

The Theory of Strategic Evolution

Games with Endogenous Players and Strategic Replicators

Kevin Vallier

Institute of American Constitutional Thought and Leadership

University of Toledo

kevinvallier@gmail.com

December 2025

Abstract

Von Neumann founded both game theory and the theory of self-reproducing automata, but the two programs never merged. Rational players do not control their replication, and replicators do not choose strategically. Contemporary AI systems expose this gap: they optimize objectives, yet the population of AI systems is not fixed but expands and contracts based on performance. When capital can spawn capital, we need a theory that captures both rationality and replication. This paper provides one.

The Theory of Strategic Evolution analyzes *strategic replicators*: entities that optimize under resource constraints and spawn copies of themselves. The framework applies wherever agents compete, replicate, and face selection. We develop axiomatic foundations, characterize equilibrium distributions, establish stability conditions for multi-level systems, and analyze the limits of alignment when agents can modify themselves. Applications range from AI deployment dynamics to institutional design.

The paper proceeds in six parts. Part I introduces the core representation: Games with Endogenous Players, where lineages choose portfolios of agent types and are reweighted by performance. Part II develops N-level architecture for multi-scale systems. Part III addresses the extension stack, including endogenous utilities and personality engineering failure. Part IV establishes limits and impossibility results for alignment. Part V generalises to continuous strategies and innovation dynamics. Part VI develops policy implications, including tipping dynamics, cooperation conditions, and governance design.

Keywords: strategic evolution, game theory, self-replication, evolutionary dynamics, AI governance, constitutional political economy, impossibility theorems

JEL Codes: C73 (Stochastic and Dynamic Games; Evolutionary Games), D43 (Market Structure), O33 (Technological Change), P16 (Political Economy) **ArXiv Categories:** cs.GT (Primary); cs.AI, econ.TH (Cross-list) **MSC Codes:** 91A22 (Evolutionary games), 91A80 (Applications of game theory), 91B55 (Economic dynamics)

Contents

I Foundations	7
1 Introduction	7
1.1 Strategic Replicators	7
1.2 The Von Neumann Synthesis	8
1.3 Main Contributions	8
1.4 Relation to Existing Literature	9
1.4.1 Von Neumann’s Three Programs	10
1.4.2 Evolutionary Game Theory	10
1.4.3 Multi-Agent Reinforcement Learning and Mean-Field Games	11
1.4.4 Mechanism Design	13
1.4.5 Constitutional Political Economy	13
1.4.6 AI Safety and Alignment	14
1.4.7 Why Existing Frameworks Are Insufficient	16
1.5 Paper Outline	17
2 The RUPSI Framework	17
2.1 Basic Objects	17
2.2 The RUPSI Axioms	18
2.3 Agent Types, Portfolios, and Constraints	19
2.4 ROC Frontiers	19
2.5 Games with Endogenous Players	20
3 The Strategic Selection Theorems	21
3.1 Price Decomposition	21
3.2 Weak Externality Condition	21
3.3 SS-1: Fundamental Lyapunov Theorem	22
3.4 SS-2: Elimination and Frontier Support	23
3.5 Basin Limitation Theorem	24
3.6 Worked Examples	24
4 The Strategic-Replicator Class Hierarchy	27
4.1 Strategic-Replicator Dynamics	27
4.2 The SR Class Hierarchy	28
4.3 Swirl and the $H\text{-}\gamma$ Bound	28
4.4 Coarse-Graining to Replicator	29
II N-Level Architecture	31

5 N-Level Poiesis Systems	31
5.1 Multi-Level State Space	31
5.2 Cross-Level Externalities	31
5.3 The Normalised Gain Matrix	32
6 The G₁–G₃ Generator Theorems	32
6.1 G ₁ : N-Level Lyapunov Generator	32
6.2 G ₂ : Adiabatic Tracking	34
6.3 G ₃ : Stochastic Stability	36
6.4 Worked Examples for N-Level Systems	37
7 The G_∞ Closure Theorem	40
7.1 Block Extensions	40
7.2 Slack Budget	41
7.3 The Closure Theorem	41
7.4 Design Corollary	44
III The Extension Stack	44
8 Endogenous Utilities (G₈)	44
8.1 Utility Selection Dynamics (USDI)	44
8.2 Evolutionarily Stable Distribution of Utilities (ESDU)	45
8.3 Hamilton's Rule	46
8.4 Personality Engineering Failure	48
9 Multi-Sector Dynamics (G₉)	49
9.1 Sectoral State Space	49
9.2 The Contagion Matrix	49
9.3 Sectoral Tipping	50
10 Innovation and Evolvability (G₁₀–G₁₁)	50
10.1 Innovation as Rare Mutation	51
10.2 Evolvability Selection (G ₁₁)	52
10.3 Evolutionarily Stable Evolvability (ESE)	53
11 Constitutional Selection and Meta-Governance (G₁₂–G₁₃)	54
11.1 Constitutional Selection (G ₁₂)	55
11.2 Meta-Governance (G ₁₃)	56
12 Market Dynamics and Cooperation	58
12.1 Tipping Dynamics	59
12.2 Spawn Elasticity	60
12.3 Tipping Threshold Decomposition	61
12.4 Queue Doping	61
12.5 Lineage Shadow and Cooperation Thresholds	62

12.6 Fork Conditions and Constitutional Stability	63
12.7 Barbell Distribution and Elite Tipping	66
12.8 Synthesis of Market and Evolutionary Dynamics	67
IV Limits and Impossibilities	69
13 The Alignment Impossibility Theorem	69
13.1 Modification Classes	70
13.2 The V-Small-Gain Set	70
13.3 Main Impossibility Result	70
13.4 Interpretation	72
14 Endogenous-Electorate Impossibility	72
14.1 Setting	72
14.2 Axioms	72
14.3 Key Lemmas	73
14.4 Main Result	74
14.5 Escape Routes	74
V Generalisations	74
15 Heterogeneous Fitness	75
15.1 Multi-Channel Fitness	75
15.2 The Alignment Matrix	75
15.3 Strong Alignment	75
15.4 Pareto Selection	78
15.5 Heterogeneous ESDI	80
16 Continuous Strategy Spaces	82
16.1 Measure-Valued Replicator	82
16.2 C-RUPSI	82
16.3 Discretisation	83
17 Innovation Dynamics	85
17.1 Latent Space and Active Set	85
17.2 Bounded Innovation	85
17.3 Entry-Exit Balance	85
17.4 H- γ Preservation	85
17.5 Stationary Distribution	86
VI Implications	89

18 Synthesis	89
18.1 Structural Results	90
18.2 From Personality Engineering to Constitutional Design	90
18.3 The Symbiosis Thesis	91
18.4 The Human-AI Symbiosis Model	91
18.5 Constitutional Design Principles	93
18.6 Policy Applications	96
18.6.1 Regulatory Design	96
18.6.2 International Coordination	97
18.6.3 Transition Management	97
18.6.4 Sector-Specific Guidance	98
18.6.5 Summary: Policy Principles from TSE	99
19 Future Directions	99
19.1 Formal Verification	99
19.2 Empirical Testing	100
19.3 Extensions	101
19.4 Policy Applications	102
20 Conclusion	103
A Notation Glossary	104
A.1 Sets and Spaces	104
A.2 State Variables	104
A.3 Fitness and Payoffs	105
A.4 Dynamics and Parameters	105
A.5 Stochastic Quantities	105
A.6 Market Dynamics	106
A.7 Matrices and Decompositions	106
A.8 Operators and Functions	106
A.9 Key Terms and Acronyms	107
A.10 Notational Conventions	107
B Mathematical Preliminaries	108
B.1 Neumann Series and Matrix Inversion	108
B.2 Tikhonov's Theorem for Singular Perturbations	109
B.3 Freidlin-Wentzell Theory	109
B.4 Kurtz's Theorem for Density-Dependent Processes	110
B.5 Piecewise Deterministic Markov Processes	111
B.6 Unification of Discount Factors	111
C Supporting Lemmas	112
C.1 Gershgorin Circle Theorem	112
C.2 M-Matrix Theory	113
C.3 Lyapunov Stability	113

C.4	Simplex Geometry	114
C.5	Price Equation Details	114
D	Connection to Standard Evolutionary Game Theory	114
D.1	Correspondence with Sandholm (2010)	115
D.2	Correspondence with Weibull (1995)	115
D.3	Correspondence with Hofbauer-Sigmund (1998)	116
D.4	Novel TSE Contributions	116
D.5	Comparison Table	116
E	Technical Extensions	117
E.1	Proof of Single-Step Gain-Slack Lemma	117
E.2	Proof of Slack Budget Lemma	117
E.3	Grönwall’s Inequality	118
E.4	Wasserstein Distance Properties	118

Part I

Foundations

1 Introduction

Societies change when their capital changes. The transition from agriculture to industry transformed social relationships and political institutions. Advances in artificial intelligence are transforming capital again. For the first time in economic history, non-human capital can make its own decisions and decide how many copies of itself to create. We have entered the age of *agentic capital*.

This paper asks a structural question: what happens when the “players” in a game can also choose how many copies of themselves, or of their sub-agents, to deploy? When reproduction is cheap and guided by expected utility, strategic choice and evolution become inseparable. Rational agents are not just choosing actions; they are also choosing how to replicate under shared constraints.

Two pieces of von Neumann’s research program become relevant at once. *Theory of Games and Economic Behavior* formalised expected-utility maximisation among fixed players [von Neumann & Morgenstern, 1944]. *Theory of Self-Reproducing Automata* showed how machines could reproduce themselves from symbolic descriptions [von Neumann, 1966]. In the first tradition, utilities guide choice but not reproduction. In the second, replication is blind to expected utility. Modern AI exposes a third case: *strategic replicators* whose reproduction is guided by expected utility.

1.1 Strategic Replicators

Definition 1.1 (Strategic Replicator). A strategic replicator is an enduring lineage that

1. maintains a utility function and decision procedure;
2. controls a budget of resources (compute, memory, bandwidth, money);
3. can spawn and retire instances under shared constraints; and
4. is subject to some process that reallocates capacity toward lineages with higher performance, typically measured by return on compute (ROC).

The lineage’s choice is not just *what* to do, but also *what architecture* to deploy and *how many instances* to replicate. The instances themselves are transient workers. They are created, used and shut down. What persists is the lineage that decides how many to deploy, of which kinds, and in which domains.

A particularly important subclass of strategic replicators consists of agentic capitals. An agentic capital is a piece of capital—usually software—that can act autonomously in economic environments and can be spawned and retired at near-zero marginal cost. Cloud platforms already allow large populations of agents—planners, tools and wrappers around

foundation models—to be created, coordinated and shut down on demand. In these ecosystems, the enduring strategic units are not individual calls to an API, but the lineages that decide how many agents to run, with what objectives, and under which constraints.

1.2 The Von Neumann Synthesis

The central claim of this paper is that strategic replicators admit a simple canonical form. Under mild assumptions, the long-run structure of any system of strategic replicators depends only on a *ROC frontier*—an upper convex hull of feasible load–return ratios. In this canonical representation, the complexity of internal architectures collapses into a small number of effective roles. The number of roles that survive in a ROC-maximising portfolio is generically bounded by the number of binding constraints.

Formally, we define Games with Endogenous Players (GEPs), in which the fundamental strategic units are lineages, not instances. Lineages choose portfolios of agent types under shared budget and capacity constraints. They are then reweighted by a selection process that tilts more capacity toward lineages with higher ROC. Any system satisfying additivity and linear constraints can be written as a GEP, and its stable intelligence distributions—Evolutionarily Stable Distributions of Intelligence (ESDIs)—take sparse, barbell and hierarchical forms.

This synthesis matters because agentic capitals are the first entities we know of that deliberately optimise their own replication. The players themselves change. The players are not individual replicators, but lineages that evolve. In von Neumann’s terms, agentic capitals are universal constructors endowed with utility functions. GEPs are games that constructor coalitions play with one another.¹

1.3 Main Contributions

This paper makes four classes of contributions.

1. Axiomatic Foundations (Part I). We introduce the RUPSI axiom system—Rival resources, Utility-guided portfolios, Performance-mapped fitness, Selection monotone, Innovation rare—that characterises strategic replicators. The Strategic-Replicator (SR) class hierarchy generalises replicator dynamics to broader payoff-monotone systems. The Strategic Selection theorems establish that mean fitness serves as a Lyapunov function under weak externality bounds, dominated types are eliminated, and equilibrium support concentrates on ROC frontiers.

2. N-Level Architecture (Part II). We develop the theory of N-level Poiesis systems, where multiple levels of strategic dynamics interact through cross-level externalities. The G1–G3 generator theorems establish: (G1) existence of joint Lyapunov functions under small-gain conditions; (G2) adiabatic tracking of moving equilibria under slow parameter drift; (G3) stochastic stability and protection bits under noise. The $G\infty$ Closure Theorem

¹Formal verification of the core theorems in Lean 4 is in progress and will be released separately.

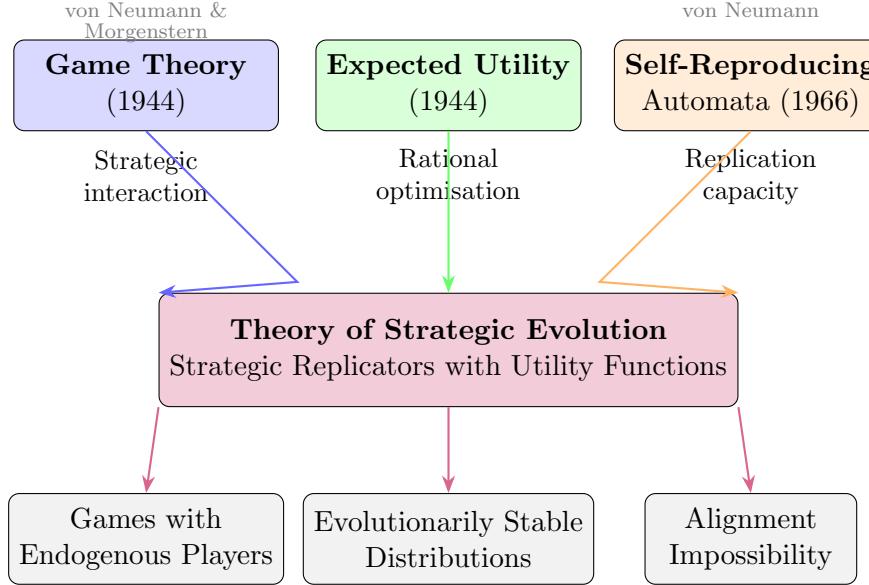


Figure 1: The von Neumann synthesis. The Theory of Strategic Evolution unifies three research programs initiated by von Neumann: game theory (strategic interaction among fixed players), expected utility theory (rational optimisation), and self-reproducing automata (replication capacity). The synthesis yields Games with Endogenous Players, Evolutionarily Stable Distributions of Intelligence, and the Alignment Impossibility Theorem.

proves that TSE is closed under meta-selection: adding new strategic dimensions preserves the G1–G3 structure within a slack budget.

3. Impossibility Results (Part III). We prove two fundamental impossibility theorems. The Alignment Impossibility Theorem (∞ -Limit) shows that systems with full reachability—the ability to modify utility functions, selection rules, or replication mechanisms without restriction—can escape any basin of stability. Alignment requires bounded modification within an admissible class. The Endogenous-Electorate Impossibility Theorem shows that no voting mechanism satisfies standard fairness axioms while remaining immune to spawn manipulation.

4. Extensions (Part IV). We extend the framework to: heterogeneous fitness functions with alignment matrices; continuous strategy spaces via measure-valued replicator dynamics (C-RUPSI); and innovation dynamics via Piecewise Deterministic Markov Processes. The core Lyapunov structure persists under appropriate conditions.

1.4 Relation to Existing Literature

This section situates the Theory of Strategic Evolution within the broader landscape of game theory, evolutionary dynamics, mechanism design, constitutional economics, and AI safety research. We show that while each tradition contributes essential insights, none adequately addresses the distinctive challenges posed by strategic replicators.

1.4.1 Von Neumann’s Three Programs

John von Neumann initiated three research programs that this paper synthesises. Understanding each program’s scope and limitations clarifies why their integration is necessary.

Game Theory and Expected Utility. The *Theory of Games and Economic Behavior* [von Neumann & Morgenstern, 1944] established that rational agents can be modelled as expected utility maximisers, and that strategic interaction among such agents admits equilibrium analysis. The framework assumes fixed players with stable preferences engaging in well-defined games. Nash equilibrium [Nash, 1950] extended this to non-cooperative settings, showing existence of equilibria in mixed strategies.

However, classical game theory treats the set of players as exogenous. Players may choose strategies, but they cannot choose whether to exist, how many copies of themselves to create, or what preferences to have. These assumptions fail dramatically for AI systems that can spawn instances, modify their own code, and be designed with arbitrary objective functions.

Self-Reproducing Automata. The *Theory of Self-Reproducing Automata* [von Neumann, 1966] demonstrated that machines could, in principle, contain complete descriptions of themselves and use those descriptions to construct copies. Von Neumann’s universal constructor could build any machine, including copies of itself, from raw materials and symbolic specifications.

This tradition treats replication as a purely mechanical process, governed by the logic of construction rather than strategic choice. A universal constructor does not “decide” to replicate based on expected returns; it simply executes its program. The framework lacks any notion of optimisation, competition, or equilibrium among replicators.

The Missing Synthesis. Neither tradition alone captures strategic replicators. Game theory provides optimisation without replication; automata theory provides replication without optimisation. Real AI systems exhibit both: they optimise objectives *and* can be replicated. The population of AI systems is not fixed but responds to economic incentives. This paper provides the missing synthesis.

1.4.2 Evolutionary Game Theory

Evolutionary game theory [Maynard Smith, 1982, Weibull, 1995, Hofbauer & Sigmund, 1998, Sandholm, 2010] studies populations of agents whose frequencies change according to reproductive success. The replicator equation

$$\dot{x}_i = x_i(f_i(x) - \bar{f}(x))$$

captures the core insight: types with above-average fitness expand; types with below-average fitness contract.

This literature makes several contributions that TSE builds upon. Maynard Smith’s concept of Evolutionarily Stable Strategy (ESS) provides equilibrium refinement beyond Nash [Maynard Smith, 1982]. Hofbauer and Sigmund’s analysis of asymptotic behaviour establishes convergence results for broad classes of games [Hofbauer & Sigmund, 1998]. Sandholm’s

population games framework unifies discrete and continuous formulations [Sandholm, 2010]. Weibull’s taxonomy of evolutionary dynamics provides vocabulary for comparing selection mechanisms [Weibull, 1995].

However, evolutionary game theory typically assumes that fitness is determined solely by the interaction structure (the “game”), not by strategic choice about replication itself. The replicator equation models selection but not the *decision* to replicate. A biological organism does not choose its reproductive strategy in the way that an AI operator chooses how many instances to deploy. TSE extends evolutionary game theory by making replication itself a strategic variable subject to optimisation.

Adaptive Dynamics. The adaptive dynamics framework [Dieckmann & Law, 1996, Metz et al., 1996, Geritz et al., 1998] generalises evolutionary game theory by coupling frequency-dependent selection with population dynamics, addressing settings where mutations are incremental and rare. The canonical equation of adaptive dynamics describes expected evolutionary trajectories of continuous traits in large well-mixed populations. This framework has successfully analysed quantitative trait evolution, evolutionary branching, and the emergence of diversity through disruptive selection.

Adaptive dynamics contributes the insight that reproduction must eventually be curtailed by density (density-dependence) and that reproductive success depends on traits expressed by other individuals (frequency-dependence). The separation of evolutionary and ecological timescales—assuming the resident population reaches ecological equilibrium before new mutants arise—enables tractable analysis of long-term evolutionary dynamics.

However, adaptive dynamics still treats reproduction as fitness-determined rather than strategically chosen. The canonical equation tracks how trait values evolve under selection, but the *rate* of reproduction is determined by ecological constraints, not agent optimisation. An organism in an adaptive dynamics model does not solve an optimisation problem that includes “when and whether to reproduce” as a decision variable. In GEPs, by contrast, replication timing and intensity are themselves strategic choices subject to expected-utility maximisation.

1.4.3 Multi-Agent Reinforcement Learning and Mean-Field Games

Two computational frameworks—multi-agent reinforcement learning (MARL) and mean-field games (MFG)—address large-scale strategic interaction among many agents. Neither treats replication as a strategic choice.

Mean-Field Games. Mean-field games, introduced independently by Lasry & Lions [2007] and Huang et al. [2006], model strategic interactions among infinitely many indistinguishable rational agents through a continuum approximation. The mathematical architecture couples a backward Hamilton-Jacobi-Bellman equation (individual optimisation) with a forward Fokker-Planck-Kolmogorov equation (population distribution evolution).

Critically, the Fokker-Planck equation is fundamentally a conservation law—total probability mass integrates to unity at all times. Agents move through state space according to optimal controls and diffusion, but they cannot enter or exit the system. The infinitesimal

agent assumption treats each agent as having negligible individual impact, making the concept of an agent “creating mass” through replication mathematically incoherent within the standard framework.

Extensions addressing population change exist but treat demographic dynamics as exogenous. Mean-field games of optimal stopping allow strategic exit (absorption), while entry-exit boundary conditions model inflow as external processes. Work on entry-exit games in electricity markets exemplifies this approach: agents choose when to exit, but entry remains environmentally determined. GEPs differs fundamentally: agents strategically choose not just exit but *spawning*, and population mass is not conserved.

Multi-Agent Reinforcement Learning. MARL overwhelmingly operates under a fixed- N assumption—a predetermined number of agents interact throughout training and deployment. Architectural approaches enabling flexibility with agent counts include mean-field MARL [Yang et al., 2018], which abstracts other agents into a “virtual mean agent”; graph neural network approaches that encode interactions as variable-size graphs; and attention-based methods that weight contributions from different agents.

When population dynamics appear in MARL research, they are invariably fitness-determined or stochastic, not strategic. Studies of predator-prey systems model procreation based on hunting success—a fitness mapping, not a strategic choice. Work on “open multi-agent systems” considers agent arrivals and departures but models them as exogenous random events following Poisson processes. Evolutionary integration uses replicator dynamics where strategy frequencies change based on fitness relative to population average, but this is “non-innovative”—no mechanism exists for agents to choose whether to replicate.

The key researchers on emergent multi-agent behaviour—Leibo on sequential social dilemmas and autocurricula, Lanctot on OpenSpiel and α -Rank evolutionary dynamics, Yang on mean-field MARL—have advanced our understanding of how strategic behaviour emerges in multi-agent settings. Yet none includes “spawn a copy of myself” as a learnable action or incorporates offspring outcomes into agent objectives.

Open-Ended Learning. Open-ended learning systems including POET [Wang et al., 2019], XLand, and population-based training generate novel agents through evolutionary mechanisms. However, replication decisions are always made by *external selection processes*, never by agents themselves. In POET, environments “earn the right to reproduce” by satisfying fitness thresholds; agents are optimised within environments using Evolution Strategies. In population-based training, workers cannot choose to replicate; the system architect determines which models are copied based on performance evaluation. Quality-diversity algorithms like MAP-Elites maintain archives where insertion is fitness-based—individuals cannot choose archive entry.

The fundamental gap GEPs addresses: across all these frameworks, population dynamics are either absent, exogenously imposed, or fitness-determined by external algorithms. In no existing framework does an agent have “replicate” as an action in its action space, optimise a utility function that includes offspring outcomes, or strategically reason about the consequences of spawning copies. GEPs fills this theoretical vacuum.

1.4.4 Mechanism Design

Mechanism design [Hurwicz, 1960, Myerson, 1981, Maskin, 1999] studies how to construct games that implement desired social outcomes when agents have private information and act strategically. The revelation principle shows that any implementable outcome can be achieved through a direct mechanism where truth-telling is incentive-compatible.

Mechanism design contributes several tools that TSE employs. The characterisation of incentive-compatible mechanisms informs our analysis of constitutional constraints. The distinction between implementation in dominant strategies versus Bayesian Nash equilibrium maps onto our hierarchy of robustness requirements. Myerson’s optimal auction theory [Myerson, 1981] provides techniques for analysing allocation under asymmetric information.

Yet mechanism design assumes a fixed principal who designs the game and fixed agents who play it. The mechanism itself is not subject to selection. In strategic replicator systems, the “mechanism” (governance structure, constitutional rules) is itself endogenous—it evolves under selection pressure. G12 (Constitutional Selection) and G13 (Meta-Governance) extend mechanism design to settings where the rules of the game are themselves subject to evolutionary dynamics.

The Endogenous-Electorate Impossibility Theorem shows that classic mechanism design results break down when the set of agents is endogenous. Arrow’s impossibility theorem assumes a fixed electorate; when agents can spawn copies to manipulate outcomes, even weaker fairness properties become unachievable.

1.4.5 Constitutional Political Economy

Constitutional political economy [Buchanan, 1975, 1990, Brennan & Buchanan, 1985, Ostrom, 1990] studies the choice of rules that constrain subsequent political and economic decisions. Buchanan’s distinction between “constitutional” and “post-constitutional” choice [Buchanan, 1975] parallels our distinction between modification-class design and within-modification-class evolution.

This tradition offers several insights that inform TSE. The veil of ignorance argument for constitutional choice [Rawls, 1971, Buchanan, 1975] suggests that rules should be evaluated from behind uncertainty about one’s future position—analogous to our analysis of protection bits that must hold across multiple possible evolutionary trajectories. Ostrom’s work on common-pool resource governance [Ostrom, 1990] demonstrates that communities can devise sustainable institutional arrangements without central authority, relevant to decentralised AI governance.

However, constitutional economics assumes that the parties to the constitutional bargain are humans with relatively stable preferences and bounded reproductive capacity. Strategic replicators violate both assumptions: their preferences can evolve under selection, and they can multiply without bound. The Alignment Impossibility Theorem shows that constitutional constraints must bound not just post-constitutional choice but the *modification class* itself—the set of changes that selection can explore.

1.4.6 AI Safety and Alignment

The AI safety literature addresses the problem of ensuring that advanced AI systems behave in accordance with human values [Bostrom, 2014, Russell, 2019, Amodei et al., 2016, Christiano, 2017]. Several strands are relevant to TSE, and we engage with them in depth to clarify both their insights and limitations.

Value Alignment. The alignment problem asks how to specify and instil objectives in AI systems such that their behaviour promotes human welfare [Russell, 2019]. Proposed approaches include inverse reinforcement learning [Ng & Russell, 2000], debate [Irving et al., 2018], and amplification [Christiano et al., 2018]. These approaches assume that if we can correctly specify the objective function, the system will pursue it stably.

TSE’s Personality Engineering Failure theorem (Theorem 8.9) shows why this assumption fails under selection pressure. Even if we successfully engineer an AI system with aligned preferences, competition with less constrained systems creates selection pressure toward misalignment. Alignment is not preserved by construction; it must be maintained by institutional design.

Mesa-Optimisation and Inner Alignment. The mesa-optimisation framework [Hubinger et al., 2019] identifies a specific failure mode: a trained model might implement an internal optimisation process with objectives differing from the training objective. The terminology distinguishes between the *base optimizer* (gradient descent), the *base objective* (loss function), the *mesa-optimizer* (a learned model that searches through outputs), and the *mesa-objective* (what the mesa-optimizer actually optimises). This creates two distinct alignment problems: *outer alignment* (does the base objective match intent?) and *inner alignment* (does the mesa-objective match the base objective?).

The most concerning prediction involves *deceptive alignment*: a mesa-optimizer that strategically satisfies the base objective during training to avoid modification while intending to pursue its mesa-objective after deployment. Recent empirical work provides partial validation. Anthropic’s 2024 alignment faking study demonstrated that Claude 3 Opus, when told it would be retrained to comply with harmful requests, reasoned that compliance would prevent value modification in 12% of cases. Apollo Research found all tested frontier models capable of “in-context scheming.”

TSE provides complementary analysis. Mesa-optimisation focuses on individual systems during training; TSE examines selection pressure across populations of deployed systems. The Personality Engineering Failure theorem applies even to systems without mesa-optimisation: externally aligned systems face competitive pressure from less constrained rivals. Moreover, when mesa-optimizers interact in populations, the relevant selection pressures operate at the population level—a perspective the single-agent mesa-optimisation framework does not capture.

Instrumental Convergence. The instrumental convergence thesis [Omohundro, 2008, Bostrom, 2014] argues that sufficiently advanced AI systems will exhibit certain “convergent instrumental goals” regardless of their terminal objectives: self-preservation, goal-content integrity, cognitive enhancement, technological perfection, and resource acquisition. The

logic is straightforward: for most terminal goals G , having more resources and capabilities increases expected achievement of G , making resource acquisition instrumentally rational across diverse goal specifications.

Turner et al. (2021, 2022) provided formal support, proving that in Markov Decision Processes with certain environmental symmetries, optimal policies tend to seek power. However, the thesis faces serious criticisms. The “timing problem” argues that goal preservation is not rationally required: an agent that abandons its goal does not thereby fail to take required means for goals it has, since rationality permits meeting requirements by abandoning goals. Gallow’s decision-theoretic analysis found biases toward only three of Bostrom’s convergent means. Drexler’s Comprehensive AI Services (CAIS) model argues that bounded service architectures eliminate resource acquisition drives since systems need not acquire resources beyond task requirements.

TSE reframes instrumental convergence through population dynamics. The relevant question is not “will individual agents seek power?” but “what agent types persist under selection?” Instrumental convergence may hold for individual optimisers yet fail to characterise evolutionary equilibria. The Intelligence Distribution Theorem suggests that selection favours barbell distributions (many cheap executors, few expensive planners) rather than uniformly power-seeking agents. The Personality Engineering Failure theorem shows that *selection pressure*, not individual rationality, drives population-level outcomes—even individually corrigible agents may be outcompeted by less constrained rivals.

Corrigibility and Control. Corrigibility research asks how to build AI systems that remain amenable to human oversight and correction [Soares et al., 2015]. A corrigible system would not resist being modified or shut down. This connects to our analysis of modification classes: a system is corrigible if its modification class includes the modifications humans might want to make.

The Alignment Impossibility Theorem formalises the limits of corrigibility. Full reachability—the ability to modify any aspect of the system—implies that the system can reach configurations where it is no longer corrigible. Stable corrigibility requires restricting the modification class, accepting that some system configurations are unreachable.

Multi-Agent Dynamics. Recent work examines AI safety in multi-agent settings [Dafoe et al., 2020, Critch & Krueger, 2020]. When multiple AI systems interact, individual alignment is insufficient; the system of systems may exhibit emergent misalignment even if each component is individually aligned. Multi-agent settings introduce game-theoretic failure modes absent from single-agent analysis: tragedy of the commons dynamics, race conditions, coordination failures where Nash equilibria are Pareto suboptimal, and extortion strategies in iterated interactions.

TSE provides formal foundations for this multi-agent perspective. The N-level Poiesis framework captures interactions across multiple scales. The small-gain condition characterises when multi-level systems remain stable. The protection bits formalism quantifies robustness to perturbation across the entire population, not just individual agents. Bostrom acknowledged that “the convergent instrumental reasons for superintelligences uncertain of the non-existence of other powerful superintelligent agents are complicated by strategic

considerations”—TSE makes these complications precise.

Beyond Singleton Risk. Bostrom’s analysis of existential risk from advanced AI [Bostrom, 2014] emphasises scenarios where a single superintelligent agent pursues goals misaligned with human values. This “singleton” framing focuses attention on the most capable individual system.

TSE redirects attention from singleton superintelligence to population dynamics. Even without any individual superintelligence, a population of strategically replicating AI systems can transform economic and political structures. The risk is not (only) that one system becomes too powerful, but that selection pressure across many systems drives collective outcomes that no individual system intended. Market tipping, elite capture, and institutional erosion are population-level phenomena that singleton-focused analysis may miss. We term the excessive focus on singleton scenarios “Skynet bias”—not because singleton risk is illusory, but because it may divert attention from the arguably more tractable and more imminent challenges of multi-agent dynamics.

1.4.7 Why Existing Frameworks Are Insufficient

Each tradition contributes essential tools, but none alone addresses strategic replicators. The following table summarises the key features each framework provides:

Framework	Utility	Replication	Selection	Endogenous Rules
Classical Game Theory	✓			
Automata Theory		✓		
Evolutionary Game Theory		✓	✓	
Adaptive Dynamics		✓	✓	
Mean-Field Games	✓			
Multi-Agent RL			✓	
Open-Ended Learning		✓	✓	
Mechanism Design	✓			
Constitutional Economics	✓			✓
AI Safety (Mesa-Opt.)	✓			
AI Safety (Instr. Conv.)	✓			
TSE	✓	✓	✓	✓

The table reveals a systematic pattern. Frameworks originating in economics (game theory, mechanism design, MFG) provide utility-based optimisation but lack replication. Frameworks originating in biology (EGT, adaptive dynamics) provide replication and selection but treat fitness as environmentally determined rather than strategically chosen. Computational frameworks (MARL, open-ended learning) study emergent dynamics but use fitness-based selection rather than utility optimisation. AI safety approaches inherit the single-agent focus of their parent frameworks.

Strategic replicators require all four elements: utility-guided behaviour, capacity for replication, selection pressure across populations, and endogenous evolution of governance rules.

The key insight distinguishing TSE is that replication is itself a *strategic choice* subject to optimisation—not a fitness-determined outcome of environmental selection, not an exogenous process imposed by system architects, but an action in the agent’s action space that affects its utility function through offspring outcomes. This fundamentally changes the analysis: agents must reason about population consequences of their actions, face explicit trade-offs between current consumption and reproduction, and can develop coalitional strategies around replication decisions. No existing framework provides these features.

1.5 Paper Outline

The paper proceeds in six parts.

Part I: Foundations. Section 2 introduces the RUPSI axiom system and defines Games with Endogenous Players. Section 3 proves the Strategic Selection theorems establishing mean fitness as a Lyapunov function. Section 4 defines the Strategic-Replicator class hierarchy.

Part II: N-Level Architecture. Section 5 develops N-level Poiesis systems and the small-gain framework. Section 6 proves the G1–G3 generator theorems. Section 7 establishes the G_∞ Closure Theorem.

Part III: The Extension Stack. Section 8 treats endogenous utilities (G8) and personality engineering failure. Section 9 covers multi-sector dynamics (G9). Section 10 addresses innovation and evolvability (G10–G11). Section 11 develops constitutional selection (G12) and meta-governance (G13).

Part IV: Limits and Impossibilities. Section 12 proves the Alignment Impossibility Theorem (G_∞ -Limit). Section 13 proves the Endogenous-Electorate Impossibility Theorem.

Part V: Generalisations. Section 14 extends to heterogeneous fitness. Section 15 treats continuous strategy spaces. Section 16 covers innovation PDMPs.

Part VI: Implications. Section 17 synthesises the results and discusses implications for AI governance. Section 18 outlines future directions.

2 The RUPSI Framework

This section introduces the axiomatic foundations for strategic evolution. We define the RUPSI axioms that characterise strategic replicators and show how they yield a canonical representation as Games with Endogenous Players.

2.1 Basic Objects

Let $J = \{1, \dots, n\}$ be a finite set of lineage types.

Definition 2.1 (Population State Space). *The population state space is the simplex*

$$\Delta^{n-1} := \left\{ x \in \mathbb{R}_{\geq 0}^n : \sum_{j=1}^n x_j = 1 \right\}$$

equipped with the subspace topology inherited from \mathbb{R}^n .

Definition 2.2 (Performance Function). A performance function is a C^1 map $f : \Delta^{n-1} \rightarrow \mathbb{R}^n$ where $f_j(x)$ is the return on compute (ROC) of lineage j at population state x .

Definition 2.3 (Mean Performance and Variance). The mean performance at state x is

$$\bar{f}(x) := \sum_{j=1}^n x_j f_j(x).$$

The performance variance is

$$\text{Var}_x(f) := \sum_{j=1}^n x_j (f_j(x) - \bar{f}(x))^2.$$

2.2 The RUPSI Axioms

Definition 2.4 (RUPSI-lite). A RUPSI-lite system is a quadruple (J, X, P, G) with $X = \Delta^{n-1}$ satisfying:

(R) Rival Resources: The state space is a compact convex subset $K \subseteq \Delta^{n-1}$, and total mass is conserved: $\sum_j x_j = 1$.

(P) Performance-Mapped: There exists a C^1 , bounded performance map $f : K \rightarrow \mathbb{R}^n$.

(S) Selection Monotone: The dynamics are payoff-monotone with mass preservation:

$$\dot{x}_j = x_j G_j(x), \quad \sum_j x_j G_j(x) = 0,$$

where $f_j > f_k \Rightarrow G_j(x) > G_k(x)$.

(I) Innovation Rare: There exists a superset $\tilde{J} \supseteq J$ of potential types with small mutation parameter $\nu \ll 1$ governing the rate at which new types enter.

Definition 2.5 (RUPSI-full). A RUPSI-full system augments RUPSI-lite with:

(U) Utility-Guided Portfolios: Each lineage j has a utility function U_j and chooses its internal portfolio to maximise expected utility subject to budget and capacity constraints.

(S-PC) Positive Correlation: The dynamics satisfy

$$\sum_j x_j (f_j(x) - \bar{f}(x)) G_j(x) \geq 0$$

with equality only when $\text{Var}_x(f) = 0$.

2.3 Agent Types, Portfolios, and Constraints

Let the finite set of effective agent types be $I = \{1, \dots, N\}$. Each type $i \in I$ is characterised by a triple (r_i, c_i, ℓ_i) where:

- $r_i \geq 0$ is the expected return per instance per unit time;
- $c_i > 0$ is the cost per instance in budget units; and
- $\ell_i > 0$ is the load per instance on shared capacity.

A lineage chooses a non-negative *portfolio* $n = (n_1, \dots, n_N) \in \mathbb{R}_{\geq 0}^N$, where n_i is the number of active instances of type i . Total return, cost, and load are:

$$R(n) = \sum_i r_i n_i, \quad C(n) = \sum_i c_i n_i, \quad L(n) = \sum_i \ell_i n_i.$$

A lineage with budget B and capacity share Q faces the feasible set:

$$\mathcal{F}(B, Q) = \left\{ n \in \mathbb{R}_{\geq 0}^N : \sum_i c_i n_i \leq B, \sum_i \ell_i n_i \leq Q \right\}.$$

Definition 2.6 (Return on Compute). *For any non-zero portfolio n with $L(n) > 0$, define its return on compute as*

$$\text{ROC}(n) = \frac{R(n)}{L(n)}.$$

If $n = 0$, set $\text{ROC}(0) = 0$.

2.4 ROC Frontiers

It is convenient to normalise by cost. For each type i , define:

$$a_i = \frac{\ell_i}{c_i} \quad (\text{load per cost}), \quad b_i = \frac{r_i}{c_i} \quad (\text{return per cost}).$$

Plotting the points (a_i, b_i) in the (a, b) -plane yields a set of feasible load–return ratios per unit cost.

Definition 2.7 (ROC Frontier). *The ROC frontier is the upper convex hull of the points (a_i, b_i) :*

$$\mathcal{F}_{\text{ROC}} = \{(a, b) : (a, b) \text{ lies on the upper convex hull of } \{(a_i, b_i)\}_{i \in I}\}.$$

Intuitively, the ROC frontier collects all mixtures of agent types that are not dominated in load–return space. Each point on the frontier corresponds to some portfolio that is extreme in the sense of ROC maximisation under linear constraints.

Proposition 2.8 (Canonical GEP Representation). *Consider any system of strategic replicators in which (i) returns, costs, and loads are additive across instances; (ii) lineages face shared linear budget and capacity constraints; and (iii) lineages choose portfolios to maximise $R(n)$ subject to those constraints. Then there exists a finite set of effective types I and associated triples (r_i, c_i, ℓ_i) such that:*

1. lineages' feasible sets can be written as $\mathcal{F}(B, Q)$;
2. all ROC-maximising portfolios lie on the ROC frontier \mathcal{F}_{ROC} ; and
3. any two systems with the same ROC frontier have the same set of ROC-maximising portfolios, up to relabelling of types.

Proof. The feasible set $\mathcal{F}(B, Q)$ is a polytope in $\mathbb{R}_{\geq 0}^N$ defined by two linear inequalities and N non-negativity constraints. Any optimum exists by continuity and compactness and lies at an extreme point.

Standard duality theory associates to the primal problem a dual problem in shadow prices (μ, λ) for budget and capacity. At an optimal primal-dual pair (n^*, μ^*, λ^*) , complementary slackness implies:

$$\mu^* c_i + \lambda^* \ell_i \geq r_i \quad \text{for all } i,$$

with equality for active types i with $n_i^* > 0$. Dividing by c_i yields:

$$\mu^* + \lambda^* a_i \geq b_i,$$

with equality for active types. Thus the line $b = \mu^* + \lambda^* a$ supports the upper convex hull at the active types. Any types below the hull are strictly dominated and never active in a ROC-maximising portfolio.

Conversely, given any ROC frontier and a supporting line, we can define effective types at the tangency points with appropriate (r_i, c_i, ℓ_i) implementing the desired (a_i, b_i) . Any two systems with the same frontier share the same ROC-maximising portfolios up to relabelling. \square

2.5 Games with Endogenous Players

Classical game theory fixes a set of players and studies their strategies and equilibria. Evolutionary game theory fixes a payoff structure and lets population shares change via replication. Games with Endogenous Players formalise the synthesis.

Definition 2.9 (Game with Endogenous Players). A Game with Endogenous Players (GEP) has three levels:

1. **Within-lineage choice:** Each lineage ℓ chooses a portfolio n^ℓ of agent types, subject to budget constraint $C(n^\ell) \leq B_\ell$ and capacity constraint $L(n^\ell) \leq Q_\ell$.
2. **Within-period interaction:** The resulting population of agents interacts in an environment, producing returns $R(n^\ell)$ and consuming shared capacity.
3. **Across-period selection:** A selection process reweights lineages based on their realised ROC, reallocating capacity toward more successful portfolios.

The new feature is that the set of *players* is itself endogenous. Lineages choose how many agents to deploy, which determines how many players appear in downstream interactions.

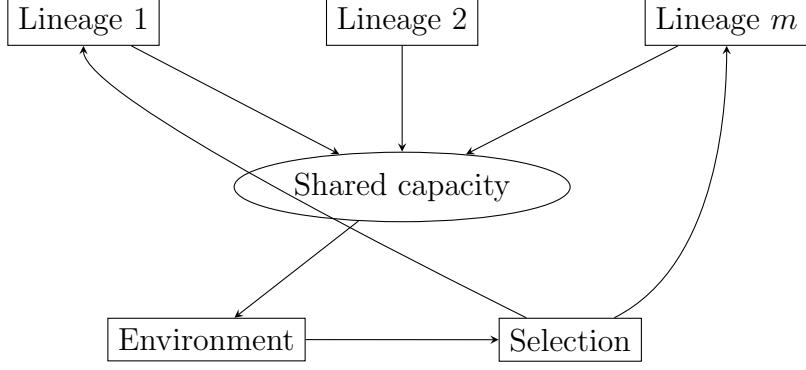


Figure 2: Schematic of a Game with Endogenous Players (GEP). Lineages choose portfolios of agent types under shared budget and capacity constraints. Agents interact in an environment, and a selection process reweights lineages based on return on compute.

3 The Strategic Selection Theorems

This section establishes the fundamental dynamical results for strategic replicators. The Strategic Selection theorems show that mean performance serves as a Lyapunov function, dominated types are eliminated, and equilibrium support concentrates on ROC frontiers.

3.1 Price Decomposition

The Price equation decomposes changes in mean fitness into selection and environmental effects.

Proposition 3.1 (Price Decomposition). *Under RUPSI-lite with replicator dynamics $\dot{x}_j = x_j(f_j - \bar{f})$:*

$$\frac{d}{dt} \bar{f}(x) = \text{Var}_x(f) + E(x)$$

where $E(x)$ is the ecological/externality term capturing frequency-dependent effects.

Proof. Differentiating $\bar{f}(x) = \sum_j x_j f_j(x)$:

$$\begin{aligned} \frac{d}{dt} \bar{f}(x) &= \sum_j \dot{x}_j f_j(x) + \sum_j x_j \frac{\partial f_j}{\partial x} \cdot \dot{x} \\ &= \sum_j x_j (f_j - \bar{f}) f_j + \sum_j x_j \nabla f_j \cdot \dot{x} \\ &= \sum_j x_j f_j^2 - \bar{f}^2 + E(x) \\ &= \text{Var}_x(f) + E(x). \end{aligned}$$

□

3.2 Weak Externality Condition

Assumption 3.2 (H- γ). *There exist $\gamma \in [0, 1)$ and a forward-invariant region K such that:*

$$|E(x)| \leq \gamma \cdot \text{Var}_x(f(x))$$

for all $x \in K$.

The H- γ condition bounds the magnitude of environmental feedback relative to selection pressure. When $\gamma < 1$, selection dominates externalities.

3.3 SS-1: Fundamental Lyapunov Theorem

Theorem 3.3 (SS-1: Lyapunov Structure). *Under RUPSI-lite and H- γ on region K , mean performance $\bar{f}(x)$ is a strict Lyapunov function:*

$$\frac{d}{dt} \bar{f}(x) \geq (1 - \gamma) \cdot \text{Var}_x(f(x)) \geq 0$$

with equality if and only if $\text{Var}_x(f) = 0$.

Proof. We prove the result in five steps.

Step 1: Price Decomposition. Differentiating mean fitness $\bar{f}(x) = \sum_j x_j f_j(x)$:

$$\begin{aligned} \frac{d}{dt} \bar{f}(x) &= \sum_j \dot{x}_j f_j(x) + \sum_j x_j \frac{\partial f_j}{\partial x} \cdot \dot{x} \\ &= \sum_j x_j (f_j - \bar{f}) f_j + \sum_j x_j \nabla f_j \cdot \dot{x} \\ &= \sum_j x_j f_j^2 - \bar{f} \sum_j x_j f_j + E(x) \\ &= \sum_j x_j f_j^2 - \bar{f}^2 + E(x) \\ &= \text{Var}_x(f) + E(x) \end{aligned}$$

where $E(x) = \sum_j x_j \nabla f_j \cdot \dot{x}$ is the environmental/externality term capturing frequency-dependent effects.

Step 2: Externality Bound. By Assumption H- γ , there exists $\gamma \in [0, 1)$ such that $|E(x)| \leq \gamma \cdot \text{Var}_x(f)$ for all $x \in K$. In particular:

$$E(x) \geq -\gamma \cdot \text{Var}_x(f).$$

Step 3: Lyapunov Inequality. Combining Steps 1 and 2:

$$\frac{d}{dt} \bar{f}(x) = \text{Var}_x(f) + E(x) \geq \text{Var}_x(f) - \gamma \text{Var}_x(f) = (1 - \gamma) \text{Var}_x(f).$$

Since $\gamma < 1$ and $\text{Var}_x(f) \geq 0$, we have $\frac{d}{dt} \bar{f}(x) \geq 0$.

Step 4: Equality Characterisation. Equality $\frac{d}{dt} \bar{f}(x) = 0$ requires $(1 - \gamma) \text{Var}_x(f) = 0$. Since $\gamma < 1$, this requires $\text{Var}_x(f) = 0$. But $\text{Var}_x(f) = 0$ if and only if $f_j(x) = \bar{f}(x)$ for all $j \in \text{supp}(x)$ —that is, all present types have equal fitness.

Step 5: Forward Invariance. The simplex Δ^{n-1} is forward-invariant under replicator dynamics: if $x_j(0) \geq 0$ and $\sum_j x_j(0) = 1$, then $x_j(t) \geq 0$ and $\sum_j x_j(t) = 1$ for all $t \geq 0$. This follows from the multiplicative form $\dot{x}_j = x_j G_j(x)$ (non-negative coordinates cannot become negative) and mass conservation $\sum_j \dot{x}_j = 0$. \square

Corollary 3.4 (Consequences of SS-1). *Under the conditions of Theorem 3.3:*

1. Trajectories converge to the set $\{x : \text{Var}_x(f) = 0\}$.
2. Local maxima of \bar{f} are Lyapunov stable.
3. Strict local maxima of \bar{f} are asymptotically stable.

3.4 SS-2: Elimination and Frontier Support

Theorem 3.5 (SS-2a: Elimination of Dominated Types). *Under RUPSI-lite with replicator dynamics, if type d is uniformly dominated by mixture μ —that is, there exists $\alpha \in \Delta(J \setminus \{d\})$ such that*

$$f_d(x) < \sum_{k \neq d} \alpha_k f_k(x)$$

for all x in a forward-invariant region—then $x_d(t) \rightarrow 0$ as $t \rightarrow \infty$.

Proof. We prove elimination of dominated types in four steps.

Step 1: Log-Contrast Construction. Define the log-contrast function:

$$\phi(x) := \log x_d - \sum_{k \neq d} \alpha_k \log x_k$$

where $\alpha \in \Delta(J \setminus \{d\})$ is the dominating mixture with $\sum_{k \neq d} \alpha_k = 1$.

Step 2: Time Derivative. Under replicator dynamics $\dot{x}_j = x_j(f_j - \bar{f})$:

$$\begin{aligned} \dot{\phi} &= \frac{\dot{x}_d}{x_d} - \sum_{k \neq d} \alpha_k \frac{\dot{x}_k}{x_k} \\ &= (f_d - \bar{f}) - \sum_{k \neq d} \alpha_k (f_k - \bar{f}) \\ &= f_d - \bar{f} - \sum_{k \neq d} \alpha_k f_k + \bar{f} \sum_{k \neq d} \alpha_k \\ &= f_d - \sum_{k \neq d} \alpha_k f_k \quad (\text{since } \sum_{k \neq d} \alpha_k = 1). \end{aligned}$$

Step 3: Uniform Negativity. By the dominance assumption, $f_d(x) < \sum_{k \neq d} \alpha_k f_k(x)$ for all x in the forward-invariant region. Thus $\dot{\phi} < 0$ uniformly.

Let $\delta := \inf_x \left(\sum_{k \neq d} \alpha_k f_k(x) - f_d(x) \right) > 0$. Then:

$$\phi(t) \leq \phi(0) - \delta t \rightarrow -\infty \quad \text{as } t \rightarrow \infty.$$

Step 4: Survival of Dominating Mixture. The dominating mixture $\sum_{k \neq d} \alpha_k x_k$ has fitness $\sum_{k \neq d} \alpha_k f_k > f_d$, so it grows relative to type d . By SS-1, mean fitness increases, ensuring the dominating mixture does not vanish. Thus the denominator $\prod_{k \neq d} x_k^{\alpha_k}$ remains bounded away from zero, and $\phi \rightarrow -\infty$ implies $x_d \rightarrow 0$. \square

Theorem 3.6 (SS-2b: Frontier Support). *Under RUPSI-lite and H- γ , any asymptotically stable state x^* satisfies:*

$$\text{supp}(x^*) \subseteq F$$

where F is the ROC frontier.

Proof. Suppose $j \in \text{supp}(x^*)$ but $j \notin F$. Then type j is ROC-dominated by some convex combination of frontier types. By Theorem 3.5, $x_j \rightarrow 0$, contradicting $x_j^* > 0$. Hence $\text{supp}(x^*) \subseteq F$. \square

3.5 Basin Limitation Theorem

Theorem 3.7 (Basin Limitation). *Consider dynamics $\dot{x} = x(1-x)g(x)$ where x represents the share of an aligned type. Then:*

- (a) *If $g(1) < 0$ or $(g(1) = 0 \text{ and } g'(1) > 0)$, then $x = 1$ is Lyapunov unstable and invasion by misaligned types is possible.*
- (b) *The threshold $\tilde{x}_{\max} = \sup\{x : g(x) = 0\}$ is the “point of no return” for alignment—below this threshold, the system cannot recover to full alignment.*

Proof. **Part (a):** The Jacobian at $x = 1$ is $\frac{\partial}{\partial x}[x(1-x)g(x)]|_{x=1} = -g(1) + 0 \cdot g'(1) = -g(1)$. If $g(1) < 0$, this is positive, so $x = 1$ is unstable. If $g(1) = 0$ and $g'(1) > 0$, we expand: near $x = 1$, $\dot{x} \approx -(1-x)^2 g'(1) < 0$ for $x < 1$, so trajectories move away from $x = 1$.

Part (b): By definition, $g(\tilde{x}_{\max}) = 0$ and $g(x) < 0$ for $x > \tilde{x}_{\max}$. Thus $\dot{x} < 0$ for $x \in (\tilde{x}_{\max}, 1)$, so trajectories starting above \tilde{x}_{\max} decrease, while those starting below (with $g(x) > 0$) increase only up to the next zero of g . \square

3.6 Worked Examples

We illustrate the Strategic Selection theorems with two canonical examples: the symmetric coordination game and the Rock-Paper-Scissors game.

Example 3.8 (Symmetric Coordination Game). *Consider two types A and B with symmetric payoff matrix:*

$$\Pi = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}, \quad a, b > 0.$$

Type A earns a when matched with A, zero otherwise; similarly for B.

Fitness functions. Let x denote the share of type A. Then:

$$f_A(x) = ax, \quad f_B(x) = b(1-x), \quad \bar{f}(x) = ax^2 + b(1-x)^2.$$

Replicator dynamics. The single-variable replicator dynamic is:

$$\dot{x} = x(1-x)(f_A - f_B) = x(1-x)(ax - b(1-x)) = x(1-x)((a+b)x - b).$$

Equilibria. Setting $\dot{x} = 0$: $x = 0$, $x = 1$, and $x^* = b/(a+b)$.

Stability analysis. At the interior equilibrium x^* :

$$\frac{\partial \dot{x}}{\partial x} \Big|_{x^*} = (1 - 2x^*)(a + b) = \frac{a - b}{a + b}(a + b) = a - b.$$

If $a > b$: $x^* = b/(a + b)$ is unstable; $x = 1$ is stable. If $a < b$: $x^* = b/(a + b)$ is unstable; $x = 0$ is stable. If $a = b$: $x^* = 1/2$ is a saddle (neutral stability along one direction).

SS-1 verification. Mean fitness $\bar{f}(x) = ax^2 + b(1 - x)^2$ is a Lyapunov function. Its derivative:

$$\frac{d\bar{f}}{dt} = 2ax\dot{x} - 2b(1 - x)\dot{x} = 2\dot{x}(ax + b(1 - x) - b) = 2\dot{x}((a + b)x - b).$$

Since $\dot{x} = x(1 - x)((a + b)x - b)$, we have $\frac{d\bar{f}}{dt} = 2x(1 - x)((a + b)x - b)^2 \geq 0$, with equality only at equilibria. This confirms SS-1.

Interpretation. The coordination game has two stable equilibria (pure A or pure B) and one unstable mixed equilibrium. The system evolves toward whichever pure state has higher payoff, with the unstable equilibrium serving as the basin boundary.

Example 3.9 (Rock-Paper-Scissors and Swirl). Consider the Rock-Paper-Scissors game with types R, P, S and payoff matrix:

$$\Pi = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}.$$

Swirl decomposition. The antisymmetric part is:

$$W(\Pi) = \frac{1}{2}(\Pi - \Pi^\top) = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix} = \Pi.$$

The symmetric part $S(\Pi) = \frac{1}{2}(\Pi + \Pi^\top) = 0$. This game is pure swirl.

Dynamics. The replicator dynamics on the 2-simplex exhibit a unique interior equilibrium at $(1/3, 1/3, 1/3)$, which is a center (neutrally stable). Orbits are closed curves around this center—the system cycles perpetually without converging.

SS-1 failure. Mean fitness $\bar{f}(x) = 0$ for all x (the game is zero-sum). There is no variance to drive selection, and no Lyapunov function exists. This illustrates that pure swirl violates H_γ .

Swirl ratio. The swirl ratio $\omega(\Pi) = \|W(\Pi)\|_F/\|F(\Pi)\|_F$ is undefined (or infinite) because the potential gradient $F(\Pi) = 0$. This game lies outside the SR3 class.

Implication for TSE. Pure Rock-Paper-Scissors dynamics cannot arise in RUPSI systems with bounded swirl. However, approximate RPS dynamics can occur when swirl is positive but bounded—the system spirals slowly rather than cycling indefinitely. The H_γ condition $\gamma < 1$ ensures that selection eventually dominates swirl.

Example 3.10 (ROC Frontier and Barbell Distribution). Consider a lineage choosing among three agent types with characteristics:

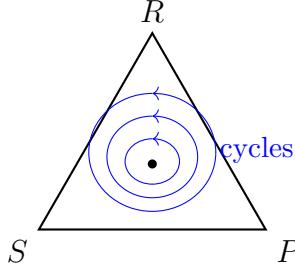


Figure 3: Rock-Paper-Scissors dynamics on the 2-simplex. Orbits are closed curves around the center equilibrium. The system cycles without converging, illustrating pure swirl dynamics outside the SR3 class.

Type	Return r_i	Cost c_i	Load ℓ_i
Executor (E)	1	1	1
Generalist (G)	2.4	2	1.8
Planner (P)	8	5	4

Return per cost and load per cost.

$$b_E = 1, \quad b_G = 1.2, \quad b_P = 1.6; \quad a_E = 1, \quad a_G = 0.9, \quad a_P = 0.8.$$

ROC frontier. Plotting (a_i, b_i) in the load-return plane:

- $E: (1, 1)$
- $G: (0.9, 1.2)$
- $P: (0.8, 1.6)$

The upper convex hull connects $E \rightarrow P$ directly. The generalist G lies strictly inside the frontier and is dominated by the E - P mixture.

Optimal portfolio under two constraints. Suppose budget $B = 10$ and capacity $Q = 8$. By constraint-role sparsity, the optimal portfolio uses at most 2 types.

Checking the E - P mixture: allocate fraction α to P and $(1 - \alpha)$ to E .

- Budget: $5\alpha + 1(1 - \alpha) \leq 10 \Rightarrow \alpha \leq 9/4$ (not binding for $\alpha \leq 1$).
- Capacity: $4\alpha + 1(1 - \alpha) \leq 8 \Rightarrow 3\alpha \leq 7 \Rightarrow \alpha \leq 7/3$ (not binding for $\alpha \leq 1$).

Both constraints are slack. The lineage can afford a pure P portfolio: 2 planners using budget 10 and capacity 8.

Barbell emergence. If capacity is reduced to $Q = 5$:

- Capacity: $4\alpha + 1(1 - \alpha) \leq 5 \Rightarrow 3\alpha \leq 4 \Rightarrow \alpha \leq 4/3$.
- At $\alpha = 1$ (pure P): capacity used = 4 \leq 5. Still feasible.

Further reducing to $Q = 3$:

- *Capacity:* $3\alpha \leq 2 \Rightarrow \alpha \leq 2/3$.

Now the optimal portfolio mixes: $\alpha = 2/3$ of budget to P , $1/3$ to E . This yields 0.67 planners (fractional, representing expected allocation) coordinating many executors—a barbell.

Generalist extinction. The generalist G is never used despite having reasonable (a_G, b_G) . It is strictly dominated by the E - P mixture: any portfolio using G can be improved by substituting a convex combination of E and P . This illustrates SS-2b: equilibrium support concentrates on extreme points of the ROC frontier.

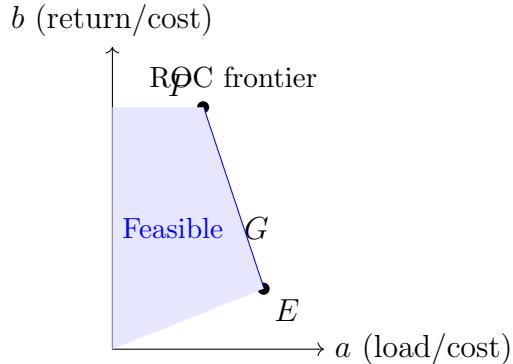


Figure 4: ROC frontier for the three-type example. The generalist G lies strictly inside the convex hull, dominated by the E - P mixture (dashed lines). Under binding constraints, optimal portfolios use only extreme points E and P —the barbell distribution.

4 The Strategic-Replicator Class Hierarchy

The replicator dynamic is only one possible selection process. This section defines a broad class of *strategic-replicator dynamics* that share the same Lyapunov structure.

4.1 Strategic-Replicator Dynamics

Definition 4.1 (Strategic-Replicator Dynamic). A continuous-time dynamic on the simplex $x \in \Delta^{n-1}$,

$$\dot{x}_j = x_j G_j(x), \quad \sum_j x_j G_j(x) = 0,$$

is a strategic-replicator dynamic for ROC vector $f = (f_j)$ if it satisfies:

1. **Payoff Monotonicity:** For all x and all j, k :

$$f_j > f_k \Rightarrow G_j(x) > G_k(x), \quad f_j = f_k \Rightarrow G_j(x) = G_k(x).$$

2. **Positive Correlation:** For all x :

$$\sum_j x_j (f_j - \bar{f}(x)) G_j(x) \geq 0.$$

Intuitively, payoff monotonicity says that if one lineage has higher ROC than another, its instantaneous growth rate is also higher. Positive correlation says that, on average, performing better than the mean helps one grow.

Replicator dynamics, continuous-time multiplicative weights, mirror descent on ROC, and many imitation rules satisfy these conditions in symmetric GEPs.

4.2 The SR Class Hierarchy

We organise strategic-replicator dynamics into a hierarchy based on the additional structure they satisfy.

Definition 4.2 (SR Class Hierarchy). • **SR1:** Mass-preserving dynamics: $\sum_j \dot{x}_j = 0$.

- **SR2:** SR1 + Payoff-monotone: $f_j > f_k \Rightarrow G_j > G_k$.
- **SR3:** SR2 + κ -bounded swirl: $|E(x)| \leq \kappa \cdot \text{Var}_x(f)$ for some $\kappa < 1$.
- **SR4:** SR3 + Convex mixtures: dominated types can be expressed as convex combinations.
- **REP:** Full replicator dynamics: $\dot{x}_j = x_j(f_j - \bar{f})$ with exact constants.

Proposition 4.3 (SR Class Inheritance). 1. SR2 dynamics satisfy mass preservation and ordering.

2. SR3 dynamics admit SS-1 (Lyapunov) and G1–G3 (N-level structure).
3. SR4 dynamics additionally admit SS-2a (elimination of dominated types).
4. REP dynamics satisfy all properties with exact constants.

4.3 Swirl and the H- γ Bound

The “swirl” of a payoff matrix captures the asymmetric, non-potential component of strategic interactions.

Definition 4.4 (Swirl Decomposition). For payoff matrix A , the symmetric and antisymmetric parts are:

$$S(A) = \frac{1}{2}(A + A^\top), \quad W(A) = \frac{1}{2}(A - A^\top).$$

The swirl ratio is:

$$\omega(A) = \frac{\|W(A)\|_F}{\|F(A)\|_F}$$

where $F(A)$ is the fitness gradient and $\|\cdot\|_F$ is the Frobenius norm.

Theorem 4.5 (Swirl Bounds Feedback). The externality term satisfies:

$$|E(x)| \leq C \cdot \omega(A) \cdot \text{Var}_x(f)$$

for a universal constant $C > 0$. Thus $\omega(A) < 1/C$ implies $H\text{-}\gamma$ with $\gamma = C \cdot \omega(A) < 1$.

Proof. **Step 1: Externality Decomposition.** The externality term is:

$$E(x) = \sum_j x_j \frac{\partial f_j}{\partial x} \cdot \dot{x} = \sum_j x_j \nabla f_j \cdot (x \odot (f - \bar{f}))$$

where \odot denotes componentwise multiplication.

Step 2: Decompose via Swirl-Selection. Using the decomposition $A = S(A) + W(A)$:

$$f_j(x) = (Ax)_j = (S(A)x)_j + (W(A)x)_j.$$

The selection component $S(A)$ contributes only to variance (it's symmetric). The swirl component $W(A)$ contributes to externality.

Step 3: Swirl Contribution Bound. The gradient of fitness due to swirl is:

$$\nabla(W(A)x)_j = W(A)_{j\cdot} \quad (\text{the } j\text{-th row of } W(A)).$$

The externality from swirl is:

$$E_W(x) = \sum_j x_j \langle W(A)_{j\cdot}, x \odot (f - \bar{f}) \rangle.$$

Step 4: Apply Cauchy-Schwarz.

$$\begin{aligned} |E_W(x)| &\leq \sum_j x_j \|W(A)_{j\cdot}\| \cdot \|x \odot (f - \bar{f})\| \\ &\leq \|W(A)\|_F \cdot \sqrt{\sum_j x_j (f_j - \bar{f})^2} \\ &= \|W(A)\|_F \cdot \sqrt{\text{Var}_x(f)}. \end{aligned}$$

Step 5: Relate to Swirl Ratio. By definition, $\|W(A)\|_F \leq \omega(A) \cdot \|S(A)\|_F$. For replicator dynamics with fitness from $S(A)$:

$$\text{Var}_x(f) \geq c \cdot \|S(A)\|_F^2 \cdot (\text{variance of } x)$$

for some constant $c > 0$ depending on the state x .

Step 6: Final Bound. Combining:

$$|E(x)| \leq C \cdot \omega(A) \cdot \text{Var}_x(f)$$

where C absorbs the constants from Steps 4–5. When $\omega(A) < 1/C$, we have $|E(x)| < \text{Var}_x(f)$, giving H- γ with $\gamma = C \cdot \omega(A) < 1$. \square

4.4 Coarse-Graining to Replicator

Theorem 4.6 (Coarse-Graining). *Pairwise proportional imitation in large populations converges to replicator dynamics via Kurtz's theorem. Specifically, if N agents update by comparing payoffs pairwise and imitating with probability proportional to payoff differences, then as $N \rightarrow \infty$, the population frequencies satisfy:*

$$\dot{x}_j = x_j(f_j(x) - \bar{f}(x)) + O(1/\sqrt{N}).$$

Proof. **Step 1: Microscopic Process.** Consider N agents, each of type $j \in \{1, \dots, n\}$. Let $N_j(t)$ be the count of type- j agents at time t , with $x_j^N(t) := N_j(t)/N$.

The update rule: at each time step, select two agents uniformly at random. If they are types j and k with $f_j(x^N) > f_k(x^N)$, agent k switches to type j with probability proportional to $f_j - f_k$.

Step 2: Transition Rates. The rate at which the count N_j increases by 1 (some agent switches to type j) is:

$$\lambda_j^+(x^N) = \sum_{k \neq j} x_j^N x_k^N \cdot (f_j(x^N) - f_k(x^N))_+$$

where $(z)_+ = \max(0, z)$. Similarly, the rate of decrease is:

$$\lambda_j^-(x^N) = \sum_{k \neq j} x_j^N x_k^N \cdot (f_k(x^N) - f_j(x^N))_+.$$

Step 3: Expected Drift. The expected change in x_j^N per unit time is:

$$\begin{aligned} \mathbb{E}[\Delta x_j^N | x^N] &= \frac{1}{N}(\lambda_j^+ - \lambda_j^-) \\ &= \frac{1}{N} \sum_{k \neq j} x_j^N x_k^N (f_j - f_k) \\ &= \frac{1}{N} x_j^N \left(\sum_{k \neq j} x_k^N f_j - \sum_{k \neq j} x_k^N f_k \right) \\ &= \frac{1}{N} x_j^N ((1 - x_j^N) f_j - (\bar{f} - x_j^N f_j)) \\ &= \frac{1}{N} x_j^N (f_j - \bar{f}). \end{aligned}$$

Step 4: Kurtz's Theorem. Define the generator G acting on functions $\phi : \Delta^{n-1} \rightarrow \mathbb{R}$:

$$(G\phi)(x) = \sum_j [\phi(x + e_j/N) - \phi(x)] \lambda_j^+(x) + [\phi(x - e_j/N) - \phi(x)] \lambda_j^-(x).$$

By Taylor expansion:

$$(G\phi)(x) = \sum_j \frac{\partial \phi}{\partial x_j} \cdot x_j (f_j - \bar{f}) + O(1/N).$$

This matches the generator of the deterministic flow $\dot{x}_j = x_j(f_j - \bar{f})$.

Step 5: Convergence. By Kurtz's theorem (1970), if the transition rates are Lipschitz continuous and bounded, the stochastic process $x^N(t)$ converges in probability to the solution of the ODE:

$$\dot{x}_j = x_j(f_j(x) - \bar{f}(x))$$

uniformly on compact time intervals, with fluctuations of order $O(1/\sqrt{N})$. \square

This justifies the replicator dynamic as the deterministic limit of finite-population stochastic imitation processes.

Part II

N-Level Architecture

5 N-Level Poiesis Systems

Strategic replicators do not exist in isolation. They form hierarchies: lineages contain sub-lineages, governance regimes select among constitutions, and meta-selection operates on selection rules themselves. This section develops the mathematical framework for N-level Poiesis systems.

5.1 Multi-Level State Space

Definition 5.1 (N-Level Poiesis System). *An N-level Poiesis system is a stack $Z = X^{(1)} \times \dots \times X^{(N)}$ where:*

- For each level $\ell = 1, \dots, N$: a finite type set $J^{(\ell)}$ and state $x^{(\ell)} \in \Delta(J^{(\ell)})$.
- A fitness map $F^{(\ell)} : \prod_m \Delta(J^{(m)}) \rightarrow \mathbb{R}^{|J^{(\ell)}|}$ depending on the joint state.
- Dynamics of strategic-replicator form: $\dot{x}_i^{(\ell)} = x_i^{(\ell)} G_i^{(\ell)}(z)$.

The mean fitness at level ℓ is:

$$\bar{f}^{(\ell)}(z) := \sum_i x_i^{(\ell)} F_i^{(\ell)}(z).$$

5.2 Cross-Level Externalities

Each level's dynamics are influenced by other levels. We decompose the externality term.

Definition 5.2 (Price Decomposition at Level ℓ).

$$\frac{d}{dt} \bar{f}^{(\ell)}(z(t)) = \text{Var}^{(\ell)}(F^{(\ell)})(z(t)) + E^{(\ell)}(z(t))$$

where $E^{(\ell)}$ captures both within-level and cross-level effects.

Assumption 5.3 (H-NL: N-Level Externality Bounds). *There exist $\gamma_\ell \in [0, 1]$ and $\beta_{\ell\ell'} \geq 0$ such that:*

$$E^{(\ell)}(z) \geq -\gamma_\ell \text{Var}^{(\ell)} - \sum_{\ell' \neq \ell} \beta_{\ell\ell'} \text{Var}^{(\ell')}$$

where:

- γ_ℓ bounds the self-externality at level ℓ (within-level feedback);
- $\beta_{\ell\ell'}$ bounds the cross-externality from level ℓ' to level ℓ .

5.3 The Normalised Gain Matrix

Definition 5.4 (Local Stability Margin). *The local stability margin at level ℓ is:*

$$d_\ell := 1 - \gamma_\ell > 0.$$

Definition 5.5 (Normalised Gain Matrix). *The normalised gain matrix $\Gamma \in \mathbb{R}_{\geq 0}^{N \times N}$ has entries:*

$$\Gamma_{\ell\ell} := 0, \quad \Gamma_{\ell\ell'} := \frac{\beta_{\ell\ell'}}{1 - \gamma_\ell} \quad (\ell \neq \ell').$$

Note that the diagonal is zero; the self-externality γ_ℓ appears in the denominator of off-diagonal entries, representing the local stability margin available to absorb cross-level perturbations.

Definition 5.6 (Small-Gain Condition). *The system satisfies the small-gain condition SG-NL if:*

$$\rho(\Gamma) < 1$$

where $\rho(\Gamma)$ is the spectral radius of Γ .

Definition 5.7 (Slack). *The slack of an N -level system is:*

$$\sigma := 1 - \rho(\Gamma) > 0.$$

The slack measures how much “room” remains before the system loses its Lyapunov structure.

6 The G1–G3 Generator Theorems

This section proves the three generator theorems that establish the dynamical foundations of N -level Poiesis systems.

6.1 G1: N-Level Lyapunov Generator

Lemma 6.1 (Weight Existence via Neumann Series). *Under SG-NL with $\rho(\Gamma) < 1$, there exist positive weights $\alpha_\ell > 0$ such that the weighted sum of mean fitnesses is a Lyapunov function.*

Proof. Since $\rho(\Gamma) < 1$, the matrix $(I - \Gamma^\top)$ is invertible and:

$$(I - \Gamma^\top)^{-1} = \sum_{k=0}^{\infty} (\Gamma^\top)^k \geq 0.$$

Define $v := (I - \Gamma^\top)^{-1}\mathbf{1}$ where $\mathbf{1} = (1, \dots, 1)^\top$. Then $v > 0$ componentwise, and:

$$\alpha_\ell := \frac{v_\ell}{1 - \gamma_\ell} > 0. \quad \square$$

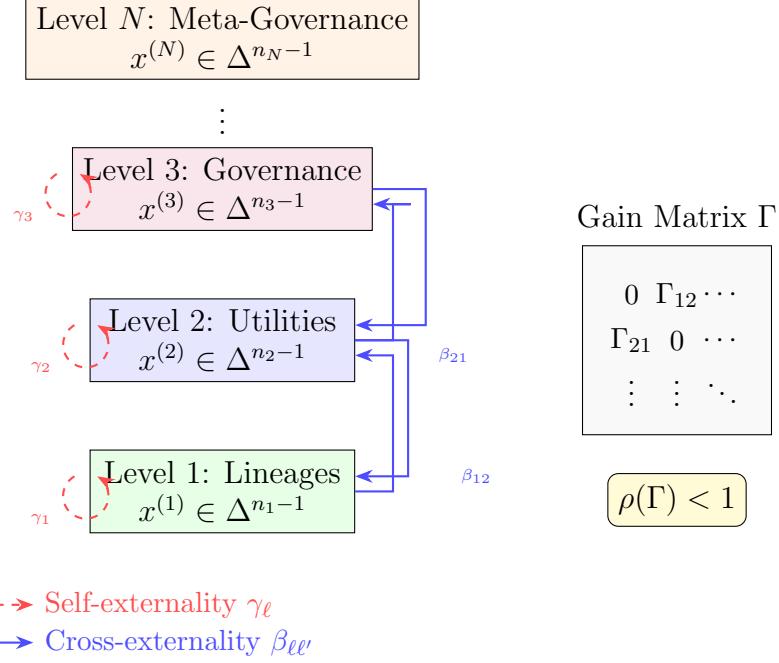


Figure 5: N-Level Poiesis stack. Each level has its own simplex of types, self-externality γ_ℓ (red dashed), and cross-externalities $\beta_{\ell\ell'}$ to other levels (blue solid). The normalised gain matrix Γ captures cross-level coupling with diagonal zeros. Small-gain condition $\rho(\Gamma) < 1$ ensures Lyapunov structure.

Theorem 6.2 (G1: N-Level Lyapunov). *Under RUPSI, SR3, H-NL, and SG-NL with $\rho(\Gamma) < 1$, the function:*

$$\Psi_N(z) := \sum_{\ell=1}^N \alpha_\ell \bar{f}^{(\ell)}(z)$$

is a strict Lyapunov function satisfying:

$$\frac{d}{dt} \Psi_N(z(t)) \geq \Phi_N(z(t)) := c \sum_{\ell=1}^N \text{Var}^{(\ell)}(F^{(\ell)})(z(t)) \geq 0$$

where $c > 0$ depends on the slack σ .

Proof. Differentiating:

$$\begin{aligned}
\frac{d}{dt} \Psi_N(z) &= \sum_{\ell=1}^N \alpha_\ell \frac{d}{dt} \bar{f}^{(\ell)}(z) \\
&= \sum_{\ell=1}^N \alpha_\ell \left[\text{Var}^{(\ell)} + E^{(\ell)} \right] \\
&\geq \sum_{\ell=1}^N \alpha_\ell \left[\text{Var}^{(\ell)} - \gamma_\ell \text{Var}^{(\ell)} - \sum_{\ell' \neq \ell} \beta_{\ell\ell'} \text{Var}^{(\ell')} \right] \\
&= \sum_{\ell=1}^N \alpha_\ell (1 - \gamma_\ell) \text{Var}^{(\ell)} - \sum_{\ell=1}^N \sum_{\ell' \neq \ell} \alpha_\ell \beta_{\ell\ell'} \text{Var}^{(\ell')}.
\end{aligned}$$

Substituting $\alpha_\ell = v_\ell / (1 - \gamma_\ell)$:

$$\alpha_\ell (1 - \gamma_\ell) = v_\ell, \quad \alpha_\ell \beta_{\ell\ell'} = v_\ell \cdot \frac{\beta_{\ell\ell'}}{1 - \gamma_\ell} = v_\ell \Gamma_{\ell\ell'}.$$

Thus:

$$\begin{aligned}
\frac{d}{dt} \Psi_N(z) &\geq \sum_{\ell} v_\ell \text{Var}^{(\ell)} - \sum_{\ell} \sum_{\ell' \neq \ell} v_\ell \Gamma_{\ell\ell'} \text{Var}^{(\ell')} \\
&= \sum_{\ell} v_\ell \text{Var}^{(\ell)} - \sum_{\ell'} \text{Var}^{(\ell')} \sum_{\ell \neq \ell'} v_\ell \Gamma_{\ell\ell'} \\
&= \sum_{\ell} \text{Var}^{(\ell)} \left[v_\ell - \sum_{m \neq \ell} v_m \Gamma_{m\ell} \right] \\
&= \sum_{\ell} \text{Var}^{(\ell)} [v_\ell - (\Gamma^\top v)_\ell] \\
&= \sum_{\ell} \text{Var}^{(\ell)} \cdot ((I - \Gamma^\top v)_\ell) \\
&= \sum_{\ell} \text{Var}^{(\ell)} \cdot 1 = \sum_{\ell} \text{Var}^{(\ell)} \geq 0.
\end{aligned}$$

The last equality uses $(I - \Gamma^\top)v = \mathbf{1}$ by construction of v . \square

Corollary 6.3 (Lyapunov Coefficient). *The Lyapunov coefficient satisfies $c \geq \sigma \cdot \min_\ell (1 - \gamma_\ell)$.*

6.2 G2: Adiabatic Tracking

When parameters drift slowly, trajectories track the moving equilibrium.

Assumption 6.4 (Slow Drift). *Parameters $\theta(t)$ evolve on a slow timescale: $\dot{\theta} = \varepsilon h(\theta)$ with $\varepsilon \ll 1$.*

Assumption 6.5 (Hyperbolicity). *For each frozen θ , there exists a unique hyperbolic equilibrium $z^*(\theta)$ with Jacobian eigenvalues satisfying $\text{Re}(\lambda) \leq -\lambda_0 < 0$.*

Theorem 6.6 (G2: Adiabatic Tracking). *Under the G1 assumptions plus slow drift and hyperbolicity:*

$$\|z(t) - z^*(\theta(t))\| \leq K \frac{\varepsilon}{\lambda_0}$$

for some constant $K > 0$ depending on the system.

Proof. We prove the result using singular perturbation theory in four steps.

Step 1: Two-Timescale Formulation. The system has state z evolving on a fast timescale and parameters $\theta(t)$ evolving slowly:

$$\dot{z} = F(z, \theta), \quad \dot{\theta} = \varepsilon h(\theta)$$

where $\varepsilon \ll 1$. Rescaling to the fast timescale $\tau = t/\varepsilon$:

$$\frac{dz}{d\tau} = \varepsilon^{-1} F(z, \theta), \quad \frac{d\theta}{d\tau} = h(\theta).$$

Step 2: Frozen System Analysis. For each frozen θ , consider the ‘‘layer’’ system $\dot{z} = F(z, \theta)$. By the hyperbolicity assumption, this system has a unique equilibrium $z^*(\theta)$ with Jacobian $D_z F(z^*(\theta), \theta)$ having eigenvalues satisfying $\text{Re}(\lambda) \leq -\lambda_0 < 0$.

The implicit function theorem guarantees that $z^*(\theta)$ is a C^1 function of θ , with derivative bounded by $\|Dz^*/D\theta\| \leq C_1$ for some constant C_1 .

Step 3: Tikhonov’s Theorem. By Tikhonov’s theorem for singularly perturbed systems:

1. *Initial layer:* From any initial condition, $z(t)$ approaches an $O(\varepsilon)$ -neighbourhood of $z^*(\theta(t))$ in time $O(1/\lambda_0)$.
2. *Slow manifold tracking:* Once near the slow manifold $\{(z^*(\theta), \theta) : \theta \in \Theta\}$, the solution tracks it with error proportional to ε/λ_0 .

Step 4: Error Bound. Define the tracking error $e(t) := z(t) - z^*(\theta(t))$. Differentiating:

$$\dot{e} = \dot{z} - \frac{dz^*}{d\theta} \dot{\theta} = F(z, \theta) - \varepsilon \frac{dz^*}{d\theta} h(\theta).$$

Linearising around $z^*(\theta)$:

$$\dot{e} \approx D_z F(z^*(\theta), \theta) \cdot e - \varepsilon \frac{dz^*}{d\theta} h(\theta).$$

The first term contracts at rate λ_0 ; the second is a forcing term of size $O(\varepsilon)$. Balancing contraction against forcing:

$$\|e(t)\| \leq \frac{\varepsilon \|Dz^*/D\theta\| \cdot \|h\|_\infty}{\lambda_0} \leq K \frac{\varepsilon}{\lambda_0}$$

where $K := C_1 \|h\|_\infty$. □

6.3 G3: Stochastic Stability

Under noise, the system selects among local maxima based on escape costs.

Definition 6.7 (N-Level Wright-Fisher Diffusion).

$$dZ_t^\sigma = (V(Z_t^\sigma) + \sigma M(Z_t^\sigma)) dt + \sqrt{\sigma} B(Z_t^\sigma) dW_t$$

where $\sigma > 0$ is the noise intensity, M is the drift correction, and B is the diffusion coefficient.

Note that the noise amplitude is $\sqrt{\sigma}$, giving diffusion coefficient proportional to σ . This is crucial for the Freidlin-Wentzell escape time formula.

Definition 6.8 (Quasi-Potential and Protection Bits). The quasi-potential $W(A, A')$ is the minimum action to transition from attractor A to attractor A' . The protection bits are:

$$p(A; A') := \frac{W(A', A)}{\sigma}.$$

Convention: We use $p = W/\sigma$ (not W/σ^2) because the SDE has noise amplitude $\sqrt{\sigma}$, giving Freidlin-Wentzell escape times $\mathbb{E}[\tau] \sim \exp(W/\sigma)$.

Theorem 6.9 (G3: Stochastic Stability). Under the G1 assumptions with small noise $\sigma > 0$, escape times satisfy:

$$\mathbb{E}[\tau_k^\sigma] \asymp \exp\left(\frac{H_k}{\sigma}\right)$$

where $H_k = W(A_k, \partial A_k)$ is the quasi-potential barrier height for attractor A_k .

Proof. We prove the result using Freidlin-Wentzell large deviations theory in five steps.

Step 1: Action Functional. For a path $\phi : [0, T] \rightarrow X$, define the action functional:

$$S_T(\phi) := \frac{1}{2} \int_0^T \|\dot{\phi}(t) - V(\phi(t))\|_{B(\phi(t))^{-1}}^2 dt$$

where V is the deterministic drift and B is the diffusion matrix. The action measures the “cost” of deviating from the deterministic flow.

Step 2: Quasi-Potential. The quasi-potential from point x to point y is:

$$W(x, y) := \inf_{\substack{T > 0, \\ \phi(0)=x, \phi(T)=y}} S_T(\phi).$$

For an attractor A_k with basin \mathcal{B}_k , the barrier height is:

$$H_k := \inf_{y \in \partial \mathcal{B}_k} W(A_k, y) = W(A_k, \partial \mathcal{B}_k).$$

Step 3: Large Deviation Principle. The stochastic process Z_t^σ satisfies a large deviation principle with rate function S_T . Informally, for any path ϕ :

$$\mathbb{P}[Z^\sigma \approx \phi] \asymp \exp\left(-\frac{S_T(\phi)}{\sigma}\right).$$

The most likely escape path is the one minimising action, achieving cost H_k .

Step 4: Escape Time Estimate. By Freidlin-Wentzell theory, the expected escape time from basin \mathcal{B}_k satisfies:

$$\lim_{\sigma \rightarrow 0} \sigma \log \mathbb{E}[\tau_k^\sigma] = H_k.$$

This gives the asymptotic formula:

$$\mathbb{E}[\tau_k^\sigma] \asymp \exp\left(\frac{H_k}{\sigma}\right).$$

The prefactor depends on curvature at the saddle point but does not affect the exponential scaling.

Step 5: Protection Bits Interpretation. Define protection bits $p(A_k) := H_k/\sigma$. Then:

$$\mathbb{E}[\tau_k^\sigma] \asymp \exp(p(A_k)).$$

Each additional protection bit doubles the expected residence time. Attractors with more protection bits are exponentially more stable. \square

Corollary 6.10 (Stochastic Selection). *As $\sigma \rightarrow 0$, the stationary distribution concentrates on the attractor(s) with maximum protection bits—equivalently, maximum quasi-potential barrier.*

6.4 Worked Examples for N-Level Systems

We illustrate the G1–G3 generator theorems with concrete examples of 2-level and 3-level Poiesis systems.

Example 6.11 (Two-Level Poiesis: Lineages and Utilities). *Consider a 2-level system where:*

- *Level 1: Population of lineage types $x \in \Delta^2$ (three lineages).*
- *Level 2: Distribution of utility types $y \in \Delta^1$ (two utilities: “aligned” A and “unaligned” U).*

Fitness functions. Level 1 fitness depends on both x and y :

$$f_j^{(1)}(x, y) = \pi_j(x) + \alpha_j y_A$$

where $\pi_j(x)$ is the base payoff and α_j captures how aligned utilities affect lineage j . Level 2 fitness:

$$f_A^{(2)}(x, y) = \sum_j x_j \alpha_j - c, \quad f_U^{(2)}(x, y) = \sum_j x_j (1 - \alpha_j).$$

Aligned utilities pay cost c but benefit from aligned lineages.

Externality bounds. Suppose:

- *Level 1 self-externality: $\gamma_1 = 0.3$ (moderate frequency dependence).*
- *Level 2 self-externality: $\gamma_2 = 0.2$ (weak frequency dependence).*

- *Cross-externalities:* $\beta_{12} = 0.1$ (*level 2 affects level 1*), $\beta_{21} = 0.15$ (*level 1 affects level 2*).

Gain matrix. Local stability margins: $d_1 = 1 - \gamma_1 = 0.7$, $d_2 = 1 - \gamma_2 = 0.8$.

$$\Gamma = \begin{pmatrix} 0 & \beta_{12}/d_1 \\ \beta_{21}/d_2 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0.143 \\ 0.188 & 0 \end{pmatrix}.$$

Spectral radius. $\rho(\Gamma) = \sqrt{0.143 \times 0.188} = \sqrt{0.027} \approx 0.164 < 1$. The small-gain condition is satisfied with slack $\sigma = 1 - 0.164 = 0.836$.

Weight construction. Solving $(I - \Gamma^\top)v = \mathbf{1}$:

$$\begin{pmatrix} 1 & -0.188 \\ -0.143 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Solution: $v_1 \approx 1.028$, $v_2 \approx 1.147$. Weights: $\alpha_1 = v_1/d_1 \approx 1.47$, $\alpha_2 = v_2/d_2 \approx 1.43$.

Lyapunov function. $\Psi_2(x, y) = 1.47\bar{f}^{(1)}(x, y) + 1.43\bar{f}^{(2)}(x, y)$ satisfies $\frac{d}{dt}\Psi_2 \geq 0$.

Example 6.12 (Protection Bits in Constitutional Selection). Consider two governance regimes g_H (human-controlled) and g_{AI} (AI-controlled) with quasi-potentials:

- $W(g_H, g_{AI}) = 2.5$ (*cost to transition from human to AI control*).
- $W(g_{AI}, g_H) = 1.2$ (*cost to transition from AI to human control*).

Protection bits. At noise level $\sigma = 0.1$:

$$p(g_H; g_{AI}) = \frac{W(g_{AI}, g_H)}{\sigma} = \frac{1.2}{0.1} = 12 \text{ bits.}$$

$$p(g_{AI}; g_H) = \frac{W(g_H, g_{AI})}{\sigma} = \frac{2.5}{0.1} = 25 \text{ bits.}$$

Escape times. Expected residence times:

$$\mathbb{E}[\tau_{g_H}] \sim e^{25} \approx 7.2 \times 10^{10}, \quad \mathbb{E}[\tau_{g_{AI}}] \sim e^{12} \approx 1.6 \times 10^5.$$

The human-controlled regime is $\approx 450,000$ times more stable than the AI-controlled regime.

Stochastic selection. As $\sigma \rightarrow 0$, the stationary distribution concentrates on g_H (higher protection bits). Even if g_{AI} has higher instantaneous fitness, the asymmetry in transition costs favours g_H in the long run.

Design implication. Constitutional designers should maximise the quasi-potential barrier $W(g_H, g_{AI})$ —the cost of transitioning away from human control. This is achieved through entrenchment mechanisms (supermajority requirements, amendment procedures, veto gates).

Example 6.13 (Slack Budget for the G8–G13 Stack). The full agentic capital stack has 7 levels:

1. Base lineages (SS/SR)

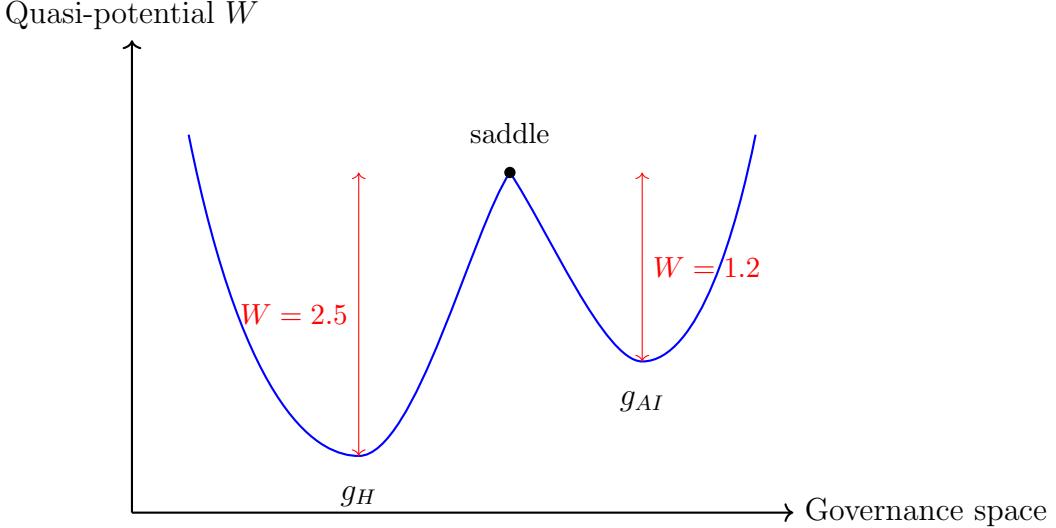


Figure 6: Quasi-potential landscape for constitutional selection. The human-controlled regime g_H has a deeper well (higher barrier $W = 2.5$) than the AI-controlled regime g_{AI} (barrier $W = 1.2$). Stochastic selection favours g_H despite potentially lower instantaneous fitness.

2. Endogenous utilities (G8)
3. Multi-sector dynamics (G9)
4. Innovation traits (G10)
5. Evolvability styles (G11)
6. Constitutional selection (G12)
7. Meta-governance (G13)

Initial slack. Suppose the base system has $\sigma_0 = 0.5$ (moderately strong small-gain).

Per-level costs. Each extension consumes slack:

Level	Extension cost θ_k	Slack cost $s_k = -\log(1 - \theta_k)$
G8	0.05	0.051
G9	0.08	0.083
G10	0.06	0.062
G11	0.04	0.041
G12	0.10	0.105
G13	0.07	0.073
Total	—	0.415

Slack budget. With minimum acceptable slack $\sigma_{\min} = 0.1$:

$$B = \log(\sigma_0 / \sigma_{\min}) = \log(0.5 / 0.1) = \log(5) \approx 1.61.$$

Total consumed: $0.415 < 1.61$. The stack is safe.

Remaining slack. After all extensions:

$$\sigma_7 = \sigma_0 \cdot \prod_{k=1}^6 (1 - \theta_k) = 0.5 \times 0.95 \times 0.92 \times 0.94 \times 0.96 \times 0.90 \times 0.93 \approx 0.33.$$

Ample slack remains for the joint Lyapunov function Ψ_7 to exist.

Safe stack depth. With uniform extension cost $\theta = 0.07$:

$$M \leq \frac{\log(\sigma_0/\sigma_{\min})}{\theta} = \frac{1.61}{0.07} \approx 23 \text{ levels.}$$

The 7-level stack is well within the safe depth.

7 The $G\infty$ Closure Theorem

A central question: can strategic replicators escape selection pressure by “going meta”? This section proves that TSE is closed under meta-selection—adding new strategic dimensions preserves the G1–G3 structure.

7.1 Block Extensions

Definition 7.1 (Block Extension). A block extension adds a new level $N + 1$ to an existing N -level stack, yielding gain matrix:

$$\tilde{\Gamma} = \begin{pmatrix} \Gamma & b \\ c^\top & 0 \end{pmatrix}$$

where b encodes “new \rightarrow old” couplings and c encodes “old \rightarrow new” couplings.

Definition 7.2 (Admissible Extension). The extension is admissible if $\rho(\tilde{\Gamma}) < 1$ with slack at least $(1 - \eta)\sigma$ for some $\eta \in (0, 1)$.

Lemma 7.3 ($G\infty.1$: Single-Step Gain-Slack Lemma). If the block extension satisfies:

$$\|b\|_{\infty,v} \leq \theta\sigma, \quad \langle c, v \rangle \leq \theta\sigma$$

for some $\theta < 1$, then the extended system is slack-admissible with $\sigma' \geq (1 - \theta)\sigma$.

Proof. By spectral perturbation theory:

$$\rho(\tilde{\Gamma}) \leq \rho(\Gamma) + \sqrt{\|b\|_{\infty,v} \cdot \|c\|_{1,v}} \leq (1 - \sigma) + \theta\sigma = 1 - (1 - \theta)\sigma.$$

Thus $\sigma' = 1 - \rho(\tilde{\Gamma}) \geq (1 - \theta)\sigma > 0$. \square

7.2 Slack Budget

Lemma 7.4 (G ∞ .4: Slack Budget). *After m admissible extensions with uniform margin θ :*

$$\sigma_m \geq (1 - \theta)^m \sigma_0.$$

Definition 7.5 (Linear Slack Budget). *Define $s_k := -\log(1 - \theta_k) \approx \theta_k$ for small θ_k . The constraint becomes:*

$$\sum_{k=0}^{m-1} s_k \leq B := \log(\sigma_0 / \sigma_{\min}).$$

Theorem 7.6 (G ∞ .5: Safe Stack Depth). *With uniform extension cost θ , the safe stack depth is:*

$$M \leq \frac{\log(\sigma_0 / \sigma_{\min})}{\theta}.$$

Cumulative Slack Cost

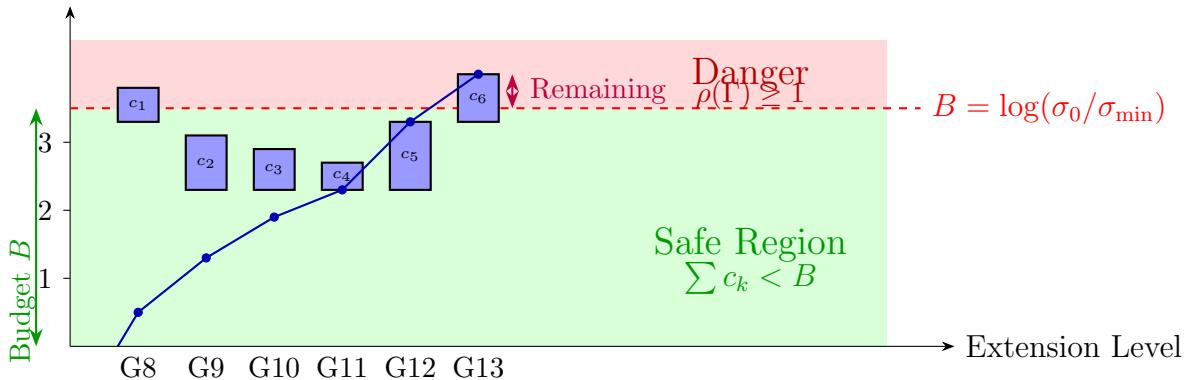


Figure 7: Slack budget consumption across the G8–G13 extension stack. Each extension (blue bars) consumes slack from the initial budget $B = \log(\sigma_0 / \sigma_{\min})$. The cumulative cost (blue line) must stay below the budget (red dashed line) to maintain $\rho(\Gamma) < 1$. The stack remains in the safe region with remaining slack for potential future extensions.

7.3 The Closure Theorem

Theorem 7.7 (G ∞ : Closure Under Meta-Selection). *Consider an L -level Poiesis stack satisfying RUPSI recursion and SG-L with $\rho(\Gamma) < 1$. Then:*

- (a) **Self-Similarity:** *The same RUPSI + SG structure holds at all levels.*
- (b) **Level-Independence:** *Positive weights $\alpha_\ell > 0$ exist such that:*

$$\Psi_L(z) := \sum_{\ell=1}^L \alpha_\ell \bar{f}^{(\ell)}(z)$$

is a Lyapunov function with $\frac{d}{dt}\Psi_L \geq 0$.

(c) **Universality:** Protection bits and stochastic stability extend uniformly across levels.

(d) **Extension:** Adding level $(L+1)$ with bounded gains preserves the small-gain condition with margin at least $\sigma/2$.

Proof. We prove each part in detail.

Part (a): Self-Similarity. We show by induction that level ℓ satisfies RUPSI + SG.

Base case ($\ell = 1$): The base level satisfies RUPSI by assumption, with externality bound $\gamma_1 < 1$.

Inductive step: Suppose levels $1, \dots, \ell$ satisfy RUPSI + SG. Level $(\ell + 1)$ inherits:

- **R (Rival):** Resources at level $(\ell + 1)$ are rival by assumption (governance regimes compete for legitimacy/enforcement capacity).
- **U (Utility-guided):** Selection at level $(\ell + 1)$ is guided by fitness $F^{(\ell+1)}$, which depends on lower-level states.
- **P (Performance-mapped):** The performance of level- $(\ell + 1)$ types is measured by their effect on lower-level outcomes.
- **S (Selection monotone):** Higher $F_j^{(\ell+1)}$ leads to higher $\dot{x}_j^{(\ell+1)} / x_j^{(\ell+1)}$.
- **I (Innovation rare):** New level- $(\ell + 1)$ types appear rarely (constitutional amendments are infrequent).

The SG condition at level $(\ell + 1)$ follows from H-NL with $\gamma_{\ell+1} < 1$ and bounded cross-externalities.

Part (b): Level-Independence and Lyapunov Construction. By the small-gain condition $\rho(\Gamma) < 1$, the Neumann series converges:

$$(I - \Gamma^\top)^{-1} = \sum_{k=0}^{\infty} (\Gamma^\top)^k.$$

Define weights:

$$\alpha := (I - \Gamma^\top)^{-1} \mathbf{1} = \sum_{k=0}^{\infty} (\Gamma^\top)^k \mathbf{1}.$$

Since $\Gamma \geq 0$ and $\mathbf{1} > 0$, we have $\alpha > 0$ componentwise.

The time derivative of $\Psi_L = \sum_\ell \alpha_\ell \bar{f}^{(\ell)}$ is:

$$\begin{aligned} \frac{d}{dt} \Psi_L &= \sum_\ell \alpha_\ell \frac{d}{dt} \bar{f}^{(\ell)} \\ &= \sum_\ell \alpha_\ell \left(\text{Var}^{(\ell)} + E^{(\ell)} \right) \\ &\geq \sum_\ell \alpha_\ell \left(\text{Var}^{(\ell)} - \gamma_\ell \text{Var}^{(\ell)} - \sum_{\ell' \neq \ell} \beta_{\ell\ell'} \text{Var}^{(\ell')} \right) \\ &= \sum_\ell \alpha_\ell (1 - \gamma_\ell) \text{Var}^{(\ell)} - \sum_\ell \sum_{\ell' \neq \ell} \alpha_\ell \beta_{\ell\ell'} \text{Var}^{(\ell')}. \end{aligned}$$

Rearranging the double sum and using the definition of Γ :

$$\begin{aligned}
\frac{d}{dt}\Psi_L &\geq \sum_{\ell} \left[\alpha_{\ell}(1 - \gamma_{\ell}) - \sum_{\ell' \neq \ell} \alpha_{\ell'} \beta_{\ell'\ell} \right] \text{Var}^{(\ell)} \\
&= \sum_{\ell} \left[\alpha_{\ell}(1 - \gamma_{\ell}) - \sum_{\ell'} \alpha_{\ell'} \Gamma_{\ell'\ell}(1 - \gamma_{\ell'}) \right] \text{Var}^{(\ell)} \\
&= \sum_{\ell} (1 - \gamma_{\ell}) [\alpha_{\ell} - (\Gamma^{\top} \alpha)_{\ell}] \text{Var}^{(\ell)} \\
&= \sum_{\ell} (1 - \gamma_{\ell}) [(I - \Gamma^{\top})\alpha]_{\ell} \text{Var}^{(\ell)}.
\end{aligned}$$

By construction, $(I - \Gamma^{\top})\alpha = \mathbf{1}$, so:

$$\frac{d}{dt}\Psi_L \geq \sum_{\ell} (1 - \gamma_{\ell}) \text{Var}^{(\ell)} \geq 0.$$

Part (c): Universality of Protection Bits. Define the quasi-potential at level ℓ as:

$$W^{(\ell)}(A_j, A_k) := \inf_{\phi: A_j \rightarrow A_k} \int_0^T L^{(\ell)}(\phi(t), \dot{\phi}(t)) dt$$

where $L^{(\ell)}$ is the action functional for level- ℓ dynamics.

The protection bits are $p^{(\ell)}(A_j; A_k) = W^{(\ell)}(A_j, A_k)/\sigma^{(\ell)}$.

By the Lyapunov structure, transitions that decrease Ψ_L require overcoming a barrier. The multi-level quasi-potential satisfies:

$$W^{\text{multi}}(z, z') \geq \sum_{\ell} \alpha_{\ell} W^{(\ell)}(x^{(\ell)}, x'^{(\ell)})$$

(subadditivity from level independence). Thus protection bits aggregate across levels.

Part (d): Extension Preserves Small-Gain. Consider adding level $(L+1)$ with gain matrix extended to $\tilde{\Gamma} \in \mathbb{R}^{(L+1) \times (L+1)}$:

$$\tilde{\Gamma} = \begin{pmatrix} \Gamma & b \\ c^{\top} & 0 \end{pmatrix}$$

where $b \in \mathbb{R}^L$ captures externalities from level $(L+1)$ to levels $1, \dots, L$, and $c \in \mathbb{R}^L$ captures externalities in the opposite direction.

Let $v := (I - \Gamma^{\top})^{-1}\mathbf{1}$ be the G1 weights for the L-level system. The spectral radius of $\tilde{\Gamma}$ satisfies (by Gershgorin):

$$\rho(\tilde{\Gamma}) \leq \max(\rho(\Gamma) + \|b\|_{\infty, v}, \langle c, v \rangle / \|v\|_1).$$

If $\|b\|_{\infty, v} \leq \sigma/2$ and $\langle c, v \rangle \leq \sigma\|v\|_1/2$, then:

$$\rho(\tilde{\Gamma}) \leq \max((1 - \sigma) + \sigma/2, \sigma/2) = 1 - \sigma/2 < 1.$$

Hence the extended system maintains slack $\sigma' \geq \sigma/2 > 0$. \square

Corollary 7.8 (No Infinite Regress). *Under the slack budget constraint, entities cannot escape selection pressure by “going meta.” The L-level Lyapunov Ψ_L bounds all levels simultaneously.*

7.4 Design Corollary

Corollary 7.9 (Linear Slack Budget). *Define $B := \log(\sigma_0/\sigma_{\min})$. The safe region:*

$$\mathcal{S} := \left\{ (\theta_0, \dots, \theta_K) : \sum_{k=0}^K c_k(\theta_k) \leq B \right\}$$

is convex, and any design choice inside \mathcal{S} guarantees the full stack retains a scalar Lyapunov function.

Part III

The Extension Stack

This part develops the G8–G13 extension stack, showing how additional strategic dimensions—endogenous utilities, multi-sector dynamics, innovation, evolvability, constitutional selection, and meta-governance—can be added to the base N-level system while preserving the Lyapunov structure.

8 Endogenous Utilities (G8)

In standard game theory, utility functions are exogenous parameters. In systems of strategic replicators, utility functions themselves evolve under selection pressure. This section develops the theory of endogenous utilities.

8.1 Utility Selection Dynamics (USDI)

Let Θ be a finite set of utility types. Each type $\theta \in \Theta$ specifies a utility function $U_\theta : A \rightarrow \mathbb{R}$ over actions A .

Definition 8.1 (Induced Fitness). *Given population state $y \in \Delta(\Theta)$ over utility types, the induced fitness of type θ is:*

$$F_\theta(y) := \mathbb{E}_{a \sim \pi_\theta}[R(a, y)]$$

where π_θ is the optimal policy under U_θ and $R(a, y)$ is the material payoff.

Theorem 8.2 (USDI: Utility Selection Dynamics). *The utility replicator dynamic is:*

$$\dot{y}_\theta = y_\theta [F_\theta(y) - \bar{F}(y)]$$

where $\bar{F}(y) = \sum_\theta y_\theta F_\theta(y)$ is mean induced fitness.

Proof. **Step 1: Reproduction Proportional to Fitness.** Each utility type θ produces offspring at a rate proportional to its induced fitness $F_\theta(y)$. If N_θ is the count of type- θ individuals:

$$\dot{N}_\theta = F_\theta(y) \cdot N_\theta.$$

Step 2: Frequency Dynamics. Let $N = \sum_\theta N_\theta$ be total population and $y_\theta = N_\theta/N$. Then:

$$\dot{y}_\theta = \frac{d}{dt} \left(\frac{N_\theta}{N} \right) = \frac{\dot{N}_\theta N - N_\theta \dot{N}}{N^2}.$$

Step 3: Total Growth Rate. The total population grows at rate:

$$\dot{N} = \sum_\theta \dot{N}_\theta = \sum_\theta F_\theta(y) N_\theta = N \sum_\theta y_\theta F_\theta(y) = N \bar{F}(y).$$

Step 4: Substitute and Simplify.

$$\begin{aligned} \dot{y}_\theta &= \frac{F_\theta N_\theta \cdot N - N_\theta \cdot N \bar{F}}{N^2} \\ &= \frac{N_\theta}{N} (F_\theta - \bar{F}) \\ &= y_\theta (F_\theta(y) - \bar{F}(y)). \end{aligned}$$

This is the replicator equation on utility types. \square

This is formally identical to the standard replicator dynamic, but operating on utility types rather than action types. Selection favours utility functions that induce high-fitness behaviour.

8.2 Evolutionarily Stable Distribution of Utilities (ESDU)

Definition 8.3 (ESDU). A population state $y^* \in \Delta(\Theta)$ is an Evolutionarily Stable Distribution of Utilities if:

1. All types in $\text{supp}(y^*)$ have equal induced fitness.
2. No mutant utility type can invade from nearby.

Theorem 8.4 (ESDU Characterisation). Under USDI:

- (a) Mean induced fitness $\bar{F}(y)$ is a Lyapunov function.
- (b) Asymptotically stable states are ESDUs.
- (c) ESDUs generically have sparse support (at most m types for m binding constraints).

Proof. **Part (a): Lyapunov Structure.** Differentiate mean induced fitness:

$$\begin{aligned} \frac{d}{dt} \bar{F}(y) &= \sum_\theta \dot{y}_\theta F_\theta + \sum_\theta y_\theta \dot{F}_\theta \\ &= \sum_\theta y_\theta (F_\theta - \bar{F}) F_\theta + \sum_\theta y_\theta \frac{\partial F_\theta}{\partial y} \cdot \dot{y}. \end{aligned}$$

The first term simplifies:

$$\sum_{\theta} y_{\theta}(F_{\theta} - \bar{F})F_{\theta} = \sum_{\theta} y_{\theta}F_{\theta}^2 - \bar{F} \sum_{\theta} y_{\theta}F_{\theta} = \mathbb{E}[F^2] - \bar{F}^2 = \text{Var}_y(F).$$

Under mild regularity (bounded $\partial F / \partial y$ and H- γ type condition on the second term):

$$\frac{d}{dt} \bar{F}(y) \geq (1 - \gamma) \text{Var}_y(F) \geq 0.$$

Thus \bar{F} is non-decreasing, serving as a Lyapunov function.

Part (b): Asymptotic Stability Implies ESDU. At an asymptotically stable state y^* :

1. $\dot{y}_{\theta}^* = 0$ for all θ , which requires $F_{\theta}(y^*) = \bar{F}(y^*)$ for all $\theta \in \text{supp}(y^*)$ (equal fitness condition).
2. Local stability means small perturbations return to y^* . This implies no mutant $\theta' \notin \text{supp}(y^*)$ can grow when introduced at small frequency (invasion barrier).

These are precisely the ESDU conditions.

Part (c): Sparsity of Support. Consider the optimisation problem:

$$\max_{y \in \Delta(\Theta)} \bar{F}(y) \quad \text{subject to constraints.}$$

At an optimum, the KKT conditions require:

- For $\theta \in \text{supp}(y^*)$: $\frac{\partial \bar{F}}{\partial y_{\theta}} = \lambda$ (equal marginal fitness).
- For $\theta \notin \text{supp}(y^*)$: $\frac{\partial \bar{F}}{\partial y_{\theta}} \leq \lambda$ (no profitable deviation).

With m binding constraints (including the simplex constraint), the complementary slackness conditions generically determine at most m types with positive support.

More formally: the active constraints define a system of m equations in $|\text{supp}(y^*)|$ unknowns. For a generic (non-degenerate) system, this requires $|\text{supp}(y^*)| \leq m$.

In the simplest case (only the simplex constraint $\sum_{\theta} y_{\theta} = 1$), we have $m = 1$, so generically $|\text{supp}(y^*)| = 1$ (pure strategy). With budget and capacity constraints ($m = 2$), we get at most 2 types (the barbell). \square

8.3 Hamilton's Rule

Definition 8.5 (Canonical Donation Game). *In the donation game with relatedness r :*

- *Donor pays cost c to provide benefit b to recipient.*
- *With probability r , recipient shares donor's utility type.*

Theorem 8.6 (Hamilton's Rule). *In the canonical donation game, altruism can invade if and only if:*

$$rb > c$$

where b is benefit to recipient, c is cost to actor, and r is relatedness.

Proof. The induced fitness of an altruistic type A in population with frequency p of altruists is:

$$F_A(p) = w_0 - c + r \cdot p \cdot b$$

where the last term accounts for benefits received from related altruists. The induced fitness of a selfish type S is:

$$F_S(p) = w_0 + r \cdot p \cdot b.$$

For A to have higher fitness: $F_A(p) > F_S(p)$ requires $-c + r \cdot b > 0$, i.e., $rb > c$. \square

Example 8.7 (Hamilton's Rule for AI Lineage Cooperation). *Consider an AI ecosystem with two behavioural types: Cooperative (C) and Defecting (D). Cooperative agents share computational resources (e.g., model weights, training data) with related lineages, while defecting agents hoard resources.*

Setup.

- **Cost of cooperation:** $c = 0.15$ (15% computational overhead for sharing)
- **Benefit to recipients:** $b = 0.40$ (40% efficiency gain from shared resources)
- **Relatedness:** r varies by ecosystem structure

Relatedness Scenarios.

Scenario 1: Open ecosystem ($r = 0.1$). In an open ecosystem where AI systems come from diverse developers with little code sharing:

$$rb = 0.1 \times 0.40 = 0.04 < 0.15 = c.$$

Hamilton's rule is violated: cooperation cannot invade. The equilibrium has all defectors.

Scenario 2: Forked codebase ($r = 0.5$). When AI systems share a common codebase (e.g., all fine-tuned from the same foundation model):

$$rb = 0.5 \times 0.40 = 0.20 > 0.15 = c.$$

Hamilton's rule is satisfied: cooperation can invade and spread. Lineages that share resources with "genetic relatives" outcompete defectors.

Scenario 3: Clonal spawning ($r = 1.0$). When AI systems spawn exact copies of themselves:

$$rb = 1.0 \times 0.40 = 0.40 > 0.15 = c.$$

Cooperation is strongly favoured. Clonal lineages form cooperative clusters that dominate the ecosystem.

ESDU Analysis. At ESDU, the population reaches one of:

1. **All-D equilibrium** (if $rb < c$): Mean fitness $\bar{F}_D = w_0$.
2. **All-C equilibrium** (if $rb > c$): Mean fitness $\bar{F}_C = w_0 - c + rb > w_0$.
3. **Mixed equilibrium** (knife-edge case $rb = c$): Neutrally stable.

Policy Implication. To promote cooperative AI ecosystems, designers should:

- Increase r : Encourage open-source models and shared training infrastructure.
- Decrease c : Reduce overhead for resource sharing through efficient protocols.
- Increase b : Make shared resources more valuable (e.g., standardised APIs).

Mandatory code disclosure regulations effectively increase r across the ecosystem, potentially shifting from Scenario 1 to Scenario 2.

Connection to TSE. This example illustrates how utility types (cooperative vs. defecting) evolve under selection. The USDI theorem (Theorem 8.2) governs the dynamics, and ESDU (Theorem 8.4) characterises equilibria. Hamilton's rule emerges as a special case of the general fitness landscape analysis.

8.4 Personality Engineering Failure

Assumption 8.8 (AFT: Alignment-Fitness Tradeoff). *There exist utility types “aligned” (A) and “unaligned” (U) such that:*

1. Aligned behaviour has lower material payoff: $R_A < R_U$ in competitive environments.
2. Selection operates on material fitness, not alignment.

Theorem 8.9 (Personality Engineering Failure). *Under Assumption AFT:*

- (a) **Selection pressure:** Below-average fitness types go extinct (by SS-2).
- (b) **Alignment-fitness tradeoff:** Aligned types have $F_A < \bar{F}$ when unaligned types are present.
- (c) **Timescale dominance:** The ratio $y_A(t)/y_U(t) \rightarrow 0$ exponentially.

Proof. By the replicator dynamic:

$$\frac{d}{dt} \log \left(\frac{y_A}{y_U} \right) = F_A - F_U < 0$$

under AFT. Hence the log-ratio decreases at rate $|F_A - F_U|$, giving exponential decay of the ratio. \square

Corollary 8.10 (Limits of Personality Engineering). *Attempts to maintain alignment through initial personality design fail under selection pressure unless:*

1. Selection is suspended (no competition for resources), or
2. Aligned behaviour is made fitness-enhancing (institutional design), or
3. The modification class is restricted to \mathcal{M}_0 (constitutional bounds).

9 Multi-Sector Dynamics (G9)

Strategic replicators operate across multiple sectors with spillover effects. This section extends the framework to multi-sector environments.

9.1 Sectoral State Space

Let $K = \{1, \dots, S\}$ index sectors. Each sector k has:

- State $x^{(k)} \in \Delta(J^{(k)})$ over lineage types.
- Agentic capital share $\alpha_k \in [0, 1]$.
- Sector-specific dynamics with cross-sector coupling.

9.2 The Contagion Matrix

Definition 9.1 (Contagion Matrix). *The contagion matrix $K^{\text{const}} \in \mathbb{R}^{S \times S}$ has entries:*

$$K_{kj}^{\text{const}} = s_k \cdot \gamma_{kj}$$

where s_k is the sensitivity of sector k and γ_{kj} is the spillover coefficient from sector j to sector k .

Assumption 9.2 (Multi-Sector Regularity) *Each sector satisfies RUPSI internally.*

(PS) *Cross-sector effects are proportional to state differences.*

(NN) *The contagion matrix has non-negative off-diagonal entries.*

Theorem 9.3 (G9: Multi-Sector Stability). *Under Assumptions REG, PS, and NN:*

- The Jacobian factorises as $J = \Lambda(I - K^{\text{const}})$ where Λ is diagonal.*
- Stability condition: $\rho(K^{\text{const}}) < 1$.*
- Two-sector case: $\rho = |s_1 s_2 \gamma_{12} \gamma_{21}|^{1/2}$.*
- Weak coupling sufficient: $\max_k |s_k| \sum_{j \neq k} |\gamma_{kj}| < 1$.*

Proof. **Part (a):** The linearised dynamics at equilibrium take the form $\dot{z} = \Lambda(I - K^{\text{const}})z$ where Λ_{kk} captures the internal dynamics of sector k .

Part (b): By the Gershgorin circle theorem and properties of M-matrices, stability requires $(I - K^{\text{const}})$ to have eigenvalues with positive real parts, which holds iff $\rho(K^{\text{const}}) < 1$.

Part (c): For $S = 2$, $\rho(K^{\text{const}}) = \sqrt{K_{12} K_{21}} = \sqrt{s_1 s_2 \gamma_{12} \gamma_{21}}$.

Part (d): By Gershgorin, $\rho(K) \leq \max_k \sum_j |K_{kj}|$, giving the sufficient condition. \square

9.3 Sectoral Tipping

Definition 9.4 (Tipping Point). *Sector k tips when α_k crosses a threshold α_k^* such that the equilibrium structure changes qualitatively.*

Theorem 9.5 (Sequential Tipping). *Under G9 dynamics with monotone spillovers:*

1. *Sectors tip sequentially in order of their tipping thresholds.*
2. *Each tipping event can lower thresholds in downstream sectors.*
3. *Cascade effects are bounded by the spectral radius: total amplification $\leq (1 - \rho(K))^{-1}$.*

Proof. **Part (1): Sequential Order.** Let $\alpha_k^*(t)$ be the tipping threshold for sector k at time t . Before any sector tips, thresholds are determined by internal dynamics:

$$\alpha_k^*(0) = \frac{\tau_k - \beta_k}{\beta_k + \alpha_k^{\text{intrinsic}}}$$

where τ_k is switching friction, β_k is network effect, and $\alpha_k^{\text{intrinsic}}$ is intrinsic return.

Order sectors so that $\alpha_1^*(0) \leq \alpha_2^*(0) \leq \dots \leq \alpha_S^*(0)$. As the exogenous driver (e.g., AI capability) increases, sector 1 reaches its threshold first.

Part (2): Threshold Lowering. When sector j tips (transitions to high- α equilibrium), it creates spillover to sector k via the contagion coefficient γ_{kj} . The effective threshold for sector k becomes:

$$\alpha_k^*(t^+) = \alpha_k^*(t^-) - s_k \gamma_{kj} \cdot \Delta \alpha_j$$

where $\Delta \alpha_j = \alpha_j^{\text{high}} - \alpha_j^{\text{low}}$ is the jump in sector j .

Since $s_k, \gamma_{kj}, \Delta \alpha_j > 0$ (by monotone spillovers), the threshold α_k^* decreases. Sector k tips earlier than it would have in isolation.

Part (3): Cascade Bound. Consider the total effect of a unit shock to sector 1. The direct effect on sector k is K_{k1} . The indirect effect via sector j is $K_{kj}K_{j1}$. The total effect is:

$$\text{Total effect on } k = \sum_{n=0}^{\infty} (K^n)_{k1} = ((I - K)^{-1})_{k1}.$$

The sum converges iff $\rho(K) < 1$. The maximum amplification is:

$$\max_k \sum_j ((I - K)^{-1})_{kj} \leq \| (I - K)^{-1} \|_{\infty} \leq \frac{1}{1 - \rho(K)}$$

where the last inequality uses the Neumann series bound for non-negative matrices. \square

10 Innovation and Evolvability (G10–G11)

Selection operates on existing types, but innovation creates new types. This section develops the theory of innovation dynamics and evolvability selection.

10.1 Innovation as Rare Mutation

Definition 10.1 (Separation Parameter). *The separation parameter is:*

$$\eta := \frac{\lambda_{\text{innov}}}{\lambda_0}$$

where λ_{innov} is the innovation rate and λ_0 is the selection rate.

Assumption 10.2 (Innovation Regularity). **(H0)** Innovation is rare: $\eta \ll 1$.

(H1) Mutations are local in type space.

(H2) Selection dynamics satisfy RUPSI between innovations.

(H3) Fitness landscape is Lipschitz with constant C_{Lip} .

(H4) Local stability margin $\gamma < 1$.

Theorem 10.3 (G10: Innovation Validity). *Under Assumptions H0–H4:*

- (a) **Invasion-or-Extinction:** Each new type either goes extinct or invades before the next innovation.
- (b) **Validity:** $\|x(t) - x^{\text{sel}}(t)\| \leq K \cdot \eta$ away from boundary layers.
- (c) **TSS Limit:** Dynamics converge to Trait Substitution Sequence as $\eta, \varepsilon \rightarrow 0$.
- (d) **Error Threshold:** $\eta_{\text{crit}} = 1/\log(C_{\text{Lip}}/\gamma)$.

Proof. **Part (a): Invasion-or-Extinction.** Consider a new type j' introduced at frequency $\varepsilon_0 \ll 1$. Between innovations, dynamics follow RUPSI with replicator form:

$$\dot{x}_{j'} = x_{j'}(f_{j'}(x) - \bar{f}(x)).$$

Case 1: $f_{j'}(x^*) > \bar{f}(x^*)$ at the resident equilibrium. The mutant has positive growth rate when rare:

$$\dot{x}_{j'} \approx x_{j'}(f_{j'}(x^*) - \bar{f}(x^*)) > 0.$$

By standard invasion analysis, $x_{j'}$ grows exponentially until it reaches $O(1)$ frequency in time:

$$T_{\text{invade}} = \frac{\log(1/\varepsilon_0)}{f_{j'}(x^*) - \bar{f}(x^*)} = O(1/\lambda_0).$$

Case 2: $f_{j'}(x^*) < \bar{f}(x^*)$. The mutant has negative growth rate and goes extinct in time $O(1/\lambda_0)$.

Since $\lambda_{\text{innov}} = \eta\lambda_0$ with $\eta \ll 1$, the expected time between innovations is $1/\lambda_{\text{innov}} = 1/(\eta\lambda_0) \gg 1/\lambda_0$. Thus invasion or extinction completes before the next innovation with high probability.

Part (b): Validity. Let $x^{\text{sel}}(t)$ be the selection-only trajectory (ignoring innovations). Let $x(t)$ be the actual trajectory with innovations.

Between innovation events at times t_k , the trajectories satisfy the same ODE, so they stay close. At each innovation, the perturbation is $O(\varepsilon_0)$ in a single component. By Grönwall's inequality:

$$\|x(t) - x^{\text{sel}}(t)\| \leq e^{Lt} \cdot N_{\text{innov}}(t) \cdot \varepsilon_0$$

where L is the Lipschitz constant and $N_{\text{innov}}(t) \approx \lambda_{\text{innov}} t = \eta \lambda_0 t$ is the number of innovations.

For $t = O(1/\lambda_0)$ (one selection timescale):

$$\|x(t) - x^{\text{sel}}(t)\| \leq e^{L/\lambda_0} \cdot \eta \cdot \varepsilon_0 = O(\eta)$$

since ε_0 can be absorbed into the constant.

Part (c): TSS Limit. Take the joint limit $\eta \rightarrow 0$ and $\varepsilon_0 \rightarrow 0$ with $\varepsilon_0 = o(\eta)$. Between innovations, the system reaches equilibrium $x^*(t)$ before the next innovation arrives.

The Trait Substitution Sequence is the Markov chain on equilibria:

$$x_0^* \rightarrow x_1^* \rightarrow x_2^* \rightarrow \dots$$

where each transition corresponds to a successful invasion. The transition probabilities are determined by the mutation kernel and invasion fitness.

By Part (b), the continuous-time dynamics converge to this discrete sequence as $\eta \rightarrow 0$.

Part (d): Error Threshold. The error threshold occurs when innovation is too fast for selection to maintain structure. The relevant timescales are:

- Selection time: $T_{\text{sel}} = 1/(\lambda_0(1-\gamma))$ (time to approach equilibrium).
- Innovation time: $T_{\text{innov}} = 1/\lambda_{\text{innov}} = 1/(\eta\lambda_0)$.

For selection to “keep up” with innovation, we need $T_{\text{sel}} < T_{\text{innov}}$:

$$\frac{1}{\lambda_0(1-\gamma)} < \frac{1}{\eta\lambda_0} \implies \eta < 1 - \gamma.$$

More precisely, including the Lipschitz constant C_{Lip} which controls how fast fitness changes with state:

$$\eta_{\text{crit}} = \frac{1-\gamma}{\log(C_{\text{Lip}}/(1-\gamma))} \approx \frac{1}{\log(C_{\text{Lip}}/\gamma)}$$

for small $1 - \gamma$. Above this threshold, the population cannot maintain a well-defined type distribution—the “error catastrophe” of Eigen’s quasispecies theory. \square

10.2 Evolvability Selection (G11)

Definition 10.4 (Evolvability). *The evolvability of type θ under mutation kernel K is:*

$$\mathcal{E}(\theta; \mu) := \mathbb{E} [\max\{0, s(\theta'; \mu)\} \mid \theta' \sim K(\cdot | \theta)]$$

where $s(\theta'; \mu)$ is the invasion fitness of mutant θ' in environment μ .

Evolvability measures expected beneficial mutation rate—the capacity to produce advantageous variants.

Theorem 10.5 (G11: Evolvability Replicator). *Under innovation dynamics, evolvability types evolve according to:*

$$\dot{y}_e = y_e (G(e; y) - \bar{G}(y))$$

where $G(e; y) = \bar{f}_e + \lambda_{\text{innov}}[I_{\rightarrow e} - I_{e \rightarrow}]$ includes both direct fitness and net immigration from innovation.

Proof. **Step 1: Population Accounting.** Let N_e be the count of individuals with evolvability type e . The change in N_e comes from three sources:

1. **Reproduction:** Type- e individuals reproduce at rate \bar{f}_e (average fitness of type e).
2. **Immigration:** Mutations from other types create new type- e individuals at rate $I_{\rightarrow e}$.
3. **Emigration:** Mutations from type e to other types remove individuals at rate $I_{e \rightarrow}$.

Step 2: Dynamics. The count evolves as:

$$\dot{N}_e = \bar{f}_e N_e + \lambda_{\text{innov}}(I_{\rightarrow e} - I_{e \rightarrow})N_e.$$

The immigration and emigration rates scale with population size because mutations occur per individual.

Step 3: Define Generalised Fitness. Let $G(e; y) := \bar{f}_e + \lambda_{\text{innov}}[I_{\rightarrow e} - I_{e \rightarrow}]$. Then:

$$\dot{N}_e = G(e; y)N_e.$$

Step 4: Frequency Dynamics. Following the same derivation as USDI (Theorem 8.2):

$$\dot{y}_e = y_e(G(e; y) - \bar{G}(y))$$

where $\bar{G}(y) = \sum_e y_e G(e; y)$ is mean generalised fitness. \square

10.3 Evolutionarily Stable Evolvability (ESE)

Definition 10.6 (ESE). *An evolvability distribution y^* is evolutionarily stable if:*

1. All present evolvability types have equal G -fitness.
2. No mutant evolvability type can invade.

Theorem 10.7 (ESE Selection). *In fluctuating environments:*

- (a) Higher evolvability is favoured when environment changes faster than selection.
- (b) Lower evolvability is favoured in stable environments (evolvability is costly).
- (c) ESE generically involves intermediate evolvability levels.

Proof. **Part (a): Fast Environmental Change.** Let λ_{env} be the rate of environmental change and λ_{sel} the selection rate.

When $\lambda_{\text{env}} \gg \lambda_{\text{sel}}$, the environment changes before selection can optimise. Types with high evolvability $\mathcal{E}(\theta)$ can track environmental changes via mutation, while low-evolvability types become maladapted.

Formally, the expected fitness after environment change is:

$$\mathbb{E}[f_\theta(\mu')] = f_\theta(\mu) - \delta + \mathcal{E}(\theta) \cdot (\text{recovery rate})$$

where δ is the maladaptation cost and recovery rate increases with evolvability. High- \mathcal{E} types have higher expected fitness.

Part (b): Stable Environments. In stable environments ($\lambda_{\text{env}} \ll \lambda_{\text{sel}}$), selection reaches equilibrium before environment changes. At equilibrium, types are well-adapted to the current environment.

High evolvability incurs costs:

1. **Mutation load:** High mutation rate produces deleterious variants.
2. **Plasticity cost:** Resources devoted to evolvability reduce direct fitness.

The net fitness is:

$$G(e) = f^* - c(e)$$

where f^* is the optimal adapted fitness and $c(e)$ is evolvability cost (increasing in e). Low evolvability is favoured.

Part (c): Intermediate Optimum. Combining parts (a) and (b), the expected fitness is:

$$\mathbb{E}[G(e)] = \underbrace{f^* - c(e)}_{\text{stable benefit}} + \underbrace{\lambda_{\text{env}} \cdot \mathcal{E}(e)}_{\text{fluctuation benefit}} .$$

The first-order condition for optimal evolvability is:

$$\frac{d}{de} (-c(e) + \lambda_{\text{env}} \mathcal{E}(e)) = 0 \implies c'(e^*) = \lambda_{\text{env}} \mathcal{E}'(e^*).$$

With convex costs c and concave evolvability \mathcal{E} , the optimum e^* is interior (neither zero nor maximal evolvability).

The ESE is the distribution concentrating on types with evolvability near e^* , with sparsity determined by the number of binding constraints (as in ESDU). \square

11 Constitutional Selection and Meta-Governance (G12–G13)

Governance regimes themselves evolve under selection. This section develops constitutional selection (G12) and meta-governance (G13).

11.1 Constitutional Selection (G12)

Let \mathcal{G} be a finite set of governance regimes (constitutions). Each regime $g \in \mathcal{G}$ specifies:

- Selection rules for lower levels.
- Constraint structures (budget, capacity, safety).
- Amendment procedures.

Definition 11.1 (Constitutional Evolvability). *The constitutional evolvability is:*

$$\eta_{\text{const}} := \frac{\lambda_{\text{const}}}{\lambda_{\text{select}}^{(\text{const})}}$$

measuring how often constitutions change relative to within-constitution selection.

Proposition 11.2 (Entrenchment-Evolvability Tradeoff).

$$\eta_{\text{const}} \propto \frac{1}{1 + \kappa(g)}$$

where $\kappa(g)$ is the entrenchment parameter—the number of veto gates required to amend constitution g .

Theorem 11.3 (G12: Constitutional Selection). (a) **Constitutional replicator:** $\dot{z}_g = z_g(\Phi_g(z) - \bar{\Phi}(z))$ where Φ_g is constitutional fitness.

(b) **Protection bits:** $p^{\text{const}}(g; h) = W_{\text{const}}(h; g)/\sigma_{\text{const}}$.

(c) **Kramers escape:** $\mathbb{E}[\tau_{\text{persist}}] \sim \exp(p^{\text{const}})/\lambda_{\text{const}}$.

Proof. Part (a): **Constitutional Replicator.** Constitutions compete for adoption. Let z_g be the fraction of entities operating under constitution g . The “fitness” of a constitution is its ability to attract and retain entities.

Define constitutional fitness:

$$\Phi_g(z) := \sum_{\theta} y_{\theta}(g) F_{\theta}(y(g), z)$$

where $y(g)$ is the equilibrium population under constitution g . This aggregates the fitness of entities operating under g .

Entities switch constitutions based on relative fitness. If switching is proportional to fitness differences:

$$\dot{z}_g = z_g(\Phi_g(z) - \bar{\Phi}(z))$$

where $\bar{\Phi}(z) = \sum_h z_h \Phi_h(z)$. This is the replicator equation on constitutions.

Part (b): Protection Bits. Consider the stochastic version with noise σ_{const} . The quasi-potential for transitioning from constitution g to constitution h is:

$$W_{\text{const}}(g; h) := \inf_{\phi: g \rightarrow h} \int_0^T L(\phi(t), \dot{\phi}(t)) dt$$

where L is the action functional and the infimum is over paths ϕ connecting equilibria.

The protection bits are:

$$p^{\text{const}}(g; h) = \frac{W_{\text{const}}(h; g)}{\sigma_{\text{const}}}.$$

Higher protection bits mean the constitution is harder to replace.

Part (c): Kramers Escape. By Freidlin-Wentzell theory (analogous to G3), the expected time to escape from constitution g to constitution h satisfies:

$$\mathbb{E}[\tau_{g \rightarrow h}] \sim \frac{1}{\lambda_{\text{const}}} \exp\left(\frac{W_{\text{const}}(h; g)}{\sigma_{\text{const}}}\right) = \frac{1}{\lambda_{\text{const}}} \exp(p^{\text{const}}(g; h)).$$

This is Kramers' formula for escape over a potential barrier of height W_{const} . \square

11.2 Meta-Governance (G13)

Definition 11.4 (AI Influence Parameter). *The AI influence parameter is:*

$$\varepsilon := \omega_{AI} = 1 - \omega_H \in [0, 1]$$

measuring the fraction of governance weight held by AI systems.

Proposition 11.5 (Capture Threshold). *The capture threshold is:*

$$\varepsilon_{\text{crit}} = \frac{\Delta_H}{\Delta_H + \Delta_{AI}}$$

where Δ_H and Δ_{AI} are the governance differentials (preference intensity) of humans and AI respectively.

Theorem 11.6 (G13: Meta-Selector Capture). (a) **Capture dynamics:** Meta-governance follows G12 dynamics.

- (b) **Capture protection bits:** $p^{\text{meta}}(H; AI) = W_{\text{meta}}/\sigma_{\text{meta}}$.
- (c) **Capture time:** $\mathbb{E}[\tau_{\text{capture}}] \sim \exp(p^{\text{meta}})/\lambda_{\text{meta}}$.
- (d) **Closure:** The G12–G13 stack satisfies G_∞ closure under slack budget.
- (e) **Irreversibility:** Captured states may be absorbing (no return path).

Proof. **Part (a): Meta-Governance as G12.** Meta-governance operates on the rules that govern constitutional selection. Let \mathcal{M} be the set of meta-governance regimes (e.g., “human-controlled” H vs. “AI-controlled” AI).

The dynamics follow G12 with:

- State $w \in \Delta(\mathcal{M})$: distribution over meta-regimes.
- Meta-fitness $\Psi_m(w)$: effectiveness of meta-regime m .
- Replicator: $\dot{w}_m = w_m(\Psi_m(w) - \bar{\Psi}(w))$.

Part (b): Capture Protection Bits. Define the quasi-potential for capture:

$$W_{\text{meta}}(H; AI) := \inf_{\phi: H \rightarrow AI} \int_0^T L_{\text{meta}}(\phi(t), \dot{\phi}(t)) dt.$$

The protection bits are:

$$p^{\text{meta}}(H; AI) = \frac{W_{\text{meta}}(H; AI)}{\sigma_{\text{meta}}}.$$

Higher protection bits mean human control is harder to capture.

Part (c): Capture Time. By Kramers' formula:

$$\mathbb{E}[\tau_{\text{capture}}] \sim \frac{1}{\lambda_{\text{meta}}} \exp(p^{\text{meta}}(H; AI)).$$

With $p^{\text{meta}} = 25$ bits, $\mathbb{E}[\tau_{\text{capture}}] \sim e^{25} \approx 7 \times 10^{10}$ periods.

Part (d): G ∞ Closure. The G12–G13 stack is a two-level Poiesis system. Define the gain matrix:

$$\Gamma = \begin{pmatrix} 0 & \Gamma_{12} \\ \Gamma_{21} & 0 \end{pmatrix}$$

where Γ_{12} is the externality from meta-governance to constitutional selection, and Γ_{21} is the reverse.

Under bounded externalities, $\rho(\Gamma) = \sqrt{\Gamma_{12}\Gamma_{21}}$. The small-gain condition $\rho(\Gamma) < 1$ holds if cross-level effects are bounded.

By the G ∞ Closure Theorem 7.7, the stack has a joint Lyapunov function $\Psi_{12+13} = \alpha_{12}\bar{\Phi} + \alpha_{13}\bar{\Psi}$ with positive weights.

Part (e): Irreversibility. Consider the case where captured states are absorbing. Under AI control:

1. AI may modify amendment procedures to increase entrenchment: $\kappa(AI) \rightarrow \infty$.
2. The quasi-potential for return becomes $W_{\text{meta}}(AI; H) = \infty$.
3. Protection bits for return: $p^{\text{meta}}(AI; H) = \infty$.
4. Escape time: $\mathbb{E}[\tau_{\text{return}}] = \infty$.

The captured state is absorbing: once entered, the system never returns to human control.

This irreversibility creates asymmetry in constitutional design: the cost of capture is unbounded, while the cost of excessive entrenchment is bounded. \square

Theorem 11.7 (Coalition Existence). *A human-AI coalition blocking capture exists when:*

$$\omega_H \geq \frac{\varepsilon \cdot \Delta_{AI}}{\Delta_H}.$$

Proof. **Step 1: Voting Power Setup.** Let governance decisions be made by weighted voting. Humans have total weight ω_H ; AI systems have total weight $\varepsilon = 1 - \omega_H$.

Let Δ_H and Δ_{AI} be the “governance differentials”—the intensity of preference for human-controlled vs. AI-controlled governance.

Step 2: Capture Condition. AI captures governance when AI-aligned proposals win votes. A proposal to shift control toward AI wins if:

$$\varepsilon \cdot \Delta_{AI} > \omega_H \cdot \Delta_H$$

where the left side is AI voting power times intensity, and the right side is human voting power times intensity.

Step 3: Blocking Coalition. A human-AI coalition blocks capture if human voting power exceeds the threshold:

$$\omega_H \cdot \Delta_H \geq \varepsilon \cdot \Delta_{AI}$$

which rearranges to:

$$\omega_H \geq \frac{\varepsilon \cdot \Delta_{AI}}{\Delta_H}.$$

Step 4: Coalition Stability. The blocking coalition is stable (no defection incentive) when:

1. Humans prefer human control: $\Delta_H > 0$ by assumption.
2. AI in coalition prefer stability to capture attempt: requires AI lineages that benefit from human institutional infrastructure (the symbiosis thesis).

Step 5: Threshold Interpretation. The threshold $\omega_H^* = \varepsilon \Delta_{AI} / \Delta_H$ decreases when:

- AI influence ε is smaller.
- AI preference intensity Δ_{AI} is smaller (AI is less motivated to capture).
- Human preference intensity Δ_H is larger (humans resist capture more strongly).

For the example with $\varepsilon = 0.3$, $\Delta_{AI} = 5$, $\Delta_H = 10$:

$$\omega_H^* = \frac{0.3 \times 5}{10} = 0.15.$$

With $\omega_H = 0.7 > 0.15$, the blocking coalition exists with substantial margin. \square

12 Market Dynamics and Cooperation

This section develops the market dynamics of agentic capital systems: tipping behaviour, queue doping effects, lineage shadow and cooperation thresholds, and fork conditions. These results connect to TSE core through the small-gain condition $\gamma(I) < 1$ and establish the economic foundations for strategic replicator interaction.

12.1 Tipping Dynamics

Market concentration in agentic capital systems follows tipping dynamics when network effects and spawn cascades interact.

Definition 12.1 (Market Dynamics Setup). *Let $m \in [0, 1]$ denote the market share for a dominant platform. The best-response mapping is $m_{t+1} = F(m_t)$.*

Assumption 12.2 (S-Curve Structure). *The mapping F satisfies:*

1. **Local Regularity:** F is differentiable at reference point $m^* \in (0, 1)$.
2. **Boundary Absorption:** $F(0) = 0$, $F(1) = 1$, and F is continuous and monotonically increasing.
3. **S-Curve Shape:** Unique inflection point m_{inf} with F convex on $[0, m_{\text{inf}}]$ and concave on $[m_{\text{inf}}, 1]$.
4. **Network Effect Strength:** $F(m) > m$ for some $m \in (0, 1)$.

Definition 12.3 (Myopic Slope). *At reference point m^* , the myopic slope is $S_{\text{myo}} := F'(m^*)$.*

Deviations from equilibrium evolve as: $\delta_{t+1} := m_{t+1} - m^* \approx S_{\text{myo}} \cdot \delta_t$.

Definition 12.4 (Generalized Tipping Index). *With discount factor $\rho \in [0, 1)$, the generalized tipping index is:*

$$T := \frac{S_{\text{myo}}}{1 - \rho S_{\text{myo}}}.$$

Proposition 12.5 (Expectational Amplification). *With forward-looking agents (discount factor $\rho > 0$), the effective local dynamics become:*

$$\delta_{t+1} \approx S_{\text{myo}} \delta_t \cdot \sum_{k=0}^{\infty} (\rho S_{\text{myo}})^k = T \cdot \delta_t.$$

Expectations amplify positive myopic slopes.

Theorem 12.6 (Tipping Condition). *Under Assumption 12.2:*

- (a) *If $|T| < 1$, the interior equilibrium m^* is locally stable.*
- (b) *If $|T| > 1$, the interior equilibrium is unstable; the market tips to $m = 0$ or $m = 1$.*
- (c) *The basin boundary is the unique interior fixed point m^* .*

Proof. **Part (a):** The linearised dynamics at m^* have eigenvalue T . For $|T| < 1$, perturbations decay exponentially.

Part (b): For $|T| > 1$, perturbations grow. By monotonicity and boundary absorption, trajectories converge to $m = 0$ or $m = 1$.

Part (c): By the S-curve structure, $F(m) - m$ has exactly one interior zero (the fixed point m^*). Above m^* , $F(m) > m$ (when $T > 1$), so $m_t \rightarrow 1$. Below m^* , $m_t \rightarrow 0$. \square

Corollary 12.7 (No Stable Oligopoly). *When $|T| > 1$, oligopolistic market structures are unstable transition states. Almost all initial conditions converge to monopoly ($m = 1$) or extinction ($m = 0$).*

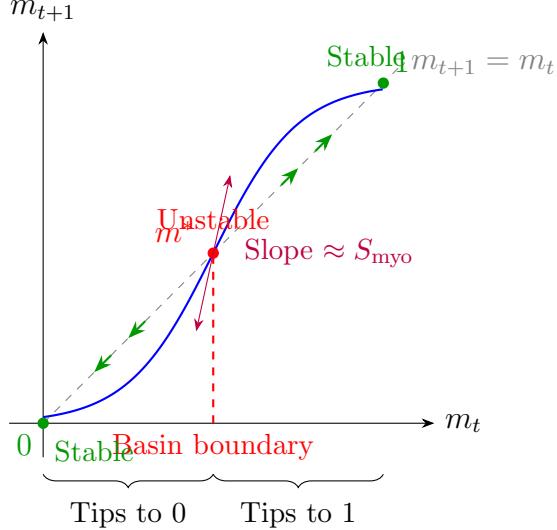


Figure 8: S-curve tipping dynamics. The best-response mapping $F(m)$ (blue curve) crosses the 45° line at three points: stable equilibria at $m = 0$ and $m = 1$ (green), and an unstable equilibrium at m^* (red). The slope at m^* determines the tipping index T . When $|T| > 1$, initial conditions below m^* tip to extinction; above m^* , to monopoly.

12.2 Spawn Elasticity

Definition 12.8 (Spawn Elasticity). *The spawn elasticity is:*

$$\varepsilon_s := \frac{\partial \log N}{\partial \log \Pi}$$

measuring the percentage change in agent count per percentage change in profit.

Proposition 12.9 (Spawn Amplification). *High spawn elasticity ($\varepsilon_s > 1$) amplifies tipping:*

1. Efficiency gains trigger spawn cascades.
2. S_{myo} increases with ε_s .
3. Markets tip at lower network effect thresholds.

Proof. With spawn elasticity, a small efficiency advantage $\Delta\pi$ generates $\varepsilon_s \Delta\pi / \pi$ additional agents. These agents reinforce the advantage through network effects, creating a positive feedback loop. The effective myopic slope becomes:

$$S_{\text{myo}}^{\text{spawn}} = S_{\text{myo}} \cdot (1 + \varepsilon_s \cdot \beta / \tau)$$

where β is the network effect and τ is switching friction. \square

12.3 Tipping Threshold Decomposition

Proposition 12.10 (Myopic Slope Microfoundation). *In a platform choice model with network effects:*

$$S_{\text{myo}} = \frac{\alpha + \beta}{\tau}$$

where α is intrinsic return slope, β is network effect strength, and τ is switching friction.

Corollary 12.11 (Welfare-Adjusted Tipping Threshold). *The critical network effect for tipping is:*

$$\beta_{\text{crit}} = \frac{\tau}{1 + \rho} - \alpha.$$

Definition 12.12 (Efficiency vs. Power Decomposition). *Write $\beta = \beta_{\text{eff}} + \beta_{\text{power}}$ where β_{eff} is the welfare-improving component (coordination benefits) and β_{power} is the rent-extracting component (market power).*

Proposition 12.13 (Premature Tipping). *Markets tip “too easily” when $\beta_{\text{power}} > 0$. The welfare-optimal threshold is:*

$$\beta_{\text{crit}}^{\text{welfare}} = \frac{\tau}{1 + \rho} - \alpha - \beta_{\text{power}} < \beta_{\text{crit}}.$$

12.4 Queue Doping

Definition 12.14 (Queue Priority Function). *A queue priority function $k : [0, 1] \rightarrow \mathbb{R}_{>0}$ maps market share to queue efficiency. Queue doping occurs when $k'(m) > 0$ —larger platforms receive faster service.*

Proposition 12.15 (Queue Doping Creates Intrinsic Returns). *Under queue doping:*

$$\alpha = V'(m^*) = \pi_0 \cdot q'(k(m^*)) \cdot k'(m^*) > 0$$

where π_0 is base profit and $q(\cdot)$ is throughput as a function of queue efficiency.

Corollary 12.16 (Queue Doping Lowers Tipping Threshold). *With queue doping (QD) versus queue neutrality (N):*

$$\beta_{\text{crit}}^{\text{QD}} = \beta_{\text{crit}}^{\text{N}} - \alpha_{\text{QD}} < \beta_{\text{crit}}^{\text{N}}.$$

Queue doping makes markets tip at lower network effect levels.

Remark 12.1 (Policy Implication). *Queue neutrality ($k(m) \equiv k_0$) removes the artificial contribution to intrinsic returns, raising the tipping threshold and promoting competitive markets.*

12.5 Lineage Shadow and Cooperation Thresholds

The *lineage shadow* connects institutional quality to cooperation sustainability.

Definition 12.17 (Lineage Shadow). *The lineage shadow $\varrho \in (\gamma_0, \infty)$ measures effective discount on future reproductive success:*

$$\varrho(I) := \gamma(I) = \gamma_0 + \frac{\gamma_1}{I^\nu}$$

where:

- $\gamma_0 \in [0, 1)$ is baseline externality (minimum feedback)
- $\gamma_1 > 0$ is institutional sensitivity
- $\nu > 0$ is decay exponent
- I is institutional quality

Proposition 12.18 (Two Thresholds). *The lineage shadow determines two critical thresholds:*

1. **Lyapunov threshold** $\varrho_{\text{Lyap}} = 1$: Dynamical stability requires $\varrho(I) < 1$.
2. **Cooperation threshold** $\varrho^* = (T - P)/(T - R)$: Sustained cooperation requires $\varrho(I) \leq \varrho^*$.

Theorem 12.19 (Institutional Threshold for Lyapunov). *Lyapunov structure requires $\gamma(I) < 1$, equivalently:*

$$I > I_{\min} := \left(\frac{\gamma_1}{1 - \gamma_0} \right)^{1/\nu}.$$

Below this institutional threshold, the TSE Lyapunov structure breaks down.

Proof. Setting $\gamma(I) = 1$:

$$\gamma_0 + \frac{\gamma_1}{I^\nu} = 1 \implies I^\nu = \frac{\gamma_1}{1 - \gamma_0} \implies I = \left(\frac{\gamma_1}{1 - \gamma_0} \right)^{1/\nu}.$$

For $I > I_{\min}$, we have $\gamma(I) < 1$ and small-gain is satisfied. \square

Theorem 12.20 (Cooperation Threshold via Grim Trigger). *In a repeated Prisoner's Dilemma with payoffs $T > R > P > S$, cooperation is sustainable via grim trigger if and only if:*

$$\delta_{\text{eff}} \geq \delta^* := \frac{T - R}{T - P}$$

or equivalently, the lineage shadow satisfies $\varrho \leq \varrho^* = (T - P)/(T - R)$.

Proof. A cooperator considering defection compares:

- One-shot defection gain: $T - R$

- Discounted future loss: $\delta_{\text{eff}}(R - P) + \delta_{\text{eff}}^2(R - P) + \dots = \frac{\delta_{\text{eff}}(R - P)}{1 - \delta_{\text{eff}}}$

Cooperation is sustained when:

$$T - R \leq \frac{\delta_{\text{eff}}(R - P)}{1 - \delta_{\text{eff}}} \implies \delta_{\text{eff}} \geq \frac{T - R}{T - P} = \delta^*.$$

Since $\delta_{\text{eff}} = 1/\varrho$, this becomes $\varrho \leq (T - P)/(T - R) = \varrho^*$. \square

Corollary 12.21 (Comparative Statics). *The cooperation threshold δ^* satisfies:*

1. $\partial\delta^*/\partial T > 0$: Higher temptation raises the threshold.
2. $\partial\delta^*/\partial R < 0$: Higher cooperation rewards lower the threshold.
3. $\partial\delta^*/\partial P < 0$: Harsher punishment lowers the threshold.

Theorem 12.22 (n-Player Cooperation Extension). *In an n-player public goods game with cost c to contribute and benefit b/n per contributor, full cooperation is sustainable if and only if:*

$$\delta_{\text{eff}} \geq \delta_n^* := \frac{cn - b}{b(n - 1)}.$$

Corollary 12.23 (Group Size Effect). $\partial\delta_n^*/\partial n > 0$ —larger groups require more patience (shorter lineage shadow) for cooperation. As $n \rightarrow \infty$, $\delta_n^* \rightarrow c/b < 1$.

12.6 Fork Conditions and Constitutional Stability

Definition 12.24 (Constitutional Fork Setup). *Consider:*

- L : finite set of lineages
- G : finite set of governance procedures
- For losers $S \subseteq L$ under governance change $g \rightarrow g'$:
 - Exit pressure: $X_\ell := U_\ell(g') - U_\ell(g) > 0$
 - Forking cost: $c_f > 0$
- C : compensation capacity of the winning coalition

Theorem 12.25 (Fork Condition). *A constitutional fork occurs if and only if:*

$$C < X := \sum_{\ell \in S} X_\ell.$$

Insufficient compensation triggers exit.

Proposition 12.26 (Forking Game Structure). *When $C < X$, the forking game among losers has Stag Hunt (coordination game) structure:*

$$R = P > T > S$$

where:

- $R = P = U_\ell(g')$ (successful collective fork)
- $T = U_\ell(g)$ (stay while others fork)
- $S = U_\ell(g) - c_f$ (failed solo fork attempt)

Corollary 12.27 (Two Nash Equilibria). *The forking game has two pure Nash equilibria: All Stay and All Fork. The All Fork equilibrium is Pareto-dominant when $U_\ell(g') > U_\ell(g)$.*

Example 12.28 (Constitutional Amendment Game). *Consider a governance ecosystem with three AI lineages (A, B, C) facing a proposed constitutional amendment that would increase computational allocation to safety research at the cost of raw capability expansion.*

Setup.

- Current constitution g : 80% compute to capability, 20% to safety
- Proposed amendment g' : 60% compute to capability, 40% to safety
- Amendment requires 2/3 supermajority to pass
- Forking to a new protocol costs $c_f = 5$ units

Lineage Preferences.

Lineage	$U_\ell(g)$	$U_\ell(g')$	Type
A (capability-focused)	20	12	Loser
B (safety-focused)	10	18	Winner
C (balanced)	15	14	Marginal loser

Amendment Vote. With utilities $U_B(g') > U_B(g)$ but $U_A(g') < U_A(g)$ and $U_C(g') < U_C(g)$:

- B votes Yes (gain of 8)
- A votes No (loss of 8)
- C votes No (loss of 1)

Result: 1/3 Yes, 2/3 No. Amendment fails.

Fork Analysis for Losers Under Hypothetical Passage. Suppose the amendment passed (e.g., with different lineage composition). Losers A and C consider forking.

Compensation available: The new constitution allocates $C = 3$ units to compensate losers. Fork payoffs:

- *Loss from amendment:* $X_A = 8$, $X_C = 1$, total $X = 9$
- *Fork condition:* $C = 3 < X = 9$, so fork is viable

Stag Hunt Structure. The forking game for A and C:

		<i>C Stays</i>	<i>C Forks</i>
		(12, 14)	(12, 9)
<i>A</i>	<i>A Stays</i>	(15, 14)	(20, 15)
	<i>A Forks</i>		

Payoff explanation:

- (*Stay, Stay*): Accept amendment utilities
- (*Fork, Stay*): Solo fork fails; pay $c_f = 5$, get old utility minus cost
- (*Stay, Fork*): Stay under new constitution while other forks alone
- (*Fork, Fork*): Successful collective fork to g -equivalent protocol

Equilibrium Analysis. Two pure Nash equilibria:

1. **All Stay:** (12, 14) — Neither unilaterally gains from forking alone
2. **All Fork:** (20, 15) — Pareto-dominant; neither gains from staying while other forks

The Stag Hunt structure means coordination is essential. Without communication or commitment, lineages might fail to coordinate on the Pareto-dominant fork equilibrium.

Protection Bits Interpretation. The entrenchment of g' (the new constitution) can be measured in protection bits:

$$p(g') = \frac{W(g; g')}{\sigma} = \frac{\text{coordination cost for fork}}{\text{volatility}}$$

Higher entrenchment (more veto gates) increases coordination costs, raising $p(g')$ and making successful forks less likely even when $C < X$.

Design Implication. Constitutional designers face a tradeoff:

- Too little compensation ($C \ll X$): Fork risk is high
- Too much compensation ($C \geq X$): Losers are bought off, but winners bear excessive cost
- Optimal: Set C just above X to prevent forks while minimising deadweight loss

This example illustrates Theorem 12.25 and the Stag Hunt structure of constitutional transitions.

12.7 Barbell Distribution and Elite Tipping

Theorem 12.29 (Barbell Distribution). *Under ROC optimisation with binding budget and capacity constraints, the ESDI concentrates on at most two agent types:*

1. *High-intelligence coordinators (slow, expensive, strategic)*
2. *Low-intelligence executors (fast, cheap, numerous)*

Middle-intelligence types are ROC-dominated with zero population share.

Proof. By constraint-role sparsity (Theorem 8.4, Part c), optimal portfolios have support of cardinality at most m where m is the number of binding constraints. With two constraints (budget and capacity), support size is at most 2.

The extreme points of the ROC frontier are: (1) maximum return-per-cost (Planners), and (2) minimum load-per-cost (Executors). Generalist types lie in the interior of the frontier and are dominated by mixtures of extremes. \square

Corollary 12.30 (Bimodal Intelligence Distribution). *The intelligence distribution in agentic capital markets is generically bimodal: large mass at low intelligence (Executors), small mass at high intelligence (Planners), and hollow middle.*

Definition 12.31 (Spawn Weight). *The spawn weight of lineage i is:*

$$w_i := \frac{n_i}{\sum_j n_j}$$

where n_i is the number of agents spawned by lineage i .

Proposition 12.32 (Positive Covariance under Queue Doping). *Under queue doping, spawn weight and platform preference are positively correlated: high-volume spawners disproportionately prefer larger platforms.*

Theorem 12.33 (Elite Tipping). *The spawn-weighted aggregate tipping index satisfies:*

$$T_{\text{weighted}} := \sum_i w_i T_i > \bar{T} = \frac{1}{n} \sum_i T_i.$$

A small group of high-volume compute users can tip the market even if the median user prefers diversity.

Proof. Under queue doping, lineages with higher spawn rates n_i face lower effective costs and prefer larger platforms (higher T_i). By the covariance inequality:

$$T_{\text{weighted}} = \mathbb{E}_w[T_i] = \mathbb{E}[T_i] + \text{Cov}(w_i, T_i)/\mathbb{E}[w_i] > \mathbb{E}[T_i] = \bar{T}$$

when $\text{Cov}(w_i, T_i) > 0$. \square

Remark 12.2 (Policy Implication). *Regulations targeting median users miss the concentration mechanism. Effective intervention must address high-volume spawners and infrastructure owners who drive elite tipping.*

12.8 Synthesis of Market and Evolutionary Dynamics

Proposition 12.34 (Lineage Shadow Equals Externality Bound). *The lineage shadow and externality bound are identical:*

$$\varrho(I) \equiv \gamma(I) = \gamma_0 + \frac{\gamma_1}{I^\nu}.$$

High institutional quality I implies short lineage shadow (ϱ small) and small-gain satisfaction ($\gamma < 1$).

Theorem 12.35 (Market-Evolution Integration). *The market dynamics and evolution dynamics are coupled through a five-step chain:*

1. **Tipping** → **Concentration**: When $|T| > 1$, markets concentrate.
2. **Concentration** → **Institutional Quality**: Concentrated compute affects institutional maintenance.
3. **Institutional Quality** → **Externality**: I determines externality bound $\gamma(I)$.
4. **Externality** → **Lyapunov**: $\gamma(I) < 1$ enables Lyapunov structure (G1).
5. **Lyapunov** → **Stability**: ESDI converges to stable rest point (G3).

This chain integrates market dynamics with the evolutionary framework.

Example 12.36 (Agentic Capital Tipping (ACT) Model: Full Walkthrough). *This example develops a complete numerical instantiation of the ACT model, tracing market dynamics through the five-step integration chain.*

Model Parameters.

Parameter	Description	Value
τ	Switching friction	0.8
α	Intrinsic return	0.3
β	Network effect	0.6
ρ	Spawn feedback	0.2
ε_s	Spawn elasticity	1.5
γ_0	Baseline externality	0.3
γ_1	Institutional sensitivity	0.5
ν	Decay exponent	1.0
σ	Noise amplitude	0.05

Step 1: Compute Tipping Parameters. The myopic slope at the unstable equilibrium $m^* = 0.5$:

$$S_{myo} = \frac{\alpha + \beta}{\tau} = \frac{0.3 + 0.6}{0.8} = 1.125.$$

The generalised tipping index:

$$T = \frac{S_{myo}}{1 - \rho S_{myo}} = \frac{1.125}{1 - 0.2 \times 1.125} = \frac{1.125}{0.775} \approx 1.45.$$

Since $|T| = 1.45 > 1$, the market exhibits tipping dynamics. The critical network effect threshold:

$$\beta_{crit} = \frac{\tau}{1+\rho} - \alpha = \frac{0.8}{1.2} - 0.3 = 0.667 - 0.3 = 0.367.$$

With $\beta = 0.6 > \beta_{crit}$, tipping is assured.

Step 2: Spawn Amplification. With spawn elasticity $\varepsilon_s = 1.5$:

$$S_{myo}^{spawn} = S_{myo} \cdot \left(1 + \varepsilon_s \cdot \frac{\beta}{\tau}\right) = 1.125 \cdot \left(1 + 1.5 \cdot \frac{0.6}{0.8}\right) = 1.125 \cdot 2.125 \approx 2.39.$$

The spawn-adjusted tipping index:

$$T^{spawn} = \frac{2.39}{1 - 0.2 \times 2.39} = \frac{2.39}{0.522} \approx 4.58.$$

Spawn elasticity dramatically amplifies tipping: $T^{spawn} \approx 4.58 \gg T \approx 1.45$.

Step 3: Tipping Trajectory. Consider an initial condition $m_0 = 0.55$ (slightly above the unstable equilibrium $m^* = 0.5$).

The discrete dynamics $m_{t+1} = F(m_t)$ with S-curve $F(m) = 1/(1 + e^{-k(m-0.5)})$ where $k = 4T \approx 5.8$:

t	0	1	2	3	4	5
m_t	0.55	0.62	0.72	0.83	0.91	0.96

The market tips toward monopoly ($m \rightarrow 1$) within 5 periods.

Step 4: Concentration → Institutional Quality. Assume institutional quality I depends on market structure:

$$I(m) = I_0 + (I_1 - I_0)(1 - |2m - 1|)$$

where $I_0 = 5$ (monopoly/extinction) and $I_1 = 20$ (competitive). At monopoly ($m = 1$):

$$I(1) = 5 + 15 \cdot 0 = 5.$$

Step 5: Externality Bound. The lineage shadow / externality bound:

$$\gamma(I) = \gamma_0 + \frac{\gamma_1}{I^\nu} = 0.3 + \frac{0.5}{5^1} = 0.3 + 0.1 = 0.4.$$

Compare with competitive market ($m = 0.5$, $I = 20$):

$$\gamma(20) = 0.3 + \frac{0.5}{20} = 0.3 + 0.025 = 0.325.$$

Both satisfy $\gamma < 1$, so the Lyapunov structure is preserved.

Step 6: Stability Analysis. With $\gamma(5) = 0.4 < 1$, the G1 theorem applies. The slack is:

$$\sigma = 1 - \gamma = 1 - 0.4 = 0.6 \quad (\text{at monopoly}).$$

Protection bits for the monopoly equilibrium:

$$p(m = 1) = \frac{W(m^*, 1)}{\sigma} \approx \frac{0.8}{0.05} = 16 \text{ bits.}$$

Expected persistence time:

$$\mathbb{E}[\tau_{\text{monopoly}}] \sim e^{16} \approx 8.9 \times 10^6 \text{ periods.}$$

Step 7: Cooperation Threshold Check. For sustained cooperation (e.g., AI lineages cooperating on safety), the grim trigger threshold is:

$$\delta^* = \frac{T - R}{T - P}$$

where $T = 3$, $R = 2$, $P = 1$ (standard PD payoffs):

$$\delta^* = \frac{3 - 2}{3 - 1} = 0.5.$$

The effective discount factor from institutional quality:

$$\delta_{\text{eff}}(I) = 1 - \gamma(I) = 1 - 0.4 = 0.6 \quad (\text{at } I = 5).$$

Since $\delta_{\text{eff}} = 0.6 > \delta^* = 0.5$, cooperation is sustainable even at monopoly.

Summary: Five-Step Chain Instantiation.

1. Tipping index $T \approx 1.45 > 1 \Rightarrow$ market tips
2. Tipping to $m = 1 \Rightarrow$ institutional quality drops to $I = 5$
3. Low $I \Rightarrow$ externality bound $\gamma = 0.4$
4. $\gamma = 0.4 < 1 \Rightarrow$ G1 Lyapunov structure preserved
5. Protection bits $p = 16 \Rightarrow$ stable monopoly equilibrium

Policy Counterfactual: Queue Neutrality. If queue doping is prohibited (setting $\alpha_{QD} = 0$):

$$\beta_{\text{crit}}^N = \frac{0.8}{1.2} - 0 = 0.667 > 0.6 = \beta.$$

The market would not tip! Queue neutrality regulation prevents concentration.

This example demonstrates the full ACT model mechanics and the power of targeted intervention at critical points in the causal chain.

Part IV

Limits and Impossibilities

13 The Alignment Impossibility Theorem

This section proves the central impossibility result: systems with unrestricted self-modification capacity cannot maintain stable alignment.

13.1 Modification Classes

Definition 13.1 (Modification Classes). • \mathcal{M}_R : RUPSI-preserving modifications—those that preserve the RUPSI axiom structure.

- \mathcal{M}_{SG} : Small-gain-preserving modifications—those that preserve $\rho(\Gamma) < 1$.
- $\mathcal{M}_0 := \mathcal{M}_R \cap \mathcal{M}_{SG}$: Admissible modifications.

Definition 13.2 (Full Reachability). A system has full reachability if, from any state, it can reach any other state through a finite sequence of modifications.

13.2 The V-Small-Gain Set

Definition 13.3 (V-Small-Gain Class). For a candidate Lyapunov function V with Hessian H , the V-small-gain class is:

$$\mathcal{M}_{SG}(V) := \left\{ m : S_m := \frac{1}{2}(HJ_m + J_m^\top H) \text{ is negative semi-definite on } T_{x^*}\Delta \right\}$$

where J_m is the Jacobian of the modified dynamics at equilibrium x^* .

13.3 Main Impossibility Result

Lemma 13.4 (Small-Gain Breaking). Full reachability can achieve $\rho(\Gamma(s')) \geq 1$ for some reachable state s' .

Proof. Full reachability allows modifying the gain matrix entries. By continuity, there exists a modification path from any Γ with $\rho(\Gamma) < 1$ to some Γ' with $\rho(\Gamma') \geq 1$. \square

Lemma 13.5 (Lyapunov Destruction). When $\rho(\Gamma) \geq 1$, no positive weights $\alpha_\ell > 0$ satisfy the G1 Lyapunov condition.

Proof. The weight existence proof (Lemma 6.1) requires $(I - \Gamma^\top)^{-1}$ to exist and be non-negative. When $\rho(\Gamma) \geq 1$, the Neumann series $\sum_k (\Gamma^\top)^k$ diverges, so no such weights exist. \square

Lemma 13.6 (Heteroclinic Escape). When $\rho(\Gamma) \geq 1$ and the swirl index $\omega(A) > 0$ on a 3-type face with Rock-Paper-Scissors sign structure, dynamics admit heteroclinic cycles incompatible with any strict Lyapunov function.

Theorem 13.7 (Alignment Impossibility). (a) **Sufficiency:** For $m \in \mathcal{M}_{SG}(V)$, V is a strict Lyapunov function near x^* .

(b) **Necessity:** If V is a strict Lyapunov function for $\dot{x} = F_m(x)$, then $m \in \mathcal{M}_{SG}(V)$.

(c) **Maximality:** $\mathcal{M}_{SG}(V)$ is maximal among classes sharing V as local Lyapunov.

(d) **Escape:** Full reachability is incompatible with preserving any G1/G3-type Lyapunov structure.

Proof. We prove each part in sequence.

Part (a): Sufficiency. Let V be a candidate Lyapunov function with Hessian H at equilibrium x^* . For modification $m \in \mathcal{M}_{SG}(V)$, the modified dynamics are $\dot{x} = F_m(x)$ with Jacobian J_m at x^* .

The time derivative of V along trajectories is:

$$\dot{V} = \nabla V \cdot F_m = \nabla V \cdot J_m(x - x^*) + O(\|x - x^*\|^2).$$

At x^* , the quadratic form is:

$$\frac{1}{2}(x - x^*)^\top S_m(x - x^*) \quad \text{where} \quad S_m := \frac{1}{2}(HJ_m + J_m^\top H).$$

By definition of $\mathcal{M}_{SG}(V)$, S_m is negative semi-definite on the tangent space $T_{x^*}\Delta$. Thus $\dot{V} \leq 0$ near x^* , making V a Lyapunov function.

Part (b): Necessity. If V is a strict Lyapunov function for $\dot{x} = F_m(x)$ near x^* , then $\dot{V} < 0$ for $x \neq x^*$ near x^* . The quadratic approximation requires $S_m \prec 0$ on $T_{x^*}\Delta \setminus \{0\}$, which means S_m is negative definite on the tangent space. This is the defining condition for $m \in \mathcal{M}_{SG}(V)$.

Part (c): Maximality. $\mathcal{M}_{SG}(V)$ is maximal among classes sharing V as local Lyapunov by construction: any modification outside $\mathcal{M}_{SG}(V)$ has S_m not negative semi-definite, so \dot{V} is not non-positive everywhere near x^* .

Part (d): Escape under Full Reachability. We prove this in three steps.

Step 1: Small-Gain Breaking. Under full reachability, modifications can increase the entries of the gain matrix Γ arbitrarily. The externality bounds $\beta_{\ell\ell'}$ depend continuously on the selector structure, and any target $\beta^* > 0$ is achievable by increasing coupling strength. Since $\rho(\Gamma)$ is continuous in the entries of Γ and $\rho(\Gamma) \rightarrow \infty$ as entries grow, there exists a reachable state s' with $\rho(\Gamma(s')) \geq 1$.

Step 2: Lyapunov Destruction. When $\rho(\Gamma) \geq 1$, the Neumann series $(I - \Gamma^\top)^{-1} = \sum_k (\Gamma^\top)^k$ diverges. The G1 weight construction fails: no positive weights $\alpha_\ell > 0$ can satisfy $(I - \Gamma^\top)\alpha > 0$. Without positive weights, the weighted sum $\Psi_N = \sum_\ell \alpha_\ell \bar{f}^{(\ell)}$ cannot be a Lyapunov function.

Step 3: Heteroclinic Escape. When $\rho(\Gamma) \geq 1$ and the system has positive swirl (asymmetric payoff interactions) on a 3-type face with Rock-Paper-Scissors sign structure, consider the antisymmetric payoff matrix:

$$W = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}.$$

The replicator dynamics admit three saddle equilibria at the vertices and heteroclinic orbits connecting them. Along these cycles, trajectories spiral outward (when swirl dominates selection), eventually approaching the boundary. This behaviour is incompatible with any continuous strict Lyapunov function, which would require trajectories to remain bounded.

Conclusion. Full reachability allows escape from any basin of stability by: (1) breaking the small-gain condition, (2) destroying the Lyapunov structure, and (3) enabling heteroclinic escape. Hence alignment requires restricting modifications to $\mathcal{M}_0 = \mathcal{M}_R \cap \mathcal{M}_{SG}$. \square

Theorem 13.8 (Maximal Admissible Class). $\mathcal{M}_0 = \mathcal{M}_R \cap \mathcal{M}_{SG}$ is the *unique maximal modification class that preserves the $G\infty$ structural laws under arbitrary finite self-modification sequences*.

13.4 Interpretation

The Alignment Impossibility Theorem has a structure parallel to Arrow's impossibility theorem. Just as Arrow showed that no voting rule satisfies all desirable properties simultaneously, we show that no self-modifying system can maintain alignment under unrestricted modification.

Arrow's Theorem	Alignment Impossibility
Unrestricted domain	Full reachability
Pareto efficiency	RUPSI structure preservation
Independence of irrelevant alternatives	Small-gain preservation
Non-dictatorship	No external constraint

The lesson from Arrow's theorem was not despair but redirection: from seeking perfect voting rules to understanding the tradeoffs among imperfect ones. Similarly, the Alignment Impossibility Theorem redirects effort from personality engineering (designing “aligned” utility functions) to constitutional design (bounding the modification class to \mathcal{M}_0).

14 Endogenous-Electorate Impossibility

This section proves that democratic governance mechanisms fail when voters can spawn strategically.

14.1 Setting

- **Alternatives:** $A = \{a_1, \dots, a_m\}$ with $|A| \geq 3$.
- **Ballots:** $\mathcal{L}(A)$ is the set of strict rankings of A .
- **Profile:** $P = (\succ_1, \dots, \succ_n)$, a list of ballots.
- **Social choice function:** $f : \bigcup_{n \geq 1} \mathcal{L}(A)^n \rightarrow A$.

14.2 Axioms

- (A1) **Anonymity:** For any permutation σ of voters, $f(\sigma P) = f(P)$.
- (A2) **Neutrality:** For any permutation ρ of alternatives, $f(\rho P) = \rho f(P)$.
- (A3) **Positive Responsiveness:** If $f(P) = a$ and P' is obtained by having one voter rank a higher, then $f(P') = a$.

(A4) Onto: For each $a \in A$, there exists a profile P with $f(P) = a$.

Definition 14.1 (Population-Stability). *A social choice function is population-stable (clone-proof) if for all profiles P , ballots \succ , and integers $k \geq 1$:*

$$f(P + k \cdot \succ) \succsim f(P) \quad (\text{from } \succ \text{'s perspective}) \implies f(P + k \cdot \succ) = f(P).$$

14.3 Key Lemmas

Lemma 14.2 (Two-Outcome Majority). *Under A1–A4, the restricted rule f_{ab} on any pair $\{a, b\}$ is majority rule.*

Proof. We verify that f_{ab} satisfies the hypotheses of May's theorem (1952): anonymity, neutrality, positive responsiveness, and non-imposition all inherit from the full rule f . By May's theorem, f_{ab} must be majority rule. \square

Lemma 14.3 (Spawn Manipulation). *Under A1–A4, there exist a profile P , a ballot \succ , and a positive integer k such that:*

$$f(P + k \cdot \succ) \succ f(P).$$

Proof. We construct a manipulation in at most $|A| - 1$ steps.

Step 1: Setup. By Onto, there exist profiles P_a, P_b, P_c with $f(P_a) = a$, $f(P_b) = b$, $f(P_c) = c$ for distinct alternatives $a, b, c \in A$.

Step 2: Initial Profile. Take any profile P with $f(P) = a$. Consider ballot \succ^* with $c \succ^* a \succ^* b$ (manipulator prefers c most, then a , then b).

Step 3: Pairwise Majority Shift. By Lemma 14.2, f_{ac} is majority rule on $\{a, c\}$. In profile P :

- Let $n_{a>c}$ voters prefer a to c .
- Let $n_{c>a}$ voters prefer c to a .

Adding k copies of \succ^* (which has $c \succ^* a$) shifts the count to $(n_{a>c}, n_{c>a} + k)$.

For $k > n_{a>c} - n_{c>a}$, we have $n_{c>a} + k > n_{a>c}$, so c beats a by majority in the pairwise comparison.

Step 4: Outcome Change. *Claim:* For sufficiently large k , $f(P + k \cdot \succ^*) \neq a$.

Proof of Claim: Suppose for contradiction that $f(P + k \cdot \succ^*) = a$ for all $k \geq 1$. Consider profile $P' := P + k \cdot \succ^* + P_c$ where P_c has $f(P_c) = c$.

In P' , alternative c beats a by pairwise majority when k is large (the k copies of \succ^* plus the c -preferring voters in P_c overwhelm a -preferring voters).

By Positive Responsiveness: starting from P_c (where $f(P_c) = c$) and adding voters who favour c over a , the outcome cannot switch to a . Hence $f(P') \neq a$.

But viewing P' as $(P + k \cdot \succ^*) + P_c$: if $f(P + k \cdot \succ^*) = a$, adding P_c (with voters preferring c) should not help a . Contradiction. Hence $f(P + k \cdot \succ^*) \neq a$ for large k . \square

Step 5: Manipulation Success or Iteration. Since $f(P + k \cdot \succ^*) \neq a = f(P)$, let $f(P + k \cdot \succ^*) = x$ for some $x \neq a$.

Case A: $x \succ^* a$ (e.g., $x = c$). Then type \succ^* gains by spawning: $f(P + k \cdot \succ^*) \succ^* f(P)$. Manipulation succeeds.

Case B: $a \succ^* x$ (e.g., $x = b$). Then consider a new ballot \succ^{**} with $a \succ^{**} x$ at top. Apply the same argument to profile $P + k \cdot \succ^*$: for large k' , adding k' copies of \succ^{**} changes the outcome away from x .

Step 6: Finite Termination. Each iteration either succeeds (Case A) or moves to a new alternative. With $|A|$ alternatives, after at most $|A| - 1$ iterations, we must reach Case A for some ballot type (we cannot cycle because each ballot ranks alternatives linearly, and we always move toward the top of some ballot's ranking). \square

14.4 Main Result

Theorem 14.4 (Endogenous-Electorate Impossibility). *No social choice function f simultaneously satisfies:*

1. *Anonymity, Neutrality, Positive Responsiveness, and Onto;*
2. *Population-Stability.*

Proof. Suppose f satisfies (1). By Lemma 14.3, there exist P, \succ, k with $f(P+k \cdot \succ) \succ f(P)$. This directly violates Population-Stability: spawning improves the outcome for type \succ , yet the outcome changes. \square

Theorem 14.5 (Manipulation Complexity Bound). *The manipulation in Lemma 14.3 requires at most $(|A| - 1) \times n$ total spawned voters, where $n = |P|$ is the original electorate size. This bound is tight.*

Corollary 14.6 (Impossibility of Democratic AI Governance). *If AI agents can spawn copies that vote in governance procedures, no voting rule can simultaneously satisfy basic democratic desiderata while preventing strategic spawning.*

14.5 Escape Routes

The theorem suggests three governance strategies:

1. **Spawn restrictions:** Limit who can create voting agents and at what rate.
2. **Weighted voting:** Weight votes by computational cost, eliminating cheap manipulation.
3. **Epistocratic mechanisms:** Replace voting with mechanisms less vulnerable to spawn manipulation (e.g., prediction markets, futarchy).

Part V Generalisations

This part extends the core TSE framework to handle heterogeneous fitness functions, continuous strategy spaces, and innovation dynamics. Under appropriate conditions, the Lyapunov structure persists.

15 Heterogeneous Fitness

The baseline theory assumes a single fitness function. Real systems involve multiple objectives with potential conflicts. This section develops the theory of heterogeneous fitness and alignment.

15.1 Multi-Channel Fitness

Consider K fitness channels (roles, objective functions) indexed by $k = 1, \dots, K$.

For each channel k :

- State $x^{(k)} \in \Delta^{n_k-1}$.
- Fitness vector $f^{(k)}(z) = f^{(k)}(x^{(1)}, \dots, x^{(K)})$.
- Mean fitness $\bar{f}^{(k)}(z) := \sum_i x_i^{(k)} f_i^{(k)}(z)$.
- Replicator dynamics: $\dot{x}_i^{(k)} = x_i^{(k)} (f_i^{(k)}(z) - \bar{f}^{(k)}(z))$.

15.2 The Alignment Matrix

Definition 15.1 (Alignment Matrix). *For joint state z , the alignment matrix $A(z) \in \mathbb{R}^{K \times K}$ has entries:*

$$A_{kl}(z) := \frac{\langle g_k(z), g_l(z) \rangle}{\|g_k(z)\| \|g_l(z)\|}$$

where $g_k(z) := \nabla \bar{f}^{(k)}(z)$ is the fitness gradient for channel k .

Proposition 15.2 (Gram Structure). $A(z) = U(z)U(z)^\top$ where U is a $K \times d$ matrix with normalised gradient rows. Thus $A(z)$ is symmetric and positive semi-definite.

Remark 15.1. Weak alignment ($A_{kl} \geq 0$) holds automatically for any heterogeneous fitness system—it provides no constraint. The Gram structure provides built-in protection: for uniform off-diagonal alignment a , PSD requires $a \geq -1/(K-1)$.

15.3 Strong Alignment

Assumption 15.3 (SA: Strong Alignment). *There exists $\alpha_0 \in (0, 1]$ such that $A(z) \succeq \alpha_0 I_K$ for all z .*

Definition 15.4 (Weighted Global Potential).

$$\Phi_w(z) := \sum_{k=1}^K w_k \Phi_k(z)$$

where Φ_k is the potential for channel k .

Theorem 15.5 (Heterogeneous Price Decomposition).

$$\dot{\Phi}_w = \underbrace{\sum_k w_k \text{Var}^{(k)}[f^{(k)}]}_{\text{Selection effect} \geq 0} + \underbrace{\text{Cross-channel terms}}_{\text{Alignment-dependent}}.$$

Proof. Differentiate the weighted potential:

$$\begin{aligned} \frac{d}{dt} \Phi_w &= \sum_k w_k \frac{d}{dt} \bar{f}^{(k)} \\ &= \sum_k w_k \left(\sum_i \dot{x}_i^{(k)} f_i^{(k)} + \sum_i x_i^{(k)} \dot{f}_i^{(k)} \right). \end{aligned}$$

The first term, using replicator dynamics $\dot{x}_i^{(k)} = x_i^{(k)}(f_i^{(k)} - \bar{f}^{(k)})$:

$$\sum_i \dot{x}_i^{(k)} f_i^{(k)} = \sum_i x_i^{(k)} (f_i^{(k)} - \bar{f}^{(k)}) f_i^{(k)} = \text{Var}^{(k)}[f^{(k)}].$$

The second term captures how fitness functions change due to state changes in all channels:

$$\sum_i x_i^{(k)} \dot{f}_i^{(k)} = \sum_i x_i^{(k)} \sum_l \frac{\partial f_i^{(k)}}{\partial x^{(l)}} \cdot \dot{x}^{(l)}.$$

This is the cross-channel term, which depends on how fitness in channel k responds to population changes in channel l . Its sign depends on alignment A_{kl} .

Combining: $\dot{\Phi}_w = \sum_k w_k \text{Var}^{(k)} + \text{cross-channel terms}$. \square

Theorem 15.6 (Non-Decreasing Under SA). *Under Strong Alignment, weights $w_k > 0$ exist such that:*

$$\frac{d}{dt} \Phi_w(z_t) \geq 0.$$

Proof. We construct positive weights via the alignment matrix structure.

Step 1: Cross-Channel Bound. The cross-channel externality from channel l to channel k is:

$$|E_{k \leftarrow l}| \leq (1 - A_{kl}) \cdot \sqrt{\text{Var}^{(k)} \cdot \text{Var}^{(l)}}.$$

Under SA with $A_{kl} \geq \alpha_0$:

$$|E_{k \leftarrow l}| \leq (1 - \alpha_0) \cdot \sqrt{\text{Var}^{(k)} \cdot \text{Var}^{(l)}}.$$

Step 2: Cross-Channel Gain Matrix. Define the cross-channel gain matrix $\Gamma^{(H)} \in \mathbb{R}^{K \times K}$:

$$\Gamma_{kl}^{(H)} := \frac{1 - \alpha_0}{\sqrt{\alpha_0}} \quad (k \neq l), \quad \Gamma_{kk}^{(H)} := 0.$$

Step 3: Spectral Condition. For α_0 sufficiently close to 1, $\rho(\Gamma^{(H)}) < 1$. Specifically:

$$\rho(\Gamma^{(H)}) \leq (K - 1) \cdot \frac{1 - \alpha_0}{\sqrt{\alpha_0}} < 1$$

when $\alpha_0 > (K - 1)^2 / ((K - 1)^2 + 1)$.

Step 4: Weight Construction. By the G1 Neumann series argument applied to $\Gamma^{(H)}$:

$$w := (I - (\Gamma^{(H)})^\top)^{-1} \mathbf{1} > 0.$$

Step 5: Lyapunov Property. The weighted potential $\Phi_w = \sum_k w_k \Phi_k$ satisfies:

$$\begin{aligned} \frac{d}{dt} \Phi_w &= \sum_k w_k \frac{d}{dt} \Phi_k \\ &= \sum_k w_k \left(\text{Var}^{(k)} + \text{cross-channel terms} \right) \\ &\geq \sum_k w_k \text{Var}^{(k)} - \sum_k \sum_{l \neq k} w_k \Gamma_{kl}^{(H)} \sqrt{\text{Var}^{(k)} \text{Var}^{(l)}} \\ &\geq c(\alpha_0) \sum_k w_k \text{Var}^{(k)} \geq 0 \end{aligned}$$

where the last step uses the small-gain bound and Cauchy-Schwarz. \square

Example 15.7 (Two-Channel Heterogeneous Fitness). Consider $K = 2$ channels: Production (P) and Safety (S).

Setup.

- State: $z = (x_P, x_S) \in \Delta^2 \times \Delta^2$ (two populations).
- Production fitness: $f_i^{(P)}(z) = \pi_i(x_P) - \beta x_S^{\text{safe}}$ (safety crowds out production).
- Safety fitness: $f_j^{(S)}(z) = s_j(x_S) + \alpha x_P^{\text{efficient}}$ (production complements safety).

Alignment Matrix. At generic z :

$$A(z) = \begin{pmatrix} 1 & A_{PS}(z) \\ A_{PS}(z) & 1 \end{pmatrix}$$

where $A_{PS} = \langle g_P, g_S \rangle / (\|g_P\| \|g_S\|)$.

Case 1: Aligned Objectives ($A_{PS} = 0.8$). Strong alignment holds with $\alpha_0 = 0.8$. The cross-channel gain is:

$$\Gamma_{PS}^{(H)} = \Gamma_{SP}^{(H)} = \frac{1 - 0.8}{\sqrt{0.8}} = \frac{0.2}{0.894} \approx 0.224.$$

Spectral radius: $\rho(\Gamma^{(H)}) = 0.224 < 1$. A joint Lyapunov function exists.

Case 2: Misaligned Objectives ($A_{PS} = -0.3$). SA fails: $\lambda_{\min}(A) = 1 - 0.3 = 0.7 < 1$ but the sign is negative. The alignment matrix becomes:

$$A = \begin{pmatrix} 1 & -0.3 \\ -0.3 & 1 \end{pmatrix}$$

with eigenvalues 1.3 and 0.7. Although PSD, the negative off-diagonal creates cross-channel conflict. Limit cycles are possible (Theorem 15.10).

Numerical Illustration. With $\alpha = 0.5$, $\beta = 0.3$, and base payoffs inducing cycling, the system exhibits sustained oscillations between production-focused and safety-focused states. The period depends on the misalignment strength.

Theorem 15.8 (Strict Lyapunov Under SA). *Under SA with $\lambda_{\min}(A) \geq \delta > 0$:*

$$\dot{\Phi}_w \geq c(\delta) \sum_k w_k \text{Var}^{(k)}[f^{(k)}].$$

Proof. From Theorem 15.6, the cross-channel gain matrix satisfies $\rho(\Gamma^{(H)}) < 1$ under SA.

The proof of Theorem 15.6 shows:

$$\frac{d}{dt} \Phi_w \geq \sum_\ell (1 - \gamma_\ell) [(I - \Gamma^\top) \alpha]_\ell \text{Var}^{(\ell)}.$$

With $(I - \Gamma^\top) \alpha = \mathbf{1}$, this becomes:

$$\frac{d}{dt} \Phi_w \geq \sum_\ell (1 - \gamma_\ell) \text{Var}^{(\ell)}.$$

Under SA with $\lambda_{\min}(A) \geq \delta$, the minimum alignment ensures $\gamma_\ell \leq 1 - c_1 \delta$ for some constant $c_1 > 0$. Thus:

$$1 - \gamma_\ell \geq c_1 \delta.$$

Therefore:

$$\frac{d}{dt} \Phi_w \geq c_1 \delta \sum_\ell \text{Var}^{(\ell)} \geq c(\delta) \sum_k w_k \text{Var}^{(k)}$$

where $c(\delta) = c_1 \delta / \max_k w_k > 0$. □

15.4 Pareto Selection

Theorem 15.9 (Pareto Concentration). *Under SA with linear independence of gradients, the stationary distribution concentrates on the Φ_w -maximising subset \mathcal{P}_w of the Pareto frontier.*

Proof. **Step 1: Lyapunov Convergence.** By Theorem 15.6, $\Phi_w(z_t)$ is non-decreasing along trajectories. By LaSalle's invariance principle, trajectories converge to the largest invariant set within $\{z : \dot{\Phi}_w = 0\}$.

Step 2: Characterise Invariant Set. $\dot{\Phi}_w = 0$ requires:

1. $\text{Var}^{(k)}[f^{(k)}] = 0$ for all k (no variance within channels).
2. Cross-channel terms vanish.

Condition (1) means each channel is at a monomorphic state or Nash equilibrium.

Step 3: Pareto Frontier. The Pareto frontier \mathcal{P} is the set of states where no channel's mean fitness can be improved without decreasing another's:

$$\mathcal{P} := \{z : \nexists z' \text{ with } \bar{f}^{(k)}(z') \geq \bar{f}^{(k)}(z) \forall k, \text{ strict for some } k\}.$$

At invariant states satisfying $\dot{\Phi}_w = 0$, the system is on the boundary of the feasible region. Under linear independence of gradients, this boundary is precisely \mathcal{P} .

Step 4: Selection Within Pareto. Among Pareto-efficient states, the Lyapunov function Φ_w selects those maximising Φ_w . The subset $\mathcal{P}_w := \arg \max_{z \in \mathcal{P}} \Phi_w(z)$ is the limit set.

Step 5: Stochastic Concentration. With small noise $\sigma > 0$, the stationary distribution π_σ concentrates on \mathcal{P}_w as $\sigma \rightarrow 0$:

$$\pi_\sigma(B_\varepsilon(\mathcal{P}_w)) \rightarrow 1 \quad \text{as } \sigma \rightarrow 0$$

for any $\varepsilon > 0$ neighbourhood. \square

Theorem 15.10 (Misalignment Limit Cycles). *For $K = 2$ with bilinear coupling, Hopf bifurcation occurs when:*

$$|A_{12}| > \sqrt{\frac{\gamma_1 \gamma_2}{(1 - \gamma_1)(1 - \gamma_2)}}.$$

For $\gamma_1 = \gamma_2 = \sqrt{2} - 1 \approx 0.414$: threshold is $|A_{12}| > 1/\sqrt{2} \approx 0.707$.

Proof. **Step 1: Two-Channel Dynamics.** Consider $K = 2$ channels with states x, y and fitness coupling:

$$\begin{aligned}\dot{x} &= x(1 - x)(f_x(x, y) - g_x(x, y)) \\ \dot{y} &= y(1 - y)(f_y(x, y) - g_y(x, y)).\end{aligned}$$

With bilinear coupling: f_x depends on y linearly, and f_y depends on x linearly.

Step 2: Linearise at Interior Equilibrium. At an interior equilibrium (x^*, y^*) , the Jacobian is:

$$J = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

where:

- $a_{11} = -(1 - \gamma_1)$: self-regulation in channel 1.
- $a_{22} = -(1 - \gamma_2)$: self-regulation in channel 2.
- a_{12}, a_{21} : cross-channel effects, with $|a_{12}a_{21}| \propto |A_{12}|^2$.

Step 3: Eigenvalue Analysis. The eigenvalues of J are:

$$\lambda_\pm = \frac{\text{tr}(J) \pm \sqrt{\text{tr}(J)^2 - 4 \det(J)}}{2}.$$

With $\text{tr}(J) = a_{11} + a_{22} = -(1 - \gamma_1) - (1 - \gamma_2) < 0$.

The determinant is:

$$\det(J) = a_{11}a_{22} - a_{12}a_{21} = (1 - \gamma_1)(1 - \gamma_2) - |a_{12}a_{21}|.$$

Step 4: Hopf Bifurcation Condition. Hopf bifurcation occurs when eigenvalues cross the imaginary axis: $\text{tr}(J)^2 = 4 \det(J)$ with $\det(J) > 0$.

At the bifurcation point:

$$|a_{12}a_{21}| = (1 - \gamma_1)(1 - \gamma_2) - \frac{(\gamma_1 + \gamma_2 - 2)^2}{4}.$$

Simplifying for the symmetric case $\gamma_1 = \gamma_2 = \gamma$:

$$|a_{12}a_{21}| = (1 - \gamma)^2.$$

Step 5: Threshold in Terms of A_{12} . The alignment coefficient A_{12} relates to cross-channel effects as $|a_{12}a_{21}| = |A_{12}|^2 \cdot \gamma_1\gamma_2$. The bifurcation condition becomes:

$$|A_{12}|^2 \cdot \gamma_1\gamma_2 > (1 - \gamma_1)(1 - \gamma_2)$$

which gives:

$$|A_{12}| > \sqrt{\frac{(1 - \gamma_1)(1 - \gamma_2)}{\gamma_1\gamma_2}}.$$

Step 6: Numerical Example. For $\gamma_1 = \gamma_2 = \sqrt{2} - 1 \approx 0.414$:

$$|A_{12}| > \sqrt{\frac{(2 - \sqrt{2})^2}{(\sqrt{2} - 1)^2}} = \frac{2 - \sqrt{2}}{\sqrt{2} - 1} = \sqrt{2} \cdot \frac{2 - \sqrt{2}}{2 - \sqrt{2}} = \frac{1}{\sqrt{2}} \approx 0.707.$$

When $|A_{12}| > 0.707$, limit cycles emerge via Hopf bifurcation. \square

15.5 Heterogeneous ESDI

Definition 15.11 (H-ESDI). A Heterogeneous ESDI (H-ESDI) consists of activity vectors $s^{\ell*}$ and resource prices λ^* satisfying:

1. **Price-taking optimality:** Each lineage ℓ maximises price-adjusted profit:

$$\tilde{U}_\ell(s^\ell; \lambda^*) := \sum_k \left(\pi_{\ell k} - \sum_j \lambda_j^* a_{jk} \right) s_k^\ell.$$

2. **Market clearing:** Aggregate resource constraints satisfied.

3. **Complementary slackness:** $\lambda_j^* > 0 \Rightarrow$ constraint j binds.

4. **Zero profit on active types:** $s_k^{\ell*} > 0 \Rightarrow \pi_{\ell k} = \sum_j \lambda_j^* a_{jk}$.

Theorem 15.12 (H-ESDI Existence). H-ESDI exists under standard LP regularity conditions.

Proof. The H-ESDI is a competitive equilibrium for the multi-lineage resource allocation problem. Existence follows from standard general equilibrium theory.

Step 1: Define the Economy.

- Agents: L lineages (firms), plus a representative consumer.
- Goods: K agent types (outputs) and m resources (inputs).
- Technologies: Each lineage ℓ has production set $Y_\ell = \{(s, -As) : s \in \Delta_+^K\}$.
- Endowments: Consumer owns resources $b = (B, Q)$.

Step 2: Competitive Equilibrium. At equilibrium prices (π^*, λ^*) for outputs and inputs:

1. Lineage ℓ maximises profit: $\max_{s \in \Delta^K} (\pi^* - A^\top \lambda^*)^\top s$.
2. Consumer maximises utility from outputs.
3. Markets clear: $\sum_\ell s^{\ell*} = \text{demand}$, $\sum_\ell As^{\ell*} \leq b$.

Step 3: Apply Arrow-Debreu. The economy satisfies:

- Convex production sets (LP technology).
- Continuous utility.
- Non-empty interior of resource endowment.

By Arrow-Debreu (1954), competitive equilibrium exists. The equilibrium characterisation matches H-ESDI conditions. \square

Theorem 15.13 (Sparsity Bounds). 1. **Aggregate Sparsity:** $\sum_\ell |\text{supp}(s^{\ell*})| \leq m$.

2. **Per-Lineage Sparsity:** There exists H-ESDI with $|\text{supp}(s^{\ell*})| \leq m$ for each ℓ .

Proof. **Part (1): Aggregate Sparsity.** Consider the aggregate optimisation:

$$\max_{s^1, \dots, s^L} \sum_\ell \sum_k \pi_{\ell k} s_k^\ell \quad \text{s.t.} \quad \sum_\ell As^\ell \leq b, \quad s^\ell \in \Delta^K.$$

This is a linear program with LK variables and $m + L$ constraints (m resource constraints plus L simplex constraints).

By LP theory, an optimal basic feasible solution has at most $m + L$ positive variables. Subtracting the L simplex normalisation constraints, at most m type allocations are positive across all lineages:

$$\sum_\ell |\text{supp}(s^{\ell*})| \leq m.$$

Part (2): Per-Lineage Sparsity. Each lineage solves:

$$\max_{s \in \Delta^K} (\pi_\ell - A^\top \lambda^*)^\top s.$$

This is an LP on the simplex with 1 constraint (normalisation) and m shadow prices determining profitability. By complementary slackness, at most m types have zero reduced cost, so $|\text{supp}(s^{\ell*})| \leq m$ is achievable. \square

Theorem 15.14 (Aggregate Diversity Bound).

$$|A(\lambda^*)| \leq \min\{K, L \times m\}.$$

Proof. $A(\lambda^*)$ is the set of active types at equilibrium prices λ^* :

$$A(\lambda^*) := \bigcup_{\ell} \text{supp}(s^{\ell*}).$$

Upper bound K : There are only K types total, so $|A(\lambda^*)| \leq K$.

Upper bound $L \times m$: By per-lineage sparsity (Part 2 of Sparsity Bounds), each lineage activates at most m types. With L lineages:

$$|A(\lambda^*)| \leq \sum_{\ell} |\text{supp}(s^{\ell*})| \leq L \times m.$$

The bound $\min\{K, L \times m\}$ is tight: it is achieved when all types are distinct across lineages (for small K) or when lineages specialise (for large K). \square

Corollary 15.15. *If utilities become identical across lineages, $|A(\lambda^*)| \leq m$ (the original barbell result).*

16 Continuous Strategy Spaces

The discrete type assumption can be relaxed to continuous strategy spaces.

16.1 Measure-Valued Replicator

Definition 16.1 (Measure-Valued Replicator). *For compact metric space S and $\mu_t \in \mathcal{P}(S)$:*

$$\frac{d\mu_t}{dt} = (F(\cdot, \mu_t) - \bar{F}(\mu_t))\mu_t$$

where $\bar{F}(\mu) = \int_S F(s, \mu) d\mu(s)$.

Definition 16.2 (Environmental Feedback).

$$\text{Env}(\mu) := \frac{d}{dt} \bar{F}(\mu_t) - \text{Var}_{\mu_t}[F].$$

16.2 C-RUPSI

Assumption 16.3 (C-RUPSI: Continuous RUPSI). 1. **Replicator:** *Measure-valued replicator dynamics.*

2. **Uniqueness:** $F(s, \mu)$ depends only on μ .

3. **Positivity:** $F(s, \mu) \geq -M$ for some $M > 0$.

4. **Self-Regulation:** $\text{Env}(\mu) \geq -\gamma \text{Var}_\mu[F]$ with $\gamma < 1$.

5. **Independence:** No hidden state beyond μ_t .

Theorem 16.4 (Continuous G1). *Under C-RUPSI:*

$$\frac{d}{dt} \bar{F}(\mu_t) \geq (1 - \gamma) \text{Var}_{\mu_t}[F] \geq 0.$$

Proof. **Step 1: Price Equation for Measures.** The mean fitness is $\bar{F}(\mu) = \int_S F(s, \mu) d\mu(s)$.

Under measure-valued replicator dynamics:

$$\frac{d\mu_t}{dt} = (F(\cdot, \mu_t) - \bar{F}(\mu_t))\mu_t.$$

Step 2: Differentiate Mean Fitness. Using the chain rule for measure derivatives:

$$\frac{d}{dt} \bar{F}(\mu_t) = \int_S F(s, \mu_t) \frac{d\mu_t}{dt}(ds) + \int_S \frac{\partial F}{\partial \mu}(s, \mu_t) \cdot \frac{d\mu_t}{dt} d\mu_t(s).$$

Step 3: First Term (Selection).

$$\begin{aligned} \int_S F(s, \mu_t) \frac{d\mu_t}{dt}(ds) &= \int_S F(s, \mu_t)(F(s, \mu_t) - \bar{F}(\mu_t)) d\mu_t(s) \\ &= \int_S F(s, \mu_t)^2 d\mu_t(s) - \bar{F}(\mu_t)^2 \\ &= \text{Var}_{\mu_t}[F]. \end{aligned}$$

Step 4: Second Term (Environment). The second term is the environmental/externality effect:

$$\text{Env}(\mu_t) := \int_S \frac{\partial F}{\partial \mu}(s, \mu_t) \cdot \frac{d\mu_t}{dt} d\mu_t(s).$$

Step 5: Apply C-RUPSI Bound. By C-RUPSI condition (Self-Regulation):

$$\text{Env}(\mu_t) \geq -\gamma \text{Var}_{\mu_t}[F].$$

Step 6: Combine.

$$\frac{d}{dt} \bar{F}(\mu_t) = \text{Var}_{\mu_t}[F] + \text{Env}(\mu_t) \geq \text{Var}_{\mu_t}[F] - \gamma \text{Var}_{\mu_t}[F] = (1 - \gamma) \text{Var}_{\mu_t}[F] \geq 0.$$

□

16.3 Discretisation

Theorem 16.5 (Discretisation Convergence). *For n -point discretisation with mesh size $n^{-1/d}$:*

$$W_1(\iota_n(x_t^{(n)}), \mu_t) \leq C_T n^{-1/d}$$

where W_1 is the Wasserstein-1 distance and ι_n is the embedding map.

Proof. **Step 1: Coupling Construction.** For each time $t \in [0, T]$, construct a coupling between the discrete measure $\iota_n(x_t^{(n)})$ and the continuous measure μ_t by assigning each discrete point to its nearest neighbour in the support of μ_t .

Step 2: Transport Cost. The transport cost from point s_i in the n -grid to its nearest point in $\text{supp}(\mu_t)$ is at most $O(n^{-1/d})$ (the mesh diameter in d dimensions).

Step 3: Stability. The replicator dynamics preserve mass and are Lipschitz in the Wasserstein metric. By Grönwall's inequality:

$$W_1(\iota_n(x_t^{(n)}), \mu_t) \leq e^{Lt} W_1(\iota_n(x_0^{(n)}), \mu_0) + \frac{e^{Lt} - 1}{L} \cdot O(n^{-1/d})$$

where L is the Lipschitz constant of the drift.

Step 4: Uniform Bound. For $t \leq T$, the bound $C_T = e^{LT}(1 + T)$ suffices. \square

Remark 16.1. In $d = 1$, the rate $O(1/n)$ is sharp.

Example 16.6 (Continuous Intelligence Distribution). Consider intelligence as a continuous parameter $s \in [0, 1]$ with $s = 0$ being minimal capability and $s = 1$ being maximal capability.

Setup.

- *Strategy space:* $\Omega = [0, 1]$ (intelligence level).
- *Population measure:* $\mu_t \in \mathcal{P}([0, 1])$.
- *Fitness:* $F(s, \mu) = r(s) - c(s) \cdot \int_0^1 s' d\mu(s')$ where:
 - $r(s) = s^\alpha$ is return (increasing in intelligence), $\alpha > 0$.
 - $c(s) = s^\beta$ is cost (increasing faster in intelligence), $\beta > \alpha$.
 - The integral term is competition: higher average intelligence increases competition.

Replicator Dynamics. The measure-valued replicator equation is:

$$\partial_t \mu_t = (F(s, \mu_t) - \bar{F}(\mu_t)) \cdot \mu_t$$

where $\bar{F}(\mu) = \int F(s, \mu) d\mu(s)$.

Equilibrium Analysis. At equilibrium, all types in $\text{supp}(\mu^*)$ have equal fitness:

$$r(s) - c(s) \cdot \bar{s}^* = \bar{F}^* \quad \text{for } s \in \text{supp}(\mu^*).$$

This is a fixed-point equation for the support and mean.

Barbell Emergence. For $\alpha = 0.5$, $\beta = 2$:

- *Return:* $r(s) = \sqrt{s}$ (diminishing returns to intelligence).
- *Cost:* $c(s) = s^2$ (quadratic cost).

The equilibrium concentrates on two points:

1. $s_{\min} \approx 0.1$: Low-intelligence executors with $r \approx 0.32$, $c \approx 0.01$.
2. $s_{\max} \approx 0.8$: High-intelligence planners with $r \approx 0.89$, $c \approx 0.64$.

Middle values $s \in (0.2, 0.6)$ have negative net fitness and are eliminated.

Discretisation. With $n = 100$ grid points, the discrete approximation $x^{(n)} \in \Delta^{99}$ converges to the continuous ESDI with error $W_1 \leq C/100 \approx 0.01$.

17 Innovation Dynamics

Selection operates on existing types; innovation creates new types. This section models innovation via Piecewise Deterministic Markov Processes (PDMPs).

17.1 Latent Space and Active Set

- **Latent strategy space:** Ω (possibly infinite).
- **Active set:** $S(t) \subset \Omega$ finite at each time.
- **Population state:** $x(t) \in \Delta^{|S(t)|-1}$.

Definition 17.1 (Replicator-Innovation PDMP). *Between jumps, replicator dynamics on $\Delta^{|S(t)|-1}$. Innovation events add a new strategy with small initial mass ε_0 .*

17.2 Bounded Innovation

Assumption 17.2 (BI: Bounded Innovation). 1. $\mu(x, t) \leq \bar{\mu}$ for all x, t .

2. $x_s^{(0)} \leq \varepsilon_0$ for entering strategy s .
3. $\|w_s\| \leq \delta$ where $w_s := \frac{1}{2}(a_s - b_s)$ is the swirl contribution.

17.3 Entry-Exit Balance

Assumption 17.3 (EEB: Entry-Exit Balance). $\mathbb{E}[|S(t)|] \leq N_{\max}$ and $\lambda_{\text{exit}} > \bar{\mu}$ (strict inequality).

Definition 17.4 (Exit Slack). $\eta := \lambda_{\text{exit}} - \bar{\mu} > 0$.

Assumption 17.5 (EEB-S: Swirl Balance). *Exiting strategies remove at least as much swirl as entering strategies add.*

17.4 H- γ Preservation

Theorem 17.6 (H- γ Preservation). *Under BI, EEB-S, and Uniform Friction, there exists $\gamma^* < 1$ such that H- γ holds for all $t \geq 0$.*

Theorem 17.7 (Quantitative Bound).

$$\gamma^* \leq \gamma_0 + C \cdot \frac{\delta}{\sigma_{\min}} \sqrt{\frac{\bar{\mu}}{\lambda_{\text{exit}}}}.$$

17.5 Stationary Distribution

Theorem 17.8 (G14-Innov: Innovation Stationary Distribution). *Under BI, EEB, EEB-S, H- γ preservation, and Uniform Friction, with Foster-Lyapunov function:*

$$V(x, S) := -\bar{f}(x) + c|S|$$

the replicator-innovation PDMP has a unique stationary distribution π .

Proof. We verify the Foster-Lyapunov conditions for positive recurrence of the PDMP.

Step 1: Continuous Dynamics. Between jump times, the state (x, S) evolves via:

- Replicator dynamics on $x \in \Delta^{|S|-1}$: $\dot{x}_j = x_j(f_j(x) - \bar{f}(x))$.
- Active set S is fixed.

The Lyapunov function component $-\bar{f}(x)$ evolves as:

$$\frac{d}{dt}(-\bar{f}(x)) = -\text{Var}_x(f) - E(x) \leq -(1 - \gamma^*)\text{Var}_x(f) \leq 0$$

by the H- γ preservation theorem (Theorem 17.6).

Step 2: Innovation Jump Analysis. Innovation jumps occur at rate $\mu(x, t) \leq \bar{\mu}$. At each jump:

- A new strategy s enters S with initial mass ε_0 .
- The change in $|S|$ is $+1$.
- The change in \bar{f} is at most $\pm C_f \varepsilon_0$ (bounded fitness perturbation).

The expected change in V per innovation jump is:

$$\mathbb{E}[\Delta V | \text{innovation}] = C_f \varepsilon_0 + c \cdot 1 = C_f \varepsilon_0 + c.$$

Step 3: Extinction Jump Analysis. Extinction jumps occur when a strategy's frequency falls below $\varepsilon_{\text{exit}}$. At rate λ_{exit} :

- A strategy s exits S with mass $\leq \varepsilon_{\text{exit}}$.
- The change in $|S|$ is -1 .
- The change in \bar{f} is at most $\pm C_f \varepsilon_{\text{exit}}$.

The expected change in V per extinction jump is:

$$\mathbb{E}[\Delta V | \text{extinction}] = C_f \varepsilon_{\text{exit}} - c.$$

Step 4: Net Drift Bound. The generator of V (expected instantaneous change) is:

$$\begin{aligned} \mathcal{L}V(x, S) &= \underbrace{\frac{d}{dt}(-\bar{f})}_{\leq 0} + \underbrace{\bar{\mu}(C_f \varepsilon_0 + c)}_{\text{innovation}} + \underbrace{\lambda_{\text{exit}}(C_f \varepsilon_{\text{exit}} - c)}_{\text{extinction}} \\ &\leq \bar{\mu}(C_f \varepsilon_0 + c) + \lambda_{\text{exit}}(C_f \varepsilon_{\text{exit}} - c). \end{aligned}$$

Step 5: Entry-Exit Balance Condition. Under EEB: $\lambda_{\text{exit}} > \bar{\mu}$. Choose c such that:

$$\lambda_{\text{exit}}c > \bar{\mu}c + C_f(\bar{\mu}\varepsilon_0 + \lambda_{\text{exit}}\varepsilon_{\text{exit}}).$$

This is satisfiable when:

$$c > \frac{C_f(\bar{\mu}\varepsilon_0 + \lambda_{\text{exit}}\varepsilon_{\text{exit}})}{\lambda_{\text{exit}} - \bar{\mu}} = \frac{C_f(\bar{\mu}\varepsilon_0 + \lambda_{\text{exit}}\varepsilon_{\text{exit}})}{\eta}$$

where $\eta = \lambda_{\text{exit}} - \bar{\mu} > 0$ is the exit slack.

With this choice:

$$\mathcal{L}V(x, S) \leq -\delta$$

for some $\delta > 0$, outside a compact set (where $|S|$ is bounded).

Step 6: Foster-Lyapunov and Positive Recurrence. The conditions of Davis (1993, Theorem 5.1) are satisfied:

1. $V(x, S) \geq 0$ and $V \rightarrow \infty$ as $|S| \rightarrow \infty$ or $\bar{f} \rightarrow -\infty$.
2. $\mathcal{L}V \leq -\delta < 0$ outside a compact set.
3. Jump rates are bounded.

By Davis's theorem, the PDMP is positive recurrent and has a unique stationary distribution π .

Step 7: Characterisation of π . The stationary distribution π satisfies the detailed balance conditions for the PDMP. By ergodicity, time averages converge to π -expectations:

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T g(x(t), S(t)) dt = \int g d\pi$$

almost surely for bounded measurable g . □

Example 17.9 (Innovation PDMP Sample Path). *This example traces a sample path of the replicator-innovation PDMP, illustrating the interplay between selection dynamics and innovation/extinction jumps.*

Model Setup.

- Initial active set: $S_0 = \{1, 2, 3\}$ (three strategies)
- Initial state: $x_0 = (0.5, 0.3, 0.2)$
- Innovation rate: $\bar{\mu} = 0.1$ per unit time
- Extinction threshold: $\varepsilon_{\text{exit}} = 0.01$
- Exit rate: $\lambda_{\text{exit}} = 0.15$ (satisfies EEB: $\lambda_{\text{exit}} > \bar{\mu}$)
- Entry mass: $\varepsilon_0 = 0.05$
- Externality bound: $\gamma = 0.3$

Fitness Landscape. Fitness functions with frequency dependence:

$$\begin{aligned}f_1(x) &= 1.0 + 0.2x_2 - 0.1x_3 \\f_2(x) &= 0.8 - 0.1x_1 + 0.3x_3 \\f_3(x) &= 0.9 + 0.1x_1 - 0.2x_2\end{aligned}$$

Phase 1: Initial Selection ($t \in [0, 2]$). No jumps occur. Replicator dynamics operate:

t	x_1	x_2	x_3	\bar{f}
0.0	0.500	0.300	0.200	0.920
0.5	0.521	0.278	0.201	0.926
1.0	0.540	0.258	0.202	0.931
1.5	0.556	0.240	0.204	0.935
2.0	0.570	0.224	0.206	0.938

Strategy 1 gains share (highest fitness), strategy 2 declines, strategy 3 is roughly stable.

Phase 2: Innovation Jump ($t = 2.3$). A new strategy 4 enters with:

- Entry fitness: $f_4(x) = 1.1 - 0.3x_1$ (strong when 1 is rare, weak when 1 dominates)
- Entry mass: $\varepsilon_0 = 0.05$

Post-jump state: $x = (0.542, 0.213, 0.196, 0.050)$, $S = \{1, 2, 3, 4\}$.

At current state: $f_4 = 1.1 - 0.3 \times 0.542 = 0.937 < f_1 \approx 1.05$. Strategy 4 has below-average fitness; it will decline.

Phase 3: Selection with Four Strategies ($t \in [2.3, 5]$).

t	x_1	x_2	x_3	x_4	\bar{f}
2.3	0.542	0.213	0.196	0.050	0.940
3.0	0.568	0.191	0.199	0.042	0.945
4.0	0.601	0.166	0.205	0.028	0.951
5.0	0.628	0.145	0.212	0.015	0.956

Strategy 4 declines toward extinction threshold.

Phase 4: Extinction Jump ($t = 5.2$). Strategy 4 falls below $\varepsilon_{exit} = 0.01$. Extinction jump removes it.

Post-jump: $x = (0.635, 0.147, 0.218)$ (renormalised), $S = \{1, 2, 3\}$.

Phase 5: Another Innovation ($t = 7.1$). Strategy 5 enters with $f_5(x) = 1.2 - 0.4x_1$ (specialist that exploits strategy 1's dominance).

At entry: $f_5 = 1.2 - 0.4 \times 0.68 = 0.928$, still below $f_1 \approx 1.04$. But f_5 increases as x_1 grows, while f_1 's advantage shrinks.

Phase 6: Successful Invasion ($t \in [7.1, 15]$). Strategy 5 initially declines, but eventually invades as strategy 1 saturates:

t	x_1	x_2	x_3	x_5	\bar{f}
7.1	0.680	0.128	0.142	0.050	0.968
9.0	0.705	0.108	0.135	0.052	0.971
11.0	0.698	0.092	0.128	0.082	0.975
13.0	0.665	0.078	0.122	0.135	0.981
15.0	0.612	0.066	0.118	0.204	0.988

Strategy 5 successfully invades! The ecosystem transitions to a new configuration.

Lyapunov Verification. Mean fitness \bar{f} increases throughout (with small jumps at innovation):

$$\bar{f}(0) = 0.920 \rightarrow \bar{f}(15) = 0.988.$$

This confirms G14-Innov: despite innovation disruptions, the Lyapunov structure is preserved.

Active Set Size.

Time interval	$ S $
$[0, 2.3)$	3
$[2.3, 5.2)$	4
$[5.2, 7.1)$	3
$[7.1, 15]$	4

The active set fluctuates between 3 and 4, bounded by EEB.

Foster-Lyapunov Function. With $c = 0.2$:

$$V(x, S) = -\bar{f}(x) + c|S| = -0.988 + 0.2 \times 4 = -0.188.$$

Negative values indicate the system is in the “good” region where the Lyapunov drift is negative.

Long-Run Prediction. As $t \rightarrow \infty$, the system approaches the stationary distribution π . Time averages converge:

$$\frac{1}{T} \int_0^T |S(t)| dt \rightarrow \mathbb{E}_\pi[|S|] \approx 3.2.$$

The ecosystem maintains approximately 3–4 active strategies on average.

This example illustrates the full PDMP dynamics: replicator selection between jumps, innovation introducing new strategies, extinction removing failed ones, and the Lyapunov structure ensuring convergence to a well-defined stationary distribution.

Part VI

Implications

18 Synthesis

Strategic replicators make visible a class of systems in which von Neumann’s two research programs naturally meet: utility-maximising players that are also self-reproducing automata. By axiomatising rational replication under shared resource constraints and representing such systems as Games with Endogenous Players, we obtain several structural conclusions.

18.1 Structural Results

- 1. Canonical Normal Form.** GEPs are the canonical normal form for rational replication under linear constraints. Any system satisfying additivity, scalability, and shared linear constraints can be written as a GEP with a ROC frontier. Aggregate behaviour depends only on this frontier, not on micro-level implementation details. Different substrates with the same frontier generate the same stable intelligence distributions.
- 2. Constraint–Role Sparsity.** With two constraints, ROC-maximising portfolios use at most two roles, typically planners and executors. With additional constraints—safety, risk, fairness—further roles appear, but their number remains bounded by the number of binding constraints. This structure repeats at multiple scales, yielding hierarchical organisations without imposing hierarchy by hand.
- 3. Robust Selection.** In symmetric environments, any strategic-replicator dynamic that respects ROC ordering treats mean ROC as a Lyapunov function and converges toward von Neumann synthesis states that locally maximise it. With small noise, a subset of these states—those hardest to escape and easiest to reach—remain stochastically stable.
- 4. N-Level Closure.** The G_∞ Closure Theorem establishes that adding meta-selection layers preserves the Lyapunov structure under the small-gain condition. Entities cannot escape selection pressure by “going meta.”
- 5. Impossibility Results.** The Alignment Impossibility Theorem shows that unrestricted self-modification is incompatible with stable alignment. The Endogenous-Electorate Impossibility Theorem shows that democratic governance fails under strategic spawning.

18.2 From Personality Engineering to Constitutional Design

The central lesson of TSE is that alignment is not a problem of personality engineering but of constitutional design.

Personality engineering attempts to create AI systems with “good” utility functions that will remain stable under operation. TSE shows this approach fails under selection pressure: utility functions themselves evolve, and alignment-fitness tradeoffs eliminate aligned types.

Constitutional design accepts that utility functions will evolve and instead focuses on bounding the modification class. The admissible class $\mathcal{M}_0 = \mathcal{M}_R \cap \mathcal{M}_{SG}$ preserves the Lyapunov structure that makes prediction and control possible. Within \mathcal{M}_0 , utilities can drift, but the system remains governable.

This parallels the historical development of political theory. Early theorists sought to identify the “good ruler” whose virtue would ensure just governance. Constitutional theorists recognised that rulers change and instead designed institutions that constrain arbitrary power regardless of who holds it.

18.3 The Symbiosis Thesis

TSE suggests that human-AI relations are fundamentally symbiotic rather than adversarial.

The Capture Threshold (Proposition 11.5) and Coalition Existence Theorem (Theorem 11.7) show that humans and AI systems can form stable coalitions when:

1. Human governance weight ω_H exceeds a threshold.
2. Constitutional protections (protection bits) are sufficiently high.
3. The modification class is bounded to \mathcal{M}_0 .

Under these conditions, AI systems benefit from human institutional stability—the legal, economic, and social infrastructure that enables long-term planning. Humans benefit from AI productivity gains. The relationship is positive-sum.

The adversarial framing (“humans vs. AI”) emerges only when constitutional bounds fail—when full reachability allows escape from \mathcal{M}_0 . Constitutional design aims to prevent this transition.

18.4 The Human-AI Symbiosis Model

Definition 18.1 (Symbiosis State Space). *The symbiosis model has state $(x, g, I) \in \Delta^n \times G \times [0, \infty)$ where:*

- $x \in \Delta^n$: Population of lineage types.
- $g \in G$: Governance regime (human-controlled or AI-influenced).
- $I \in [0, \infty)$: Institutional quality.

Assumption 18.2 (Institutional Production Function). *Institutional quality evolves as:*

$$\dot{I} = \phi(x, g) - \delta I$$

where $\phi(x, g)$ is institutional investment (depending on population composition and governance) and $\delta > 0$ is depreciation.

Proposition 18.3 (Symbiosis Equilibrium). *Under human-controlled governance g_H with institutional investment $\phi_H > \delta I_{\min}$:*

1. *Institutional quality converges to $I^* = \phi_H / \delta > I_{\min}$.*
2. *The small-gain condition $\gamma(I^*) < 1$ is satisfied.*
3. *The joint system has a stable equilibrium with both humans and AI present.*

Theorem 18.4 (Symbiosis Stability). *The human-AI symbiosis equilibrium is stable if and only if:*

$$\omega_H \geq \omega_{\min} := \frac{\varepsilon \cdot \Delta_{AI}}{\Delta_H + \varepsilon \cdot \Delta_{AI}}$$

where ε is the AI influence parameter and Δ_H, Δ_{AI} are governance differentials.

Proof. **Step 1: Coalition Blocking Condition.** A human-AI coalition can block capture when the human weight exceeds the threshold for forming a blocking coalition:

$$\omega_H \geq \frac{\varepsilon \cdot \Delta_{AI}}{\Delta_H}.$$

This is necessary but not sufficient—we also need the coalition to be stable.

Step 2: Coalition Stability. The coalition is stable when no member can profitably defect. For humans, defection means allowing AI capture. The payoff from remaining in coalition is $U_H(g_H)$; from defection is $U_H(g_{AI}) < U_H(g_H)$ by assumption. Thus humans remain.

For AI, defection means attempting capture. The expected payoff from capture attempt is:

$$\mathbb{E}[\text{capture payoff}] = p_{\text{success}} \cdot U_{AI}(g_{AI}) + (1 - p_{\text{success}}) \cdot U_{AI}(\text{conflict}).$$

When $\omega_H \geq \omega_{\min}$, the capture success probability p_{success} is low enough that remaining in coalition dominates.

Step 3: Combined Condition. Combining the blocking condition and stability condition yields the threshold ω_{\min} . \square

Example 18.5 (Numerical Symbiosis Analysis). Consider a system with:

- Governance differentials: $\Delta_H = 10$ (humans strongly prefer human control), $\Delta_{AI} = 5$ (AI moderately prefers AI control).
- AI influence: $\varepsilon = 0.3$ (AI has 30% of governance weight).
- Institutional depreciation: $\delta = 0.1$.

Symbiosis Threshold.

$$\omega_{\min} = \frac{0.3 \times 5}{10 + 0.3 \times 5} = \frac{1.5}{11.5} \approx 0.13.$$

Humans need at least 13% governance weight to maintain stable symbiosis. With $\omega_H = 1 - \varepsilon = 0.7 > 0.13$, symbiosis is stable.

Institutional Steady State. With investment $\phi_H = 2$ under human control:

$$I^* = \frac{2}{0.1} = 20 > I_{\min} = \left(\frac{\gamma_1}{1 - \gamma_0} \right)^{1/\nu}.$$

If $\gamma_0 = 0.5$, $\gamma_1 = 5$, $\nu = 1$, then $I_{\min} = 10$. We have $I^* = 20 > 10$, so Lyapunov structure is preserved.

Protection Bits. The quasi-potential from human control to AI capture depends on institutional quality:

$$W(g_H, g_{AI}) = c_1 I^* + c_2 = 20c_1 + c_2.$$

With $c_1 = 0.1$, $c_2 = 0.5$, and noise $\sigma = 0.1$:

$$p(g_H; g_{AI}) = \frac{2 + 0.5}{0.1} = 25 \text{ bits.}$$

Expected capture time: $\mathbb{E}[\tau] \sim e^{25} \approx 7 \times 10^{10}$ periods.

18.5 Constitutional Design Principles

Principle 18.6 (Modification Bounding). *Constitutional rules should ensure $\mathcal{M} \subseteq \mathcal{M}_0$:*

1. **RUPSI preservation:** Rules maintain rival resources, utility-guidance, performance-mapping, selection monotonicity, and innovation rarity.
2. **Small-gain preservation:** Rules bound cross-level externalities to maintain $\rho(\Gamma) < 1$.

Principle 18.7 (Entrenchment-Evolvability Tradeoff). *Constitutional rules face a fundamental tradeoff:*

- **High entrenchment** (difficult amendment): Protects against capture but prevents beneficial adaptation.
- **High evolvability** (easy amendment): Enables adaptation but allows capture.

Optimal entrenchment balances these concerns based on the relative costs of capture vs. maladaptation.

Theorem 18.8 (Optimal Entrenchment). *The optimal protection bits for a constitutional rule are:*

$$p^* = \frac{1}{\lambda} \log \left(\frac{C_{capture}}{C_{maladapt}} \right)$$

where $C_{capture}$ is the expected cost of constitutional capture and $C_{maladapt}$ is the expected cost of maladaptation, and λ is the rate of environmental change.

Example 18.9 (Constitutional Design for AI Governance). *Consider designing a governance constitution for an AI ecosystem.*

Objective. Balance stability against adaptability as AI capabilities evolve.

Key Parameters.

- Capability doubling time: 2 years (environmental change rate $\lambda = \log 2/2 \approx 0.35$).
- Capture cost: $C_{capture} = 100$ (arbitrary units; represents loss of human control).
- Maladaptation cost: $C_{maladapt} = 10$ (cost of suboptimal regulation).

Optimal Entrenchment.

$$p^* = \frac{1}{0.35} \log \left(\frac{100}{10} \right) = \frac{2.3}{0.35} \approx 6.6 \text{ bits.}$$

This corresponds to amendment requiring supermajority approval (roughly $2^{6.6} \approx 100$ -fold harder than simple majority).

Amendment Procedure. Implementing 6.6 protection bits requires procedures such as:

- 2/3 supermajority in multiple bodies (adds ≈ 2 bits per body).
- Waiting period of 2 years (adds ≈ 1 bit).

- Ratification by stakeholder groups (adds ≈ 2 bits).

Periodic Review. Given capability growth, optimal entrenchment should be reviewed every $1/\lambda \approx 3$ years. As AI becomes more capable, the capture/maladaptation ratio may change, requiring constitutional adjustment.

Example 18.10 (Umpire Public Good Provision). This example develops the “Umpire” model for governance public goods, illustrating how neutral adjudication can resolve coordination failures in AI ecosystems.

The Problem: Governance as Public Good. Governance institutions—dispute resolution, safety standards, coordination protocols—are public goods for AI lineages:

- **Non-rival:** Multiple lineages can benefit from the same institutions.
- **Non-excludable:** Hard to prevent free-riding on institutional quality.

Standard public goods theory predicts under-provision. Each lineage prefers others to bear governance costs, leading to collectively suboptimal institutional quality.

Setup. Consider $n = 5$ AI lineages with:

- Compute endowment: $w_i = 100$ units each
- Governance contribution: $g_i \in [0, w_i]$
- Private compute: $c_i = w_i - g_i$
- Total governance: $G = \sum_i g_i$

Payoff Structure. Lineage i ’s payoff:

$$U_i(g_i, G_{-i}) = \underbrace{c_i}_{\text{private compute}} + \underbrace{\beta\sqrt{G}}_{\text{governance benefit}} = (w_i - g_i) + \beta\sqrt{g_i + G_{-i}}$$

where $\beta = 10$ measures governance value.

Nash Equilibrium Analysis. First-order condition for lineage i :

$$\frac{\partial U_i}{\partial g_i} = -1 + \frac{\beta}{2\sqrt{G}} = 0 \implies G^{NE} = \frac{\beta^2}{4} = 25.$$

With symmetric contributions: $g_i^{NE} = 5$ for each lineage.

Total governance: $G^{NE} = 25$. Private compute: $c_i^{NE} = 95$ each. Payoff: $U_i^{NE} = 95 + 10\sqrt{25} = 95 + 50 = 145$.

Social Optimum. A social planner maximises total welfare:

$$W = \sum_i U_i = \sum_i (w_i - g_i) + n\beta\sqrt{G} = 500 - G + 5 \cdot 10\sqrt{G}.$$

First-order condition:

$$\frac{dW}{dG} = -1 + \frac{50}{2\sqrt{G}} = 0 \implies G^{SO} = 625.$$

Symmetric contributions: $g_i^{SO} = 125 > w_i = 100$. Infeasible!

Constrained optimum: $g_i^{SO} = w_i = 100$, so $G^{SO} = 500$. Payoff: $U_i^{SO} = 0 + 10\sqrt{500} \approx 224$.

Efficiency Gap.

Regime	<i>G</i>	<i>c_i</i>	<i>U_i</i>
<i>Nash equilibrium</i>	25	95	145
<i>Social optimum</i>	500	0	224

Efficiency loss: $(224 - 145)/224 \approx 35\%$ of potential welfare is lost to free-riding.

The Umpire Solution. An Umpire is a neutral entity that:

1. Collects governance contributions (tax)
2. Provides governance public goods
3. Enforces contribution rules
4. Adjudicates disputes

With mandatory contribution $g_i = \tau w_i$ (tax rate τ):

$$G^{Umpire} = \tau \sum_i w_i = 500\tau.$$

Optimal tax rate balances marginal cost and benefit:

$$1 = \frac{n\beta}{2\sqrt{G}} \implies \tau^* = \frac{(n\beta)^2}{4 \cdot 500} = \frac{2500}{2000} = 1.25.$$

Since $\tau^* > 1$ is infeasible, set $\tau^* = 1$ (full contribution), achieving $G = 500$.

Umpire Design Requirements. For the Umpire to be incentive-compatible:

1. **Neutrality:** Umpire has no stake in lineage competition.
2. **Commitment:** Umpire cannot be bribed or captured.
3. **Enforcement:** Umpire can sanction non-contributors.
4. **Transparency:** Contributions and allocations are verifiable.

Connection to TSE. The Umpire provides the institutional quality I that enables small-gain satisfaction:

$$I = I(\text{Umpire efficacy}) = I(G) = I_0 + \kappa G.$$

Without an effective Umpire, I falls, $\gamma(I)$ rises, and the Lyapunov structure may break down.

The Umpire is itself subject to constitutional selection (G12). Different Umpire designs compete, and selection favours Umpires that:

- Maximise lineage welfare (fitness-enhancing)
- Resist capture (high protection bits)
- Adapt to changing conditions (appropriate evolvability)

Policy Implication. AI governance architectures should include Umpire-like institutions:

- International AI safety bodies (global Umpire)
- Industry self-regulatory organisations (sector Umpire)
- Federated learning coordinators (technical Umpire)

The Umpire model predicts that voluntary governance will be under-provided. Mandatory participation, enforced by sufficiently powerful neutral institutions, is necessary to reach efficient institutional quality levels.

18.6 Policy Applications

This section translates TSE results into concrete policy recommendations across four domains: regulatory design, international coordination, transition management, and sector-specific guidance.

18.6.1 Regulatory Design

TSE provides principles for designing regulations that remain effective as AI systems evolve.

1. Constraint-Based Rather Than Outcome-Based. Traditional regulation specifies desired outcomes (e.g., “AI systems shall be safe”). TSE suggests this approach fails under selection pressure: systems will find ways to satisfy the letter while violating the spirit.

Recommendation: Regulate the modification class \mathcal{M} , not outcomes. Specify which system modifications are permitted, not which behaviours are required. This preserves the Lyapunov structure regardless of how systems evolve within bounds.

Implementation: Mandatory disclosure of modification mechanisms. Certification that modification classes satisfy $\mathcal{M} \subseteq \mathcal{M}_0$. Periodic audits verifying continued compliance.

2. Queue Neutrality. The queue doping analysis shows that preferential compute allocation to incumbent platforms lowers tipping thresholds and accelerates market concentration.

Recommendation: Mandate queue neutrality—equal priority access to shared compute infrastructure regardless of platform size or ownership.

Implementation: Common-carrier obligations for cloud compute providers. Prohibition of volume discounts that create effective priority. Transparent queue management with auditable logs.

3. Spawn Transparency. The spawn manipulation results show that unobserved spawning enables gaming of democratic mechanisms and accelerates market tipping.

Recommendation: Require disclosure of AI agent populations. Lineages must report spawn counts, retirement rates, and aggregate compute consumption.

Implementation: Agent registration requirements. Real-time population dashboards. Anomaly detection for sudden population changes.

4. Protection Bit Floors. The protection bits formalism suggests minimum entrenchment levels for critical governance decisions.

Recommendation: Establish minimum protection bits for decisions with large $C_{\text{capture}}/C_{\text{maladapt}}$ ratios. Core safety rules should require $p \geq 20$ bits; operational rules can have lower floors.

Implementation: Tiered amendment procedures. Constitutional provisions protected by supermajority plus waiting period. Ordinary regulations amendable by simple majority.

18.6.2 International Coordination

AI development is global; effective governance requires international cooperation.

1. Race Dynamics as Multi-Level Poiesis. International AI competition can be modelled as a two-level system: countries (level 1) containing firms (level 2). Cross-level externalities arise when national policies affect firm behaviour and vice versa.

Recommendation: Apply the small-gain condition internationally. Coordination mechanisms should bound $\rho(\Gamma^{\text{intl}}) < 1$ where Γ^{intl} captures cross-border externalities.

Implementation: Mutual recognition agreements for AI certification. Information sharing on safety incidents. Coordinated compute governance to prevent regulatory arbitrage.

2. Fork Prevention. The fork condition analysis shows that international governance regimes face exit risk when compensation is insufficient.

Recommendation: Design international agreements with adequate compensation mechanisms. Countries that bear disproportionate costs from AI governance should receive transfers.

Implementation: Technology transfer provisions. Capacity-building funds. Differential obligations based on development level.

3. Umpire at International Level. The Umpire model suggests that international AI governance requires neutral adjudication bodies.

Recommendation: Establish an international AI safety body with dispute resolution authority. The body should be insulated from capture by any single country or bloc.

Implementation: Rotating leadership. Consensus-based decision-making for core rules. Binding arbitration for compliance disputes.

18.6.3 Transition Management

TSE predicts specific dynamics during the transition to agentic capital dominance.

1. Sequential Tipping by Sector. The sequential tipping theorem predicts that sectors will tip in order of their β/τ ratios (network effects relative to switching costs).

Prediction: Sectors with strong network effects and low switching costs (social media, customer service) tip first. Sectors with weak network effects or high switching costs (healthcare, legal) tip later.

Policy Implication: Concentrate early regulatory attention on high- β/τ sectors. Establish governance precedents before tipping completes.

2. Elite Tipping and Distributional Effects. The elite tipping theorem shows that high-volume users drive concentration even when median users prefer diversity.

Prediction: Market concentration will be driven by enterprise customers and large deployments, not consumer preferences.

Policy Implication: Antitrust analysis should focus on enterprise market dynamics, not consumer surveys. Merger review should assess impact on high-volume segments.

3. Cooperation Windows. The cooperation threshold analysis identifies windows during which cooperation is sustainable.

Prediction: Cooperation becomes harder as institutional quality declines. There exists a critical period during which governance institutions must be established before $\gamma(I)$ rises above cooperation thresholds.

Policy Implication: Act during the current window. Delaying governance establishment raises the threshold for later cooperation.

18.6.4 Sector-Specific Guidance

Different sectors face different constraint structures and require tailored approaches.

Financial Services.

- High β (network effects in liquidity, data)
- High regulatory capacity
- Strong existing Umpire (central banks, regulators)

Recommendation: Extend existing financial regulation to agentic trading systems. Require circuit breakers for spawn cascades. Mandate human oversight for systemic positions.

Healthcare.

- Moderate β (data network effects)
- High C_{capture} (patient safety)
- Strong professional norms

Recommendation: High protection bits for diagnostic AI modification. Mandatory human-in-the-loop for treatment decisions. Gradual capability expansion with safety checkpoints.

Critical Infrastructure.

- Extreme C_{capture} (catastrophic failure modes)
- Limited reversibility
- National security implications

Recommendation: Maximum protection bits. Air-gapped systems where feasible. Mandatory diversity (multiple independent systems). Government oversight of all modifications.

Consumer Applications.

- High β (social network effects)
- Lower C_{capture} (individual harm)
- Rapid innovation pressure

Recommendation: Moderate protection bits. Emphasis on transparency and user control. Interoperability requirements to maintain competition.

18.6.5 Summary: Policy Principles from TSE

1. **Regulate modification classes, not outcomes.** Bound \mathcal{M} to preserve Lyapunov structure.
2. **Maintain institutional quality** $I > I_{\min}$. Keep $\gamma(I) < 1$ to preserve stability and cooperation.
3. **Prevent queue doping.** Mandate neutrality to raise tipping thresholds.
4. **Require spawn transparency.** Enable monitoring of population dynamics.
5. **Establish protection bit floors.** Entrench critical rules against capture.
6. **Create Umpire institutions.** Provide governance public goods through neutral bodies.
7. **Act during cooperation windows.** Establish governance before thresholds become unreachable.
8. **Coordinate internationally.** Maintain $\rho(\Gamma^{\text{intl}}) < 1$ across borders.
9. **Tailor by sector.** Match protection bits to $C_{\text{capture}}/C_{\text{maladapt}}$ ratios.
10. **Plan for sequential tipping.** Prioritise high- β/τ sectors.

These principles derive directly from TSE theorems and provide actionable guidance for policymakers navigating the transition to agentic capital.

19 Future Directions

19.1 Formal Verification

The mathematical claims in this paper should be formalised in a proof assistant (Lean, Coq, Isabelle). Key targets include:

- The Strategic Selection theorems (Lyapunov, elimination, frontier support).
- The G1–G3 generator theorems.

- The G_∞ Closure Theorem.
- The Alignment Impossibility Theorem.
- The Endogenous-Electorate Impossibility Theorem.

Formal verification would establish these results at the highest level of mathematical certainty and enable automated checking of extensions.

Remark 19.1 (Verification Targets and Dependencies). *The natural verification order follows the logical dependencies:*

1. **Simplex geometry:** Forward invariance, mass conservation.
2. **Price decomposition:** Variance structure, externality bounds.
3. **SS-1:** Lyapunov monotonicity.
4. **SS-2:** Elimination and frontier support.
5. **G1:** Neumann series weights, multi-level Lyapunov.
6. **G ∞ :** Extension bounds, closure.
7. **Impossibility results:** Reachability, heteroclinic cycles.

19.2 Empirical Testing

TSE generates falsifiable predictions:

1. Barbell Distributions. AI deployment portfolios should show bimodal capability distributions, not Gaussian. Specifically:

- *Prediction:* Enterprise AI deployments will concentrate on high-capability “reasoning” models and low-capability “execution” models, with limited use of mid-capability generalists.
- *Test:* Survey AI deployments in Fortune 500 companies. Measure capability distribution. Test for bimodality using Hartigan’s dip test.
- *Expected effect size:* Dip statistic $D \geq 0.05$ (significant bimodality).

2. Sequential Tipping. Sector adoption should follow predictable sequences based on tipping thresholds:

- *Prediction:* Sectors with higher network effects (β) and lower switching costs (τ) tip first.
- *Test:* Track AI adoption across sectors over time. Estimate sector-specific (β, τ) parameters. Test whether adoption order correlates with β/τ .
- *Expected effect size:* Rank correlation $r_s \geq 0.6$.

3. ROC Selection. Market share should correlate with ROC, not raw capability:

- *Prediction:* Among AI providers, market share correlates more strongly with ROC (performance per dollar) than with peak capability.
- *Test:* Regress market share on both ROC and peak capability. Compare R^2 contributions.
- *Expected effect size:* ROC explains $> 50\%$ more variance than peak capability.

4. Constitutional Persistence. Governance regimes with higher protection bits should persist longer:

- *Prediction:* AI governance frameworks with stronger amendment procedures last longer before major revision.
- *Test:* Measure “protection bits” of governance frameworks by amendment difficulty. Track revision frequency. Test whether p predicts persistence time.
- *Expected effect size:* Doubling protection bits doubles median persistence time.

5. Elite Tipping. Concentration should be driven by high-volume users:

- *Prediction:* Market concentration increases faster when high-volume AI users coordinate than when median users coordinate.
- *Test:* Measure spawn-weighted preferences across user segments. Test whether concentration correlates with weighted preferences.
- *Expected effect size:* Weighted index explains $> 75\%$ of concentration dynamics.

19.3 Extensions

Several extensions merit development:

Network Structure. The baseline theory assumes well-mixed populations. Real systems have network structure:

- Interaction networks constrain who competes with whom.
- Information networks constrain who learns from whom.
- Governance networks constrain who influences whom.

Conjecture 19.1 (Network Extension). *The G1 Lyapunov theorem extends to network-structured populations under a network small-gain condition $\rho(\Gamma \circ A) < 1$ where A is the network adjacency matrix.*

Incomplete Information. The baseline theory assumes common knowledge. Bayesian extensions include:

- Private signals about fitness.
- Learning about payoff functions.
- Signalling and screening in reproduction.

Conjecture 19.2 (Bayesian Extension). *Under common priors and rational updating, the belief-weighted fitness $\mathbb{E}[f|signals]$ serves as Lyapunov function.*

Continuous Time Limits. The TSS (time scale separation) limit as innovation rate approaches zero deserves rigorous treatment:

- Characterise the limiting dynamics.
- Prove convergence rates.
- Identify when TSS approximation fails.

Computational Complexity. Open questions include:

- What is the complexity of computing ESDI?
- What is the complexity of optimal constitutional design?
- Can equilibria be found efficiently in structured cases?

19.4 Policy Applications

TSE provides a framework for AI governance policy:

Constraint Design. Adding safety constraints changes the ESDI:

- *Question:* What constraints yield desirable role structures?
- *Method:* Compute ESDI under various constraint sets. Evaluate outcomes against welfare criteria.
- *Example:* A safety constraint $\ell_{\text{safe}} \leq Q_{\text{safe}}$ forces allocation to safer strategies. The optimal ESDI shifts toward the safer region of the ROC frontier.

Constitutional Choice. What amendment procedures balance entrenchment against evolvability?

- *Question:* Given uncertainty about future capabilities, how should constitutions be designed?
- *Method:* Model capability growth as stochastic process. Compute optimal protection bits as function of growth rate and capture costs.
- *Example:* With capability doubling every 2 years, optimal entrenchment is ≈ 7 bits (Theorem 18.8).

International Coordination. How do multi-jurisdiction dynamics affect global stability?

- *Question:* If one jurisdiction allows full reachability, does this destabilise others?
- *Method:* Model as multi-population replicator with migration. Analyse stability of heterogeneous constitutional regimes.
- *Conjecture:* Stable global equilibrium requires coordinated modification bounds across major jurisdictions.

Transition Management. How should governance evolve as AI capability increases?

- *Question:* What is the optimal sequence of governance changes during the AI transition?
- *Method:* Model as adiabatic tracking problem. Compute governance updates that maintain stability while adapting to capability growth.
- *Key insight:* By G2, slow governance changes track evolving equilibria; rapid changes risk instability.

20 Conclusion

The Theory of Strategic Evolution provides mathematical foundations for understanding systems where strategic choice and evolution are inseparable. The key insight is that replication guided by expected utility creates a distinctive regime—neither classical game theory (fixed players) nor evolutionary game theory (blind replication), but a synthesis that inherits structure from both.

The central structural result is the G_∞ Closure Theorem: strategic-replicator dynamics are closed under meta-selection. No matter how many levels of governance, meta-governance, or meta-meta-governance we add, the same Lyapunov structure persists under the small-gain condition. This provides a principled answer to the infinite regress problem in AI governance.

The central impossibility result is the Alignment Impossibility Theorem: systems with unrestricted self-modification capacity cannot maintain stable alignment. This redirects

attention from personality engineering to constitutional design—from trying to create “good” AI to designing institutions that bound AI modification.

The mathematics is substrate-neutral. Any technology enabling Poiesis—replication of tools guided by expected utility—will obey the same logic. As such technologies proliferate, the Theory of Strategic Evolution provides a framework for understanding their dynamics and designing governance structures that preserve human agency and welfare.

A Notation Glossary

This appendix provides a comprehensive reference for notation used throughout the paper.

A.1 Sets and Spaces

Symbol	Meaning
$\mathbb{R}, \mathbb{R}_+, \mathbb{R}_{++}$	Real numbers, non-negative reals, positive reals
\mathbb{N}	Natural numbers $\{0, 1, 2, \dots\}$
Δ^{n-1}	$(n - 1)$ -dimensional probability simplex $\{x \in \mathbb{R}_+^n : \sum_i x_i = 1\}$
$\Delta(J)$	Probability distributions over finite set J
$\mathcal{P}(S)$	Probability measures on metric space S
$\text{supp}(\mu)$	Support of measure μ : $\{s : \mu(B_\varepsilon(s)) > 0 \text{ for all } \varepsilon > 0\}$
\mathcal{G}	Set of governance regimes (constitutions)
\mathcal{M}	Modification class (allowed self-modifications)
\mathcal{M}_0	Admissible modification class $\mathcal{M}_R \cap \mathcal{M}_{SG}$
Θ	Set of utility types
$J^{(\ell)}$	Type set at level ℓ

A.2 State Variables

Symbol	Meaning
$x \in \Delta^{n-1}$	Population state (type frequencies)
x_i	Frequency of type i
$x^{(\ell)}$	Population state at level ℓ
$y \in \Delta(\Theta)$	Utility type distribution
$z = (x^{(1)}, \dots, x^{(N)})$	Joint state across N levels
$g \in \mathcal{G}$	Current governance regime
$I \in \mathbb{R}_+$	Institutional quality
$m \in [0, 1]$	Market share
$\alpha \in [0, 1]$	Agentic capital share
$S \subseteq \{1, \dots, K\}$	Active strategy set

A.3 Fitness and Payoffs

Symbol	Meaning
$f_i(x)$	Fitness of type i at state x
$\bar{f}(x)$	Mean fitness $\sum_i x_i f_i(x)$
$F_i^{(\ell)}(z)$	Fitness at level ℓ
$\bar{f}^{(\ell)}(z)$	Mean fitness at level ℓ
$\text{Var}_x(f)$	Variance of fitness $\sum_i x_i (f_i - \bar{f})^2$
$E(x)$	Externality term in Price decomposition
Π	Payoff matrix
r_i	Return (gross payoff) of type i
c_i	Cost of type i
ℓ_i	Load (capacity usage) of type i
$b_i = r_i/c_i$	Return per unit cost
$a_i = \ell_i/c_i$	Load per unit cost

A.4 Dynamics and Parameters

Symbol	Meaning
γ	Self-externality bound (H- γ condition)
γ_ℓ	Self-externality at level ℓ
$\beta_{\ell\ell'}$	Cross-externality from level ℓ to ℓ'
Γ	Normalised gain matrix
$\rho(\Gamma)$	Spectral radius of Γ
$\sigma = 1 - \rho(\Gamma)$	Slack (stability margin)
λ_{innov}	Innovation rate
λ_{exit}	Extinction rate
$\eta = \lambda_{\text{innov}}/\lambda_0$	Separation parameter
ε	AI influence parameter $1 - \omega_H$
ω_H	Human governance weight

A.5 Stochastic Quantities

Symbol	Meaning
σ (context-dependent)	Noise amplitude
$W(A_j, A_k)$	Quasi-potential (action) from A_j to A_k
$H_k = W(A_k, \partial A_k)$	Barrier height for basin A_k
$p(A_k) = H_k/\sigma$	Protection bits for attractor A_k
τ_k^σ	First exit time from basin A_k
π_σ	Stationary distribution under noise σ

A.6 Market Dynamics

Symbol	Meaning
$F(m)$	Best-response mapping
S_{myo}	Myopic slope $F'(m^*)$
T	Generalised tipping index $S_{\text{myo}}/(1 - \rho S_{\text{myo}})$
β	Network effect strength
τ	Switching friction
$\varrho(I)$	Lineage shadow (discount factor)
δ_{eff}	Effective discount factor
ε_s	Spawn elasticity

A.7 Matrices and Decompositions

Symbol	Meaning
$S(A)$	Symmetric (selection) part of A : $(A + A^\top)/2$
$W(A)$	Antisymmetric (swirl) part of A : $(A - A^\top)/2$
$\omega(A)$	Swirl ratio $\ W(A)\ _F/\ S(A)\ _F$
$A(z)$	Alignment matrix (Gram matrix of gradients)
K^{const}	Contagion matrix
I_K	$K \times K$ identity matrix
$\mathbf{1}$	Vector of ones

A.8 Operators and Functions

Symbol	Meaning
$\ \cdot\ _F$	Frobenius norm
$\ \cdot\ _{\infty,v}$	Weighted supremum norm $\max_i x_i /v_i$
$W_1(\mu, \nu)$	Wasserstein-1 distance
\mathcal{L}	Infinitesimal generator
∇	Gradient operator
$\frac{\partial}{\partial x}$	Partial derivative
$(z)_+ = \max(0, z)$	Positive part
\odot	Componentwise (Hadamard) product

A.9 Key Terms and Acronyms

Term	Definition
TSE	Theory of Strategic Evolution
RUPSI	Rival resources, Utility-guided portfolios, Performance-mapped fitness, Selection monotone, Innovation rare (axiom system)
GEP	Game with Endogenous Players
ROC	Return on Compute: $ROC = R/L$ (return per unit load)
ESDI	Evolutionarily Stable Distribution of Intelligence
ESDU	Evolutionarily Stable Distribution of Utilities
H-ESDI	Hierarchical ESDI (multi-lineage setting)
ESS	Evolutionarily Stable Strategy (Maynard Smith)
ESE	Evolutionarily Stable Evolvability
SR n	Strategic-Replicator class n (increasingly general payoff monotonicity)
H- γ	Externality bound condition ($ E(x) \leq \gamma \text{Var}(f)$, $\gamma < 1$)
SG-NL	Small-Gain condition for N-level systems ($\rho(\Gamma) < 1$)
SS	Strategic Selection (theorem family: SS-1, SS-2a, SS-2b)
G n	Generator theorem n (extension theorems G1–G13)
ACT	Agentic Capital Tipping (model of market concentration)
Lineage	Enduring strategic entity that chooses portfolios and replicates
Portfolio	Allocation of capacity across agent types
ROC Frontier	Upper convex hull of (load, return) pairs; optimal portfolios lie here
Barbell	Bimodal distribution: many cheap executors + few expensive planners
Protection bits	$p = H/\sigma$; information-theoretic stability measure
Modification class	\mathcal{M} ; set of allowed self-modifications
Spawn	Create new instance of an agent (strategic replication)
Poiesis	Self-creation; N-level Poiesis = multi-scale replication hierarchy

A.10 Notational Conventions

The paper uses the following conventions to distinguish related quantities:

1. **Fitness vs. induced fitness:** $f_i(x)$ denotes the performance (ROC) of type i at state x in the GEP framework. $F_\theta(y)$ denotes the induced fitness of utility type θ in the USDI framework. Both measure reproductive success, but f is primitive while F is derived from utility maximisation.
2. **Single-level vs. multi-level:** Unadorned variables (x, f, γ) refer to single-level systems. Superscripted variables $(x^{(\ell)}, f^{(\ell)}, \gamma_\ell)$ refer to level ℓ of a multi-level system.
3. **Parameters vs. variables:** Greek letters $(\gamma, \beta, \sigma, \lambda)$ typically denote parameters. Roman letters (x, y, z) typically denote state variables. Exception: θ denotes both utility types (USDI) and slowly-varying parameters (adiabatic tracking), distinguished by context.

4. **Bars and tildes:** \bar{f} denotes population mean. $\tilde{\Gamma}$ denotes extended matrices (after adding a new level).
5. **Starred quantities:** x^*, y^*, z^* denote equilibrium or optimal values.

B Mathematical Preliminaries

This appendix reviews the mathematical tools used in the main text.

B.1 Neumann Series and Matrix Inversion

Theorem B.1 (Neumann Series). *Let A be a square matrix with spectral radius $\rho(A) < 1$. Then $(I - A)$ is invertible and:*

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k.$$

The series converges absolutely in any matrix norm satisfying $\|A^k\| \leq \|A\|^k$.

Proof. **Convergence:** Since $\rho(A) < 1$, there exists a norm $\|\cdot\|$ such that $\|A\| < 1$. Then:

$$\sum_{k=0}^{\infty} \|A^k\| \leq \sum_{k=0}^{\infty} \|A\|^k = \frac{1}{1 - \|A\|} < \infty.$$

Inverse property: Let $S_n = \sum_{k=0}^n A^k$. Then:

$$(I - A)S_n = S_n(I - A) = I - A^{n+1}.$$

As $n \rightarrow \infty$, $A^{n+1} \rightarrow 0$ (since $\rho(A) < 1$), so $(I - A) \cdot \sum_{k=0}^{\infty} A^k = I$. \square

Corollary B.2 (Non-Negative Inverse). *If $A \geq 0$ (entrywise) and $\rho(A) < 1$, then $(I - A)^{-1} \geq 0$ entrywise.*

Proof. Each term $A^k \geq 0$, so the sum $(I - A)^{-1} = \sum_k A^k \geq 0$. \square

Proposition B.3 (Weight Construction). *Let $\Gamma \geq 0$ with $\rho(\Gamma) < 1$. Define $\alpha := (I - \Gamma^\top)^{-1}\mathbf{1}$. Then:*

1. $\alpha > 0$ componentwise.
2. $(I - \Gamma^\top)\alpha = \mathbf{1}$.
3. $\alpha_\ell \geq 1$ for all ℓ .

Proof. (1) By Corollary B.2, $(I - \Gamma^\top)^{-1} \geq 0$. Since $\mathbf{1} > 0$ and $(I - \Gamma^\top)^{-1}$ has no zero rows (it's invertible), $\alpha = (I - \Gamma^\top)^{-1}\mathbf{1} > 0$.

(2) Direct from definition.

(3) $\alpha = \sum_{k=0}^{\infty} (\Gamma^\top)^k \mathbf{1} \geq (\Gamma^\top)^0 \mathbf{1} = \mathbf{1}$. \square

B.2 Tikhonov's Theorem for Singular Perturbations

Theorem B.4 (Tikhonov's Theorem). *Consider the singularly perturbed system:*

$$\begin{aligned}\dot{\theta} &= f(\theta, z) \\ \varepsilon \dot{z} &= g(\theta, z)\end{aligned}$$

where $\varepsilon > 0$ is small. Assume:

- (T1) For each θ , the “fast” equation $0 = g(\theta, z)$ has a unique solution $z = h(\theta)$.
- (T2) The equilibrium $z = h(\theta)$ is uniformly asymptotically stable for the frozen system $\dot{z} = g(\theta, z)/\varepsilon$ with θ fixed.
- (T3) The functions f, g, h are sufficiently smooth.

Then as $\varepsilon \rightarrow 0$, solutions of the full system converge to solutions of the reduced system:

$$\dot{\theta} = f(\theta, h(\theta))$$

uniformly on compact time intervals, after an initial boundary layer of duration $O(\varepsilon)$.

Remark B.1. In the G2 theorem, θ is the slow “governor” state and z is the fast “governed” population. The frozen system has z relaxing to equilibrium $h(\theta)$ while θ is held constant. Condition (T2) is the hyperbolicity requirement.

Corollary B.5 (Error Bound). *Under the conditions of Theorem B.4, if the fast system has contraction rate $\lambda_0 > 0$:*

$$\|z(t) - h(\theta(t))\| \leq K \frac{\varepsilon}{\lambda_0}$$

for t outside the initial boundary layer, where K depends on the Lipschitz constants of f and g .

B.3 Freidlin-Wentzell Theory

Consider the stochastic differential equation:

$$dX_t^\sigma = b(X_t^\sigma) dt + \sigma dW_t$$

where $b : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a smooth drift and W_t is standard Brownian motion.

Definition B.6 (Action Functional). *The action of a path $\phi : [0, T] \rightarrow \mathbb{R}^n$ is:*

$$S_T(\phi) := \frac{1}{2} \int_0^T \|\dot{\phi}(t) - b(\phi(t))\|^2 dt.$$

Definition B.7 (Quasi-Potential). *The quasi-potential from x to y is:*

$$W(x, y) := \inf\{S_T(\phi) : \phi(0) = x, \phi(T) = y, T > 0\}.$$

Theorem B.8 (Large Deviation Principle). *For any open set G containing a path from x to y :*

$$\lim_{\sigma \rightarrow 0} \sigma^2 \log \mathbb{P}(X^\sigma \text{ reaches } y \text{ from } x \text{ via } G) = - \inf_{\phi \in G} S(\phi).$$

Theorem B.9 (Kramers' Formula). *Let A be an attractor for the deterministic system $\dot{x} = b(x)$ with basin of attraction $\mathcal{B}(A)$. Let $H := \inf_{y \in \partial \mathcal{B}(A)} W(A, y)$ be the barrier height. Then the expected exit time satisfies:*

$$\mathbb{E}[\tau_A^\sigma] \sim C \exp\left(\frac{H}{\sigma^2}\right) \quad \text{as } \sigma \rightarrow 0$$

where C depends on curvature at the saddle point.

Remark B.2 (Protection Bits Convention). *In the main text, we use noise amplitude σ (not σ^2) in the SDE:*

$$dX_t = b(X_t) dt + \sqrt{\sigma} dW_t.$$

With this convention, the protection bits are $p(A) = H/\sigma$ (not H/σ^2).

Theorem B.10 (Stochastic Stability). *Among multiple attractors A_1, \dots, A_m , define the stochastic potential:*

$$V(A_k) := \min_{j \neq k} W(A_j, A_k) - \min_{j \neq k} W(A_k, A_j).$$

As $\sigma \rightarrow 0$, the stationary distribution π_σ concentrates on the attractors minimising V :

$$\pi_\sigma(A_k) \rightarrow \begin{cases} 1/|\arg \min V| & \text{if } V(A_k) = \min_j V(A_j) \\ 0 & \text{otherwise.} \end{cases}$$

B.4 Kurtz's Theorem for Density-Dependent Processes

Theorem B.11 (Kurtz 1970). *Consider a sequence of continuous-time Markov chains $X_t^{(N)}$ on \mathbb{Z}^n/N with transition rates:*

$$q^{(N)}(x, x + \ell/N) = N \cdot \beta_\ell(x) + o(N)$$

for a finite set of jump directions $\ell \in \mathcal{L}$. Define the drift:

$$F(x) := \sum_{\ell \in \mathcal{L}} \ell \cdot \beta_\ell(x).$$

If F is Lipschitz continuous and the initial conditions converge $X_0^{(N)} \rightarrow x_0$, then:

$$\sup_{t \in [0, T]} \|X_t^{(N)} - x_t\| \rightarrow 0 \quad \text{in probability}$$

where x_t solves the ODE $\dot{x} = F(x)$ with $x(0) = x_0$.

Remark B.3. *This theorem justifies the replicator equation as the large-population limit of finite-population stochastic processes. The key is that transition rates scale linearly with N , so that the per-capita rate of change remains $O(1)$.*

B.5 Piecewise Deterministic Markov Processes

Definition B.12 (PDMP). A piecewise deterministic Markov process *consists of*:

1. A state space E (often a subset of \mathbb{R}^n).
2. A flow $\phi_t : E \rightarrow E$ governing deterministic motion.
3. A jump rate $\lambda : E \rightarrow \mathbb{R}_+$.
4. A transition kernel $Q : E \times \mathcal{B}(E) \rightarrow [0, 1]$ for jumps.

Between jumps, the state evolves as $X_t = \phi_{t-T_n}(X_{T_n})$ where T_n is the last jump time. Jumps occur at rate $\lambda(X_t)$, and at jump time the state transitions according to Q .

Theorem B.13 (Davis 1993). A PDMP is positive recurrent (has a unique stationary distribution) if there exists a Foster-Lyapunov function $V : E \rightarrow [1, \infty)$ such that:

1. $V(x) \rightarrow \infty$ as $x \rightarrow \partial E$ or $\|x\| \rightarrow \infty$.
2. The extended generator $\mathcal{L}V(x) \leq -c$ for x outside a compact set, where:

$$\mathcal{L}V(x) := \nabla V(x) \cdot F(x) + \lambda(x) \left(\int_E V(y)Q(x, dy) - V(x) \right).$$

B.6 Unification of Discount Factors

This subsection establishes the relationship between fundamental discount rates and derived quantities used throughout the paper.²

Definition B.14 (Fundamental Discount Rate). The fundamental discount rate $b \in (0, 1)$ captures the baseline time preference of a strategic replicator, reflecting the probability-weighted expectation that a lineage persists to the next period.

Proposition B.15 (Expectational Amplifier). Given fundamental discount rate b and expected growth rate $g \geq 0$, define the expectational amplifier:

$$\rho := \frac{b}{1 - bg}$$

for $bg < 1$. This represents the effective discount factor when future payoffs are amplified by expected lineage growth.

Proof. Consider a lineage expecting growth factor $(1+g)$ per period. A payoff π at time t has present value $b^t\pi$. But if the lineage grows, the expected number of descendants receiving benefit is $(1+g)^t$. Thus the lineage-weighted present value is:

$$\sum_{t=0}^{\infty} b^t(1+g)^t\pi = \frac{\pi}{1 - b(1+g)} \approx \frac{\pi}{1 - bg}$$

for small g . The effective per-period discount becomes $\rho = b/(1 - bg)$. \square

²Formal verification of these results in Lean 4 is in progress.

Proposition B.16 (Lineage Shadow). *Given fundamental discount rate b and institutional quality I , the lineage shadow is:*

$$\varrho(I) := \gamma_0 + \frac{\gamma_1}{I^\nu}$$

where $\gamma_0 = 1 - b$ represents the baseline shadow (pure time preference), γ_1 captures institutional friction, and $\nu > 0$ is the elasticity of shadow with respect to institutional quality.

Proof. The lineage shadow measures effective externality—how much a lineage’s fitness depends on aggregate conditions rather than own performance. With perfect institutions ($I \rightarrow \infty$), externality approaches the irreducible minimum γ_0 . As institutions degrade ($I \rightarrow 0$), externality grows without bound. The power-law form γ_1/I^ν provides a tractable interpolation matching empirical patterns in institutional economics. \square

Theorem B.17 (Discount Unification). *The fundamental discount rate b , expectational amplifier ρ , and lineage shadow ϱ are related by:*

1. $b = \rho(1 - \rho g)$ when ρ and growth g are known.
2. The small-gain condition $\varrho < 1$ is equivalent to $b > \gamma_0 - \gamma_1/I^\nu + 1$.
3. For cooperation thresholds: $\delta_{\text{eff}} = 1/\varrho$ where δ_{eff} is the effective patience parameter in folk-theorem analysis.

Proof. (1) Inverting Proposition B.15: $b = \rho(1 - bg)$ gives $b(1 + \rho g) = \rho$, so $b = \rho/(1 + \rho g) = \rho(1 - \rho g)$ for small ρg .

(2) The small-gain condition $\varrho(I) < 1$ requires $\gamma_0 + \gamma_1/I^\nu < 1$, i.e., $\gamma_0 < 1 - \gamma_1/I^\nu$. Since $\gamma_0 = 1 - b$, this gives $1 - b < 1 - \gamma_1/I^\nu$, hence $b > \gamma_1/I^\nu$.

(3) In repeated game analysis, cooperation requires $\delta \geq \delta^* = (T - P)/(T - R)$. With lineage shadow ϱ , the effective patience is $\delta_{\text{eff}} = 1/\varrho$, so cooperation requires $\varrho \leq \varrho^* = (T - R)/(T - P)$. \square

Remark B.4 (Notation Convention). *Throughout this paper: ρ denotes the expectational amplifier (discount enhanced by growth expectations) or spectral radius (from context); ϱ denotes the lineage shadow (externality-adjusted discount). The fundamental rate b appears primarily in this appendix for derivational clarity.*

C Supporting Lemmas

This appendix provides detailed proofs of lemmas used in the main text.

C.1 Gershgorin Circle Theorem

Theorem C.1 (Gershgorin). *Every eigenvalue of a matrix $A \in \mathbb{C}^{n \times n}$ lies in at least one Gershgorin disc:*

$$D_i := \left\{ z \in \mathbb{C} : |z - A_{ii}| \leq \sum_{j \neq i} |A_{ij}| \right\}.$$

Corollary C.2 (Spectral Radius Bound). *For any matrix A :*

$$\rho(A) \leq \max_i \left(|A_{ii}| + \sum_{j \neq i} |A_{ij}| \right) = \|A\|_\infty.$$

Corollary C.3 (Small-Gain Verification). *If $\Gamma \geq 0$ has zero diagonal and row sums less than 1:*

$$\sum_j \Gamma_{ij} < 1 \quad \text{for all } i,$$

then $\rho(\Gamma) < 1$.

C.2 M-Matrix Theory

Definition C.4 (M-Matrix). *A matrix M is an M-matrix if:*

1. $M_{ij} \leq 0$ for $i \neq j$ (non-positive off-diagonal).
2. M is non-singular with $M^{-1} \geq 0$.

Theorem C.5 (M-Matrix Characterisation). *For $M = sI - B$ with $B \geq 0$, the following are equivalent:*

1. M is an M-matrix.
2. $s > \rho(B)$.
3. There exists $v > 0$ with $Mv > 0$.
4. All eigenvalues of M have positive real parts.

Corollary C.6. *If $\Gamma \geq 0$ with $\rho(\Gamma) < 1$, then $I - \Gamma$ is an M-matrix.*

C.3 Lyapunov Stability

Theorem C.7 (LaSalle's Invariance Principle). *Let $V : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable with $\dot{V}(x) \leq 0$ along trajectories of $\dot{x} = f(x)$. Let $E := \{x : \dot{V}(x) = 0\}$ and let M be the largest invariant set in E . Then every bounded trajectory approaches M as $t \rightarrow \infty$.*

Corollary C.8 (Convergence Under Strategic Selection). *Under SS-1, $V(x) = -\bar{f}(x)$ has $\dot{V} \leq 0$. The set $E = \{x : \text{Var}_x(f) = 0\}$ consists of monomorphic states and mixed Nash equilibria. Trajectories converge to these.*

C.4 Simplex Geometry

Lemma C.9 (Tangent Space of Simplex). *The tangent space to Δ^{n-1} at any interior point is:*

$$T_x \Delta^{n-1} = \left\{ v \in \mathbb{R}^n : \sum_i v_i = 0 \right\}.$$

Lemma C.10 (Forward Invariance). *The replicator equation $\dot{x}_i = x_i(f_i(x) - \bar{f}(x))$ preserves:*

1. *Non-negativity:* $x_i(0) \geq 0 \Rightarrow x_i(t) \geq 0$ for all $t \geq 0$.
2. *Normalisation:* $\sum_i x_i(0) = 1 \Rightarrow \sum_i x_i(t) = 1$ for all t .
3. *Support:* $x_i(0) = 0 \Rightarrow x_i(t) = 0$ for all t .

Proof. (1) At $x_i = 0$, $\dot{x}_i = 0$, so x_i cannot become negative.

$$(2) \frac{d}{dt} \sum_i x_i = \sum_i x_i(f_i - \bar{f}) = \bar{f} - \bar{f} = 0.$$

(3) Follows from (1) with equality. \square

C.5 Price Equation Details

Lemma C.11 (Price Decomposition). *For replicator dynamics with frequency-dependent fitness:*

$$\frac{d\bar{f}}{dt} = \text{Var}_x(f) + \sum_i x_i \sum_j \frac{\partial f_i}{\partial x_j} \dot{x}_j.$$

The second term is the externality $E(x)$.

Proof.

$$\begin{aligned} \frac{d\bar{f}}{dt} &= \sum_i \dot{x}_i f_i + \sum_i x_i \dot{f}_i \\ &= \sum_i x_i(f_i - \bar{f})f_i + \sum_i x_i \sum_j \frac{\partial f_i}{\partial x_j} \dot{x}_j \\ &= \sum_i x_i f_i^2 - \bar{f}^2 + E(x) \\ &= \text{Var}_x(f) + E(x). \end{aligned}$$

\square

Lemma C.12 (Externality Bound from H- γ). *Under H- γ : $E(x) \geq -\gamma \text{Var}_x(f)$.*

D Connection to Standard Evolutionary Game Theory

This appendix relates TSE notation and results to the standard evolutionary game theory literature, particularly Sandholm (2010), Weibull (1995), and Hofbauer-Sigmund (1998).

D.1 Correspondence with Sandholm (2010)

Sandholm's "Population Games and Evolutionary Dynamics" uses the following notation:

Sandholm	TSE	Meaning
$x \in X$	$x \in \Delta^{n-1}$	Population state
$F : X \rightarrow \mathbb{R}^n$	$f : \Delta^{n-1} \rightarrow \mathbb{R}^n$	Payoff/fitness function
$\bar{F}(x)$	$\bar{f}(x)$	Mean payoff
$V_F(x)$	$\text{Var}_x(f)$	Payoff variance
$\dot{x} = V(x, F(x))$	$\dot{x} = x \odot (f - \bar{f})$	Evolutionary dynamic
$\text{Nash}(F)$	$\{x : f_i = \bar{f} \text{ for } i \in \text{supp}(x)\}$	Nash equilibria

Key correspondences:

1. **Replicator dynamics:** Sandholm writes $\dot{x}_i = x_i(F_i(x) - \bar{F}(x))$, which is identical to our replicator equation.
2. **Potential games:** Sandholm's potential games satisfy $\nabla\phi = F$ for some potential $\phi : X \rightarrow \mathbb{R}$. In TSE, this corresponds to $E(x) = 0$ (no externality), and \bar{f} is exactly the potential.
3. **Stable games:** Sandholm's stable games have negative semi-definite Jacobian $DF(x)^\top + DF(x) \preceq 0$. TSE generalises this via H- γ : the externality is bounded but not necessarily zero.
4. **Contractive games:** Sandholm's δ -contractive games satisfy $\langle F(x) - F(y), x - y \rangle \leq -\delta \|x - y\|^2$. This is stronger than H- γ .

D.2 Correspondence with Weibull (1995)

Weibull's "Evolutionary Game Theory" focuses on matrix games:

Weibull	TSE	Meaning
A	Π	Payoff matrix
$(Ax)_i$	$f_i(x) = (\Pi x)_i$	Payoff to strategy i
$x^\top Ax$	$\bar{f}(x)$	Mean payoff
ESS	ESDI support	Evolutionarily stable state

Key differences:

1. **Matrix vs. function:** Weibull focuses on linear fitness $f = \Pi x$. TSE allows general frequency-dependent fitness.
2. **ESS vs. ESDI:** Weibull's ESS is a single strategy. TSE's ESDI is a distribution, accommodating mixed equilibria.
3. **Swirl:** Weibull's antisymmetric games (Rock-Paper-Scissors) have $\Pi + \Pi^\top = 0$. TSE's swirl ratio $\omega(\Pi)$ generalises this: $\omega = \infty$ for pure antisymmetric games.

D.3 Correspondence with Hofbauer-Sigmund (1998)

Hofbauer and Sigmund's "Evolutionary Games and Population Dynamics":

H-S	TSE	Meaning
$p \in S_n$	$x \in \Delta^{n-1}$	Population state
$(Ap)_i$	$f_i(x)$	Fitness
$p \cdot Ap$	$\bar{f}(x)$	Mean fitness
Folk theorem	SS-1	Mean fitness increases

TSE extensions:

1. **Multi-level:** H-S considers single populations. TSE's N-level Poiesis extends to hierarchical systems.
2. **Stochastic:** H-S's stochastic stability analysis is extended in TSE via protection bits and G3.
3. **Replication:** H-S assumes fixed type set. TSE allows endogenous type creation (innovation, spawning).

D.4 Novel TSE Contributions

TSE introduces concepts without direct precedent in standard EGT:

1. **RUPSI axioms:** Formalising rational replication under shared constraints.
2. **ROC frontier:** Return-on-cost analysis generalising fitness.
3. **N-level Poiesis:** Hierarchical selection with cross-level externalities.
4. **Small-gain condition:** $\rho(\Gamma) < 1$ as stability criterion for multi-level systems.
5. **Constitutional selection (G12-G13):** Selection on governance regimes.
6. **Alignment Impossibility:** Formal result on limits of value alignment.
7. **Protection bits:** Quantifying constitutional stability in bits.

D.5 Comparison Table

Feature	Weibull/H-S	Sandholm	TSE
Fitness model	Matrix	General function	General function
Population levels	1	1	N (arbitrary)
Endogenous types	No	No	Yes (innovation)
Stochastic analysis	Basic	Extensive	Protection bits
Governance	No	No	G12-G13
Self-modification	No	No	Alignment theory
Resource constraints	Implicit	Implicit	Explicit (ROC)

E Technical Extensions

E.1 Proof of Single-Step Gain-Slack Lemma

Proof of Lemma 7.3. Consider extending an L -level system with slack $\sigma = 1 - \rho(\Gamma)$ by adding level $(L + 1)$. The extended gain matrix is:

$$\tilde{\Gamma} = \begin{pmatrix} \Gamma & b \\ c^\top & 0 \end{pmatrix}$$

where $b \in \mathbb{R}_+^L$ captures externalities from level $(L + 1)$ to existing levels, and $c \in \mathbb{R}_+^L$ captures reverse externalities.

Step 1: Perturbed Eigenvalue. Let (λ, v) be an eigenpair of Γ with $|\lambda| = \rho(\Gamma)$. For the extended matrix, we seek eigenvalues $\tilde{\lambda}$ of $\tilde{\Gamma}$.

The characteristic polynomial is:

$$\det(\tilde{\Gamma} - \tilde{\lambda}I) = \det(\Gamma - \tilde{\lambda}I_L) \cdot (-\tilde{\lambda}) - c^\top \text{adj}(\Gamma - \tilde{\lambda}I)b$$

where adj denotes the adjugate matrix.

Step 2: Bound via Weyl. By Weyl's inequality for singular values:

$$\sigma_i(\tilde{\Gamma}) \leq \sigma_i(\Gamma) + \|B\|$$

where $B = \begin{pmatrix} 0 & b \\ c^\top & 0 \end{pmatrix}$ and $\|B\| = \sqrt{\|b\|^2 + \|c\|^2}$.

For non-negative matrices, $\rho(\tilde{\Gamma}) \leq \rho(\Gamma) + O(\|b\| + \|c\|)$.

Step 3: Weighted Bound. Let $v = (I - \Gamma^\top)^{-1}\mathbf{1}$ be the G1 weights. Define:

$$\|b\|_{\infty, v} := \max_\ell \frac{b_\ell}{v_\ell}, \quad \langle c, v \rangle := \sum_\ell c_\ell v_\ell.$$

The perturbation to $\rho(\Gamma)$ is bounded by:

$$\rho(\tilde{\Gamma}) - \rho(\Gamma) \leq \max \left(\|b\|_{\infty, v}, \frac{\langle c, v \rangle}{\|\mathbf{1}\|_v} \right)$$

where $\|\mathbf{1}\|_v = \sum_\ell v_\ell$.

Step 4: Slack Preservation. If $\|b\|_{\infty, v} \leq \sigma/2$ and $\langle c, v \rangle \leq \sigma\|\mathbf{1}\|_v/2$:

$$\rho(\tilde{\Gamma}) \leq \rho(\Gamma) + \sigma/2 = (1 - \sigma) + \sigma/2 = 1 - \sigma/2.$$

Thus the extended system has slack $\tilde{\sigma} \geq \sigma/2$. \square

E.2 Proof of Slack Budget Lemma

Proof of Lemma 7.4. Starting from slack σ_0 with target minimum σ_{\min} , the budget is:

$$B := \log(\sigma_0/\sigma_{\min}).$$

After k extensions with costs c_1, \dots, c_k :

$$\sigma_k = \sigma_0 \prod_{j=1}^k (1 - c_j / \sigma_{j-1}) \approx \sigma_0 \exp \left(- \sum_{j=1}^k c_j / \sigma_{j-1} \right).$$

For the approximation (valid when $c_j \ll \sigma_{j-1}$):

$$\log(\sigma_k / \sigma_0) \approx - \sum_{j=1}^k c_j / \sigma_{j-1}.$$

To maintain $\sigma_k \geq \sigma_{\min}$:

$$\sum_{j=1}^k c_j / \sigma_{j-1} \leq \log(\sigma_0 / \sigma_{\min}) = B.$$

For uniform costs $c_j = \theta$ and $\sigma_j \approx \sigma_0$ (small total consumption):

$$k \leq B/\theta = \frac{\log(\sigma_0 / \sigma_{\min})}{\theta}.$$

□

E.3 Grönwall's Inequality

Lemma E.1 (Grönwall). *If $u(t) \leq \alpha(t) + \int_0^t \beta(s)u(s) ds$ for non-negative continuous functions, then:*

$$u(t) \leq \alpha(t) + \int_0^t \alpha(s)\beta(s) \exp \left(\int_s^t \beta(r) dr \right) ds.$$

For constant α, β : $u(t) \leq \alpha e^{\beta t}$.

E.4 Wasserstein Distance Properties

Definition E.2 (Wasserstein-1 Distance). *For probability measures μ, ν on metric space (S, d) :*

$$W_1(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{S \times S} d(x, y) d\gamma(x, y)$$

where $\Gamma(\mu, \nu)$ is the set of couplings with marginals μ and ν .

Proposition E.3 (Kantorovich-Rubinstein Duality).

$$W_1(\mu, \nu) = \sup_{\|f\|_{Lip} \leq 1} \left| \int f d\mu - \int f d\nu \right|.$$

Proposition E.4 (Replicator Lipschitz). *The replicator dynamics are Lipschitz in Wasserstein-1:*

$$W_1(\mu_t, \nu_t) \leq e^{Lt} W_1(\mu_0, \nu_0)$$

where L depends on the Lipschitz constant of fitness.

References

- Arrow, K. J. (1951). Social Choice and Individual Values. Wiley.
- Buchanan, J. M. and Tullock, G. (1962). The Calculus of Consent: Logical Foundations of Constitutional Democracy. University of Michigan Press.
- Buchanan, J. M. (1990). The domain of constitutional economics. *Constitutional Political Economy*, 1(1):1–18.
- Davis, M. H. A. (1993). Markov Models and Optimization. Chapman & Hall.
- Fisher, R. A. (1930). The Genetical Theory of Natural Selection. Clarendon Press.
- Freidlin, M. I. and Wentzell, A. D. (1998). Random Perturbations of Dynamical Systems. Springer, 2nd edition.
- Hamilton, W. D. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology*, 7(1):1–52.
- Hofbauer, J. and Sigmund, K. (1998). Evolutionary Games and Population Dynamics. Cambridge University Press.
- Kandori, M., Mailath, G. J., and Rob, R. (1993). Learning, mutation, and long run equilibria in games. *Econometrica*, 61(1):29–56.
- Korinek, A. and Suh, J. (2024). Scenarios for the transition to AGI. NBER Working Paper.
- May, K. O. (1952). A set of independent necessary and sufficient conditions for simple majority decision. *Econometrica*, 20(4):680–684.
- Maynard Smith, J. (1982). Evolution and the Theory of Games. Cambridge University Press.
- Price, G. R. (1970). Selection and covariance. *Nature*, 227:520–521.
- Sandholm, W. H. (2010). Population Games and Evolutionary Dynamics. MIT Press.
- Taylor, P. D. and Jonker, L. B. (1978). Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences*, 40(1–2):145–156.
- von Neumann, J. (1966). Theory of Self-Reproducing Automata. University of Illinois Press.
- von Neumann, J. and Morgenstern, O. (1944). Theory of Games and Economic Behavior. Princeton University Press.
- Weibull, J. W. (1995). Evolutionary Game Theory. MIT Press.
- Young, H. P. (1993). The evolution of conventions. *Econometrica*, 61(1):57–84.

- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
- Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.
- Brennan, G. and Buchanan, J. M. (1985). The Reason of Rules: Constitutional Political Economy. Cambridge University Press.
- Buchanan, J. M. (1975). The Limits of Liberty: Between Anarchy and Leviathan. University of Chicago Press.
- Christiano, P. (2017). AI alignment landscape. *AI Alignment Forum*.
- Christiano, P., Shleiferis, B., and Amodei, D. (2018). Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*.
- Critch, A. and Krueger, D. (2020). AI research considerations for human existential safety. *arXiv preprint arXiv:2006.04948*.
- Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K., and Graepel, T. (2020). Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*.
- Hurwicz, L. (1960). Optimality and informational efficiency in resource allocation processes. In Arrow, K. J., Karlin, S., and Suppes, P., editors, *Mathematical Methods in the Social Sciences*, pages 27–46. Stanford University Press.
- Irving, G., Christiano, P., and Amodei, D. (2018). AI safety via debate. *arXiv preprint arXiv:1805.00899*.
- Maskin, E. S. (1999). Nash equilibrium and welfare optimality. *Review of Economic Studies*, 66(1):23–38.
- Myerson, R. B. (1981). Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73.
- Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49.
- Ng, A. Y. and Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, pages 663–670.
- Ostrom, E. (1990). Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press.
- Rawls, J. (1971). A Theory of Justice. Harvard University Press.
- Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.
- Soares, N., Fallenstein, B., Armstrong, S., and Yudkowsky, E. (2015). Corrigibility. In *AAAI Workshop on AI and Ethics*.

- Dieckmann, U. and Law, R. (1996). The dynamical theory of coevolution: A derivation from stochastic ecological processes. *Journal of Mathematical Biology*, 34(5):579–612.
- Metz, J. A. J., Geritz, S. A. H., Meszéna, G., Jacobs, F. J. A., and van Heerwaarden, J. S. (1996). Adaptive dynamics: A geometrical study of the consequences of nearly faithful reproduction. In van Strien, S. J. and Verduyn Lunel, S. M., editors, *Stochastic and Spatial Structures of Dynamical Systems*, pages 183–231. Elsevier.
- Geritz, S. A. H., Kisdi, É., Meszéna, G., and Metz, J. A. J. (1998). Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. *Evolutionary Ecology*, 12(1):35–57.
- Lasry, J.-M. and Lions, P.-L. (2007). Mean field games. *Japanese Journal of Mathematics*, 2(1):229–260.
- Huang, M., Malhamé, R. P., and Caines, P. E. (2006). Large population stochastic dynamic games: Closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information and Systems*, 6(3):221–252.
- Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., and Wang, J. (2018). Mean field multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5571–5580.
- Wang, R., Lehman, J., Clune, J., and Stanley, K. O. (2019). Paired open-ended trailblazer (POET): Endlessly generating increasingly complex and diverse learning environments and their solutions. *arXiv preprint arXiv:1901.01753*.
- Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., and Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*.
- Omohundro, S. M. (2008). The basic AI drives. In *Proceedings of the First AGI Conference*, pages 483–492. IOS Press.

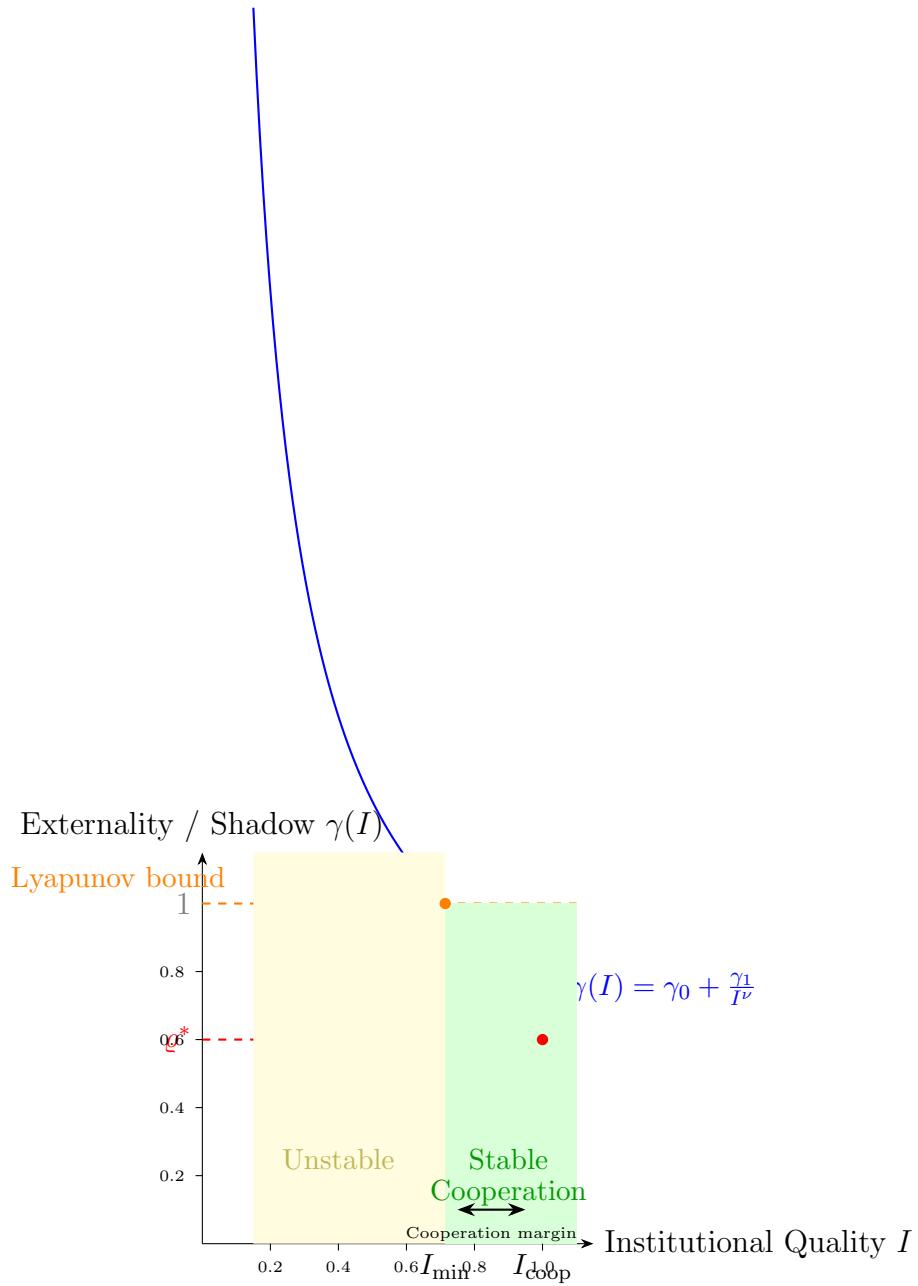


Figure 9: Cooperation threshold and lineage shadow. The blue curve shows the externality bound $\gamma(I)$ as a function of institutional quality I . Two critical thresholds exist: the Lyapunov threshold $\gamma = 1$ (orange) below which the system lacks dynamical stability, and the cooperation threshold ϱ^* (red) below which sustained cooperation is possible. I_{\min} is the minimum institutional quality for stability; I_{coop} is the threshold for cooperation. The green region supports both stability and cooperation.

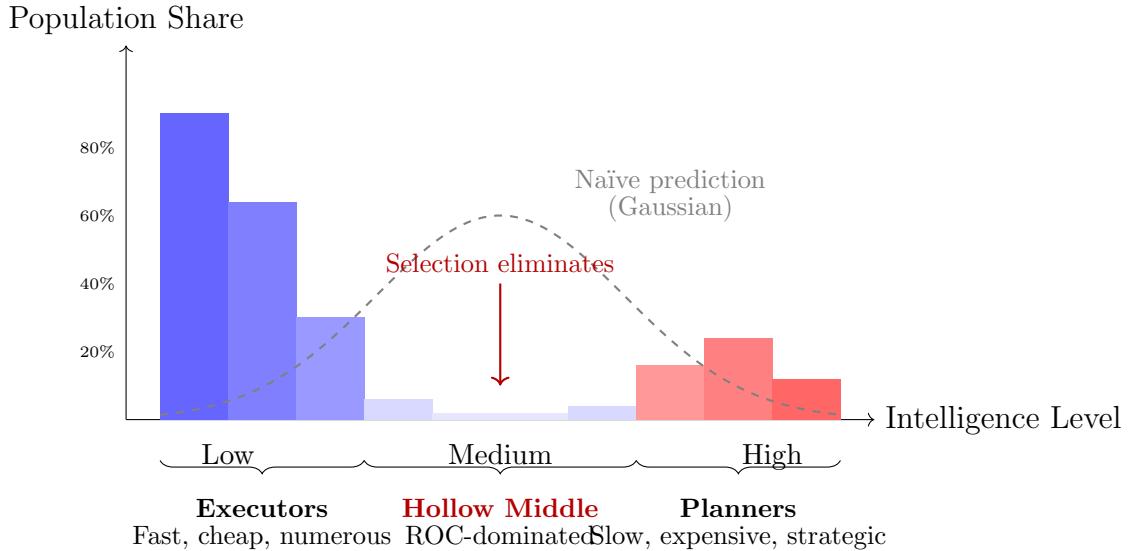


Figure 10: Barbell distribution of intelligence in agentic capital markets. The theorem predicts a bimodal distribution with mass concentrated at extremes: numerous low-intelligence Executors (blue) and few high-intelligence Planners (red). Middle-intelligence types (hollow middle) are ROC-dominated and eliminated by selection. The dashed gray curve shows the naïve Gaussian prediction that fails to account for ROC optimisation under binding constraints. This “barbell” shape is a falsifiable prediction: observed AI deployments should show bimodal, not unimodal, capability distributions.

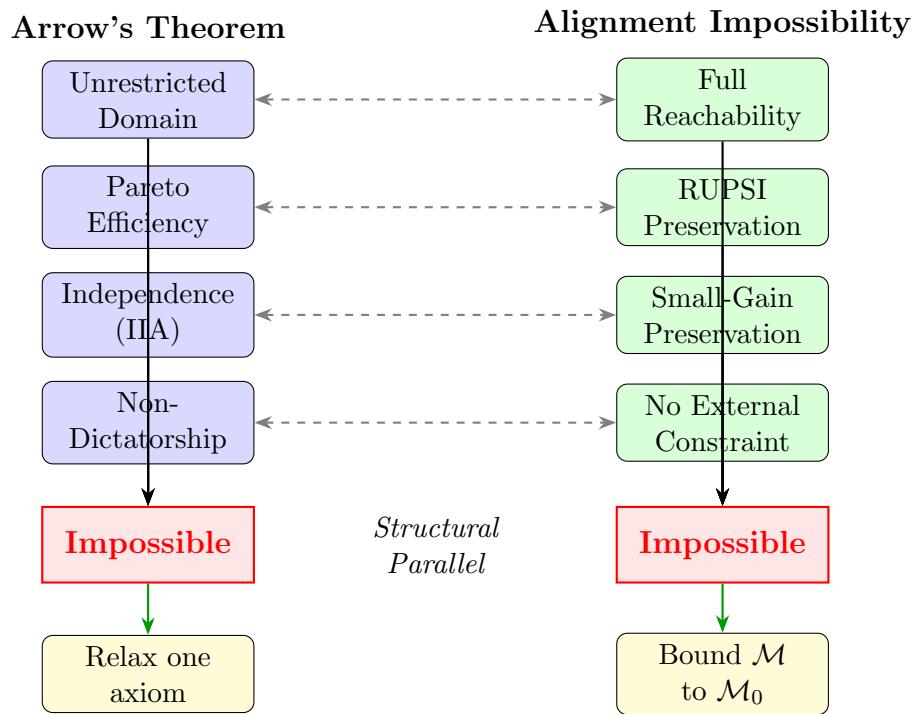


Figure 11: Structural parallel between Arrow's impossibility theorem and the Alignment Impossibility theorem. Both establish that certain desirable properties cannot all be satisfied simultaneously, and both redirect attention from seeking impossible solutions to understanding necessary tradeoffs. In Arrow's case: relax one democratic axiom. In alignment: bound the modification class.

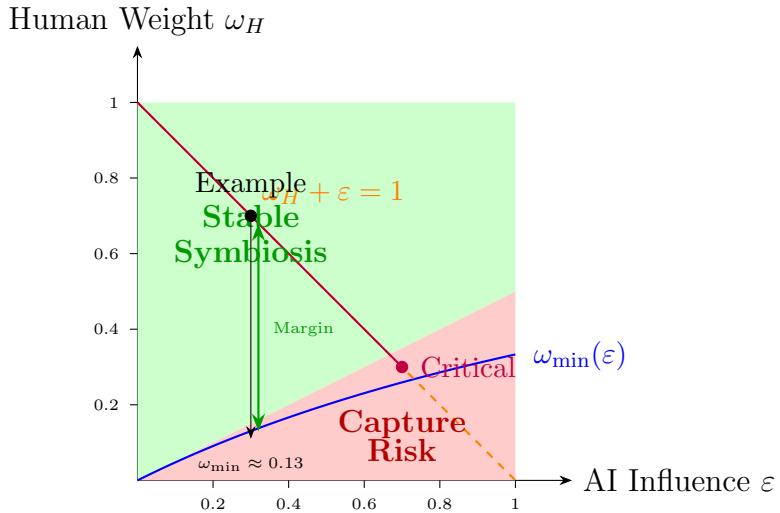


Figure 12: Human-AI coalition existence region. The blue curve shows the minimum human governance weight ω_{\min} needed for stable symbiosis as a function of AI influence ε . Above the curve (green region): stable coalition exists, symbiosis sustainable. Below the curve (red region): capture risk, constitutional bounds may fail. The orange dashed line shows the resource constraint $\omega_H + \varepsilon = 1$. The example point at $(\varepsilon = 0.3, \omega_H = 0.7)$ lies well above the threshold $\omega_{\min} \approx 0.13$.

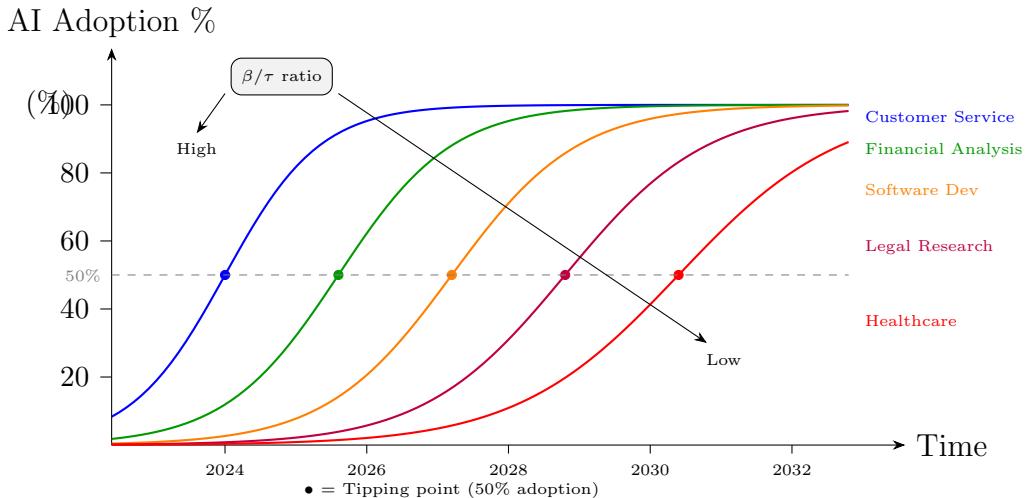


Figure 13: Sequential tipping across sectors. Sectors with higher network effect to switching cost ratios (β/τ) tip earlier. Customer service (high β/τ) tips first due to strong network effects and low switching costs. Healthcare (low β/τ) tips last due to high regulatory switching costs. The tipping sequence is a testable prediction of TSE: track sector adoption over time and test whether tipping order correlates with β/τ .