics. *Physical Review*, 124(4):983–992, 1961. doi: 10.1103/PhysRev.124.983.

Table 1: Machine-checked theorems and their correspondence to paper results.

Lean theorem	**Ref.**	**Description**
ROM/Basic.lean — Core algebraic results		
`row_stochastic_sum`	§4.1	Row-stochastic kernel preserves total mass
`rom_simplex_invariant`	§4.1	$\sum_i \dot{p}_i = 0$ (simplex preservation)
`identity_kernel_selection`	Rmk. 4.1	$M = I$ collapses to pure selection
`identity_kernel_row_stochastic`	Rmk. 4.1	Identity kernel is row-stochastic
`detailed_balance_zero_flow`	App. D	Detailed balance $\Rightarrow$ zero net flow
`asymmetric_kernel_no_uniform_balance`	App. D	Asymmetric M violates detailed balance
`rps_skew_symmetric`	App. D	$A_{ij} = -A_{ji}$ for RPS
`rps_not_potential`	App. D	RPS violates integrability
`skew_symmetric_nonzero_not_potential`	App. D	Nonzero skew-entry $\Rightarrow$ non-potential
`consent_survival_mono_legitimacy`	§5	$\partial\rho/\partial L > 0$
`consent_survival_anti_friction`	§5	$\partial\rho/\partial F < 0$
ROM/Advanced.lean — Survival function properties		
`consent_survival_nonneg`	§5	$L \geq 0, F \geq 0 \Rightarrow \rho \geq 0$
`consent_survival_zero_legitimacy`	§5	$L = 0 \Rightarrow \rho = 0$
`consent_survival_scale_legitimacy`	§5	$\rho(cL, F) = c \cdot \rho(L, F)$
`consent_survival_le_legitimacy`	§5	$\rho(L, F) \leq L$ when $F \geq 0$
`consent_survival_pos`	§5	$L > 0, F \geq 0 \Rightarrow \rho > 0$
`row_stochastic_sum_const`	§4.1	Constant source through row-stochastic kernel
ROM/Transfers.lean — Dynamic equilibrium results		
`rom_movingEquilibrium`	§4–5	ROM path admits moving equilibrium
`rom_no_static_if_path_varies`	§4–5	Varying *L*/*F* precludes static equilibrium
`rom_path_boundedChase_zero`	§4–5	Exact tracking yields zero chase error
`rom_dissensus_of_positive_discrepancy`	§5.3	Positive discrepancy $\Rightarrow$ dissensus
`romPath_nonneg`	§5	Nonneg *L*/*F* signals $\Rightarrow$ nonneg path
ROMEthics/ — Welfare-friction bridge		
`ethicalSurvival_mono_benefit`	§5	$\partial\rho_{\text{eth}}/\partial b > 0$
`ethicalSurvival_anti_harm`	§5	$\partial\rho_{\text{eth}}/\partial h < 0$
`ethicalSurvival_eq_rom_survival_of_zero_harm`	§5	Zero harm recovers ROM survival
`ethicalSurvival_mono_alignment_via_friction`	§5	Better α increases survival via lower F
`ethicalSurvival_eq_welfare_of_zero_friction`	§5	Zero friction recovers raw welfare
`ethicalSurvival_le_welfare`	§5	Ethical survival $\leq$ welfare