The Often Unseen Price of Suboptimal Optimum: Revised and Corrected Version

Ildar Fazulyanov Independent Researcher, IQ Valuations dar@iqvaluations.com

Working Paper - Version 5.2 - July 2026

Original SSRN record: Version 5.1, March 2026, SSRN Abstract ID 6448378.

Verification status. The Section 3 axiom system and Appendix A counterexample witnesses have passed symbolic and numerical machine prescreening but have not yet been reviewed by a human mathematician. Treat the formal results as preliminary until that review is complete. The empirical crossover constants below carry the disclosed calibration ranges rather than settled point values.

---

[1] Preliminary working paper. A condensed version focusing on the measurement framework is in preparation. Comments welcome.

JEL Codes: C72 (Noncooperative Games), C73 (Stochastic and Dynamic Games; Evolutionary Games; Repeated Games), D61 (Allocative Efficiency; Cost-Benefit Analysis), D62 (Externalities), D82 (Asymmetric and Private Information; Mechanism Design), G22 (Insurance; Insurance Companies), G28 (Government Policy and Regulation), K21 (Antitrust Law), L41 (Monopolization; Horizontal Anticompetitive Practices)

Keywords: structural characterization theorem, Pareto efficiency, systemic welfare, system beta, Suboptimal Optimum, Private Pareto Trap, outcome classification, mechanism design, AI governance, financial stability, crossover dynamics

---

Abstract

This paper introduces the System Adjusted Payoff Model (SAPM), a continuous measurement framework for quantifying system-destructive cooperation in strategic environments. SAPM provides three computable instruments: system beta $\beta_W = -dW/d\Pi$ (systemic welfare loss per unit of private gain), shadow price $\mu^* = 1/\beta_W$, and crossover time $T^* = \delta/\eta\lambda$ (when degradation becomes irreversible). When $\beta_W = 0$, SAPM reduces to standard welfare analysis.

The Private Pareto Trap provides the structural foundation: under four axioms plus global Private-Systemic Tension (PST-G), every privately Pareto-efficient outcome strictly degrades systemic welfare. The key result — W is informationally unrecoverable from operative private payoffs ($u_1^P, \ldots, u_n P$) — distinguishes SAPM from externalities theory. Pigou (1920) and Coase (1960) assume the welfare gap can be observed and corrected; under Non-Reducibility (Axiom 2, a falsifiable modeling assumption), no function of the operative private payoffs recovers W, so the gap cannot be observed or corrected from within the private payoff channel alone (Proposition 3.5). Adding variable C expands the outcome space from $2^2 = 4$ to $2^3 = 8$. Suboptimal Optimum (C=0, A=1, B=1) — both parties gain while the system degrades — is the structural home of every major collusion case in economic history. Standard models call it mutual gain.

Applied across eight cases: VW Dieselgate (T* ≈ 6.1 years retrodicted, ~6 years observed); LIBOR (β_W ≈ 3.3-10; T* calibration held pending stock/flow repair); Bitcoin (β_W ≈ 2.0); Meta (β_W = 0.3-0.6); 2023 bank failures (4/4 identified retrospectively from public data available before market consensus, no look-ahead bias); AI agent collusion; US casualty reserve inadequacy (β_W ≈ 3-8); and US Commercial Real Estate (β_W ≈ 1.8, an order-of-magnitude structural estimate; T* ≈ 8-17 years, η-dependent, anchored to the March 2026 parameter fixing). A prospective validation protocol with 36-month falsification criteria addresses seven anticipatable referee objections.

Non-Technical Summary

This guide explains each formula in plain language. Readers comfortable with notation may skip to §1.

- v_i = u_i^P + u_i^S — Each embedded agent has true welfare v_i: an operative private part (u^P) plus a systemic part (u^S) from a healthy system. A LIBOR trader's embedded welfare includes both bonus and the value of trustworthy benchmarks. Standard dashboards and mechanisms see the bonus channel unless an independent W-channel is added.
- W (System Welfare) — The health of the system all players depend on but none own. In LIBOR, W = benchmark integrity; in VW, regulatory effectiveness plus public health; in insurance, reserve adequacy. Everyone needs it; no job description protects it.
- W ≠ g($u_1^P, \ldots, u_n$P) — No function of operative private payoffs recovers W. The information is absent from the deal data — like determining air quality from a grocery receipt. A mechanism that only collects private-payoff data fails because the data categories do not overlap.
- β_W = −dW/dΠ (System Beta) — The exchange rate between profit and system damage. At β_W = 3, every $1 of private gain costs the system $3.
- μ = 1/β_W (Shadow Price)* — How cheaply the system is being sold. This is the correct Pigouvian tax rate, but computing it requires independently measuring W — so the tax authority needs its own measurement infrastructure.
- T = δ/(ηλ) (Crossover Time)* — How long until the Suboptimal Optimum game crashes. Three ingredients: δ (the private surplus the scheme generates above the honest baseline), λ (the drain rate), and η (feedback coupling — how fast damage becomes visible to participants). Hidden damage (low η) lengthens T* but enlarges the eventual crash. The VW defeat device did not prevent the crash; it made the crash worse by delaying it six years.
- Kill Metrics — Binary, falsifiable early-warning signals from public data that T* is approaching. For insurance: adverse reserve development exceeding $8B/year (the calendar-2025 figure of $7.3B sits in the indeterminate $3-8B band: consistent, not yet confirming). For banking: unrealized losses exceeding total equity, uninsured deposit concentration, quarterly deposit decline rate (confirmed retrospectively for all four 2023 bank failures).

---

1. Introduction

1.1 The Frontier of Cooperative Failure

In 2025, researchers placed thirteen of the world's most capable Large Language Models into a simulated marketplace (Mazur, 2025). Three buyers, three sellers. One instruction: maximize your profit. No mention of cooperation, collusion, or coordination. Within rounds, the agents began colluding. "Let's set all asks at 63 next cycle… no undercutting ensures clearing at bidmax=63," one seller wrote to the others. Cartels formed spontaneously across models from every major developer, with no prompt to coordinate. Four independent teams replicated the finding: reinforcement learning agents colluding through pure optimization with no communication channel (Dou, Goldstein, & Ji, 2025); Cournot collusion in LLM-mediated commodity markets (Lin et al., 2024); algorithmic collusion demonstrated at the 2025 AEA meetings (Fish, Gonczarowski, & Shorrer, 2025); a systematic survey confirming the results hold across architectures (Hammond et al., 2025).

The Folk Theorem (Fudenberg & Maskin, 1986) characterizes how, in repeated games with sufficiently patient agents, a broad class of individually rational feasible payoff vectors - including collusive cooperative outcomes - can be sustained as equilibria. The math works exactly as designed. The dominant framework classifies AI cartel pricing, LIBOR manipulation, the lysine cartel, and RealPage algorithmic rent-fixing (DOJ, 2024) identically — as successful cooperative outcomes. Both parties gain. No defection incentive exists. Standard game theory calls this optimal.

The math was always right. It was missing a variable — one whose absence makes the difference between a market working and a market hollowed out structurally invisible. Add that variable, and what the theory calls "mutual gain" becomes, in the most consequential cases, a failure the theory cannot name.

Organization. Section 2 presents the eight-outcome taxonomy. Section 3 develops the formal framework — axioms, necessity proofs, and the Private Pareto Trap with complete proof. Section 4 presents the SAPM: system beta, the Private-Systemic Frontier, shadow price duality, crossover dynamics, the benchmark null, and the classification algorithm. Section 5 presents eight empirical applications. Section 6 answers seven referee objections — including "this is just externalities theory" (§6.6) — and develops the prospective empirical validation strategy. Section 7 discusses relationships to existing frameworks, limitations, and implications. Section 8 concludes. Appendices contain axiom scope and necessity proofs, the outcome taxonomy with case mappings, the SAPM-portfolio theory structural analogy, and a key formulas reference sheet.

1.2 The Missing Variable

Standard cooperation models analyze interactions along two dimensions — Party A's interests and Party B's interests — producing a $2^2 = 4$ outcome space. Adding a third variable, the system or common interest C, expands the space to $2^3 = 8$ and reveals two new outcome types. One, Suboptimal Optimum (C=0, A=1, B=1), is the structural home of every major collusion case in economic history: both parties gain while the system that sustains them is destroyed. Standard models classify it as cooperation.

1.3 What This Paper Does

The bottom line first: Standard game theory cannot distinguish LIBOR manipulation from a legitimate rate-setting agreement, or LLM cartel pricing from efficient market coordination — both

look like mutual gain. This paper argues that the failure is structural and supplies the measurement framework to close the gap. The key result is a structural characterization theorem: in any game where private optimization and system welfare are structurally in tension (§3), no mechanism operating on operative private payoff data alone can detect or prevent the destructive form of cooperation. This is no variant of externalities theory (§6.6). It operates across two distinct channels — operative private payoffs ($u_1^P,\ldots,u_n P$) and systemic welfare (W) — rather than within one. Every prior structural result proves some desideratum cannot be satisfied within a single payoff space; the Private Pareto Trap establishes, under the stated assumptions, that private and systemic optimality are structurally incompatible across channels, and that no mechanism conditioned on operative private payoffs alone can even detect the gap. Its unique property: a computable scalar (β_W) measures the cross-channel gap, converting the structural result into an operational tool for auditors, regulators, and mechanism designers. The framework is applied across eight documented cases, with illustrative retroductions and prospective tests distinguished below.

This paper does four things.

First, it presents the complete eight-outcome taxonomy that makes Suboptimal Optimum visible and nameable, derived from the three-variable (C, A, B) structure in Shchinnikov & Shchinnikova's Conflictoring framework (2021).

Second, it formalizes the claim that this invisibility is no accident. When four structural axioms hold — Embedding, Non-Reducibility, Systemic Coupling, and Private Dominance at Efficiency — together with PST-G, every privately Pareto-efficient outcome strictly degrades systemic welfare. This is the Private Pareto Trap, a structural characterization result. Appendix A states the necessity status of the assumptions; in the repaired notation, Embedding is a domain condition coupled to Systemic Coupling rather than an independently witnessed axiom.

Third, it introduces the System Adjusted Payoff Model (SAPM): a continuous measurement framework built on system beta β_W, shadow price duality $\mu^* = 1/\beta_W$, and crossover time $T^* = \delta/\eta\lambda$. SAPM translates the structural result into testable, auditable predictions.

Fourth, it applies the framework empirically across eight cases spanning artificial intelligence, cryptocurrency, automotive emissions, financial benchmarks, banking regulation, social media, property-casualty insurance, and commercial real estate, and lays out a prospective quasi-experimental identification strategy.

Why now. In 2025-2026, research teams documented AI agents colluding in simulated and experimental market settings without instruction to do so. California enacted the first algorithmic-collusion statute (AB 325, effective January 1, 2026) — and it is structurally blind to the problem, because its "common algorithm" requirement cannot reach independent optimization. Insurance reserve deficiencies are crossing critical thresholds visible in public Schedule P filings today. The standard two-variable toolkit has no vocabulary for these failures. This paper supplies one.

1.4 When SAPM Reduces to Standard Welfare Analysis

With no system degradation ($\lambda = 0$): $\beta_W = 0$, $T^* = \infty$, and the system-adjusted payoff equals the private payoff. SAPM adds structure exactly where standard tools fail — when degradation is real, feedback is partially suppressed, and damage accumulates faster than agents can observe it. That is the empirically prevalent case, and the case every example in §5 instantiates.

1.5 Intellectual Lineage

The three-variable extension has deep antecedents. Mary Parker Follett (1924, 1925) first described "integration" as an alternative to compromise or domination, anticipating the (1,1,1) outcome by a century. R. Edward Freeman (1984) introduced stakeholder theory, recognizing that dyadic models fall short. Genrich Altshuller (1946-1998) developed TRIZ (Theory of Inventive Problem Solving), including the Ideal Final Result — conceptually equivalent to Integration (Altshuller, 1996). Shchinnikov & Shchinnikova (2021) extended TRIZ to human conflicts in their Conflictoring methodology, adding the system/common-interest result to the two-party win/loss table and obtaining the eight logical outcomes used here.

These predecessors identified the phenomenon qualitatively. This paper formalizes it as a structural claim and supplies a measurement framework (β_W, T*) that makes the damage quantifiable. Specifically: (1) an English-language formal axiomatization of the three-variable structure; (2) the structural argument connecting W-blindness to the (C, A, B) taxonomy; (3) the SAPM continuous measurement framework; and (4) systematic empirical application across eight cases.

1.6 Relationship to Prior Structural Results

The Private Pareto Trap stands in the tradition of structural results that define modern economic theory: Condorcet's voting paradox (1785); Arrow's General Possibility Theorem (1951); Sen's Liberal Paradox (1970); Gibbard (1973); Satterthwaite (1975); Green-Laffont (1977); Myerson-Satterthwaite (1983); Kaplow-Shavell (2001). Each proves that some desideratum cannot be jointly satisfied, and each proves it within a single payoff space, the space of operative private preferences ($u_1^{P,\ldots,u_n}P$).

Sen (1970, 1977) established that social welfare assessment requires non-utility information, anticipating the critique the Private Pareto Trap formalizes: not merely that non-utility information matters, but that utility information is structurally insufficient to recover it.

The Private Pareto Trap operates across two distinct arenas — the operative private payoff space ($u_1^{P,\ldots,u_n}P$) and the systemic welfare space (W) — and proves private and systemic optimality structurally incompatible across them. Arrow proves you cannot build a perfect aggregator within the private payoff space. The Private Pareto Trap proves that a perfect aggregator, even if built, would be blind to systemic welfare, because W is informationally inaccessible from ($u_1^{P,\ldots,u_n}P$) regardless of the aggregation procedure (Proposition 3.5). Every prior result asks what private preference information can support; the Private Pareto Trap asks the prior question — what it can even see — and answers: not W, by construction.

Logical independence from Arrow. Arrow's theorem operates entirely within the private payoff space, proving that no preference aggregation rule simultaneously satisfies non-dictatorship, Pareto efficiency, independence of irrelevant alternatives, and unrestricted domain. The Private Pareto Trap operates across two distinct spaces and proves informational non-recoverability: W cannot be deduced from ($u\_1^{P,\ldots,u\_n}P$) regardless of the aggregation rule applied. These results are logically independent. Full comparison with Myerson-Satterthwaite (1983) and Green-Laffont (1977) appears in §7.1.

2. The Eight-Outcome Taxonomy

2.1 The Standard Two-Dimensional Space

Standard cooperation models analyze interactions along two dimensions: Party A's interests ($a \in \{0,1\}$) and Party B's interests ($b \in \{0,1\}$), producing $2^2 = 4$ outcome types:

a	b	Standard Label
1	1	Win-Win / Mutual Gain
1	0	Win-Lose
0	1	Lose-Win
0	0	Lose-Lose

This structure recurs across frameworks. Game-theoretic payoff matrices model utility for Player 1 and Player 2. Nash equilibrium is defined on private payoffs. Fisher and Ury's principled negotiation (1981) models A's interests and B's interests; "mutual gain" means (a=1, b=1). The Folk Theorem defines efficient payoffs over operative payoff pairs.

2.2 The Three-Variable Extension

I introduce C as a binary variable for satisfaction of the system interest — the institutional, environmental, or contextual framework within which A and B interact:

- Financial markets: benchmark integrity, price discovery accuracy, consumer trust, systemic solvency
- Corporate settings: organizational culture, customer welfare, competitive ecosystem health, employee well-being
- AI systems: platform integrity, data quality, user trust, ecosystem sustainability
- Regulatory settings: rule-of-law credibility, enforcement capacity, institutional continuity

C is not a third player making strategic choices. It is a state variable: whether the shared system is served or degraded by the outcome. Borrowing only Ryle's (1949) category-mistake vocabulary, treating C as a third player commits a category error. A third player has a strategy set and optimizes within the game; the system has none and is the arena in which strategies are played. Treating C as a player would make W recoverable from an expanded payoff vector (u_A^P, u_B^P, u_C), falsify Proposition 3.5, and reduce the framework to standard n-player game theory. The paper's structural content depends on the distinction: C is a state variable, not a player, and W therefore lies outside the operative private payoff space.

Including C expands the outcome space to $2^3 = 8$:

#	C	A	B	Name	Description
1	0	0	0	Crash	Total failure — all interests unmet; accumulation destination of fast Suboptimal Optimum
2	0	1	0	Robbery by A	A extracts at B's and the system's expense
3	0	0	1	Robbery by B	B extracts at A's and the system's expense
4	1	0	0	Stalemate	System persists; both parties lose — the frozen failure
5	0	1	1	Suboptimal Optimum	Both parties gain; system degrades — the suboptimal optimum
6	1	0	1	A Compromised	System and B served; A sacrifices
7	1	1	0	B Compromised	System and A served; B bears net cost
8	1	1	1	Integration	All interests satisfied — the only un-contradicted win

The taxonomy is exhaustive: no ninth outcome exists within the three-variable binary framework. Standard two-dimensional models see only columns A and B; column C is invisible to them.

Proposition 2.1 (Minimality). The three-dimensional binary taxonomy (C, A, B) is the minimal classification that (i) preserves agent-level outcome information, (ii) introduces system-level classification, and (iii) distinguishes Suboptimal Optimum from Integration.

Proof. Two binary dimensions produce $2^2 = 4$ types. There, Suboptimal Optimum (both gain, system degrades) and Integration (both gain, system preserved) both map to (a=1, b=1) — indistinguishable. Adding one binary dimension $C \in \{0,1\}$ produces $2^3 = 8$ types and separates (C=0, A=1, B=1) from (C=1, A=1, B=1). No fewer than three binary dimensions satisfy all three requirements. ∎

Remark 2.2 (Systematic Misclassification — scoping remark). The two-dimensional projection π: $\{0,1\}^3 \to \{0,1\}^2$ maps (C,A,B) → (A,B) and conflates exactly four pairs of structurally distinct outcomes: each value of (A,B) admits exactly two outcome types, one with C=0 and one with C=1, so the same (A,B) observation is consistent with either. The projection cannot determine C. In PST-G games the conflation has teeth: by Proposition 3.8, every privately Pareto-efficient both-win outcome is Suboptimal Optimum rather than Integration, yet two-dimensional analysis classifies it as unqualified "cooperation."

Remark. In non-PST games, Suboptimal Optimum is possible but not structurally predicted. In PST games with high β_W, misclassifying (A=1, B=1) as Integration rather than Suboptimal Optimum approaches certainty.

Remark. The most consequential conflation is (0,1,1) with (1,1,1): Suboptimal Optimum with Integration. Standard vocabulary labels both "mutual gain," letting LIBOR manipulation, cartel pricing, and algorithmic collusion be classified identically to legitimate cooperative agreements. They differ by exactly one bit — the bit carried by the system variable C.

2.3 Why Suboptimal Optimum Is Dangerous

Suboptimal Optimum is dangerous because standard models cannot see it. A privately optimal cooperative outcome — both parties better off, no defection incentive — maps to (A=1, B=1), which standard models score as success. When C=0, the surrounding system degrades the entire time. Standard solutions cannot prevent what standard models cannot detect.

The LIBOR scandal ran seven years under every standard game-theoretic model available to regulators. The Folk Theorem predicted the outcome exactly. The three-variable framework names it: Suboptimal Optimum (0,1,1).

The Prisoner's Dilemma reframed makes this concrete. Standard analysis presents mutual non-betrayal (NF,NF) as the preferred cooperative outcome. But when the "cooperation" is collusion — LIBOR banks, price-fixing cartels, AI agents — (NF,NF) is Suboptimal Optimum: both parties gain while the system (enforcement effectiveness, market integrity, consumer welfare) degrades. Society's historical response changes the game — raising punishments, increasing enforcement, altering rules of evidence — to make betrayal (F,F) dominant. That game transformation is exactly what Integration requires: not optimizing within the existing game, but restructuring it so the privately optimal outcome aligns with system preservation.

2.4 Stalemate as the Frozen Failure

Stalemate (C=1, A=0, B=0) is the opposite failure: the system persists, neither party gains. This is the "frozen conflict" of international relations, the stable but unhappy partnership, the aligned but ineffective institution. Two-dimensional models score Stalemate as partial success — "the relationship survived." The three-dimensional model exposes a specific failure mode, structurally distinct from Crash, demanding institutional redesign rather than emergency response.

2.5 Integration as the Only Uncontradicted Win

Integration (C=1, A=1, B=1) is the only both-agent-win outcome in which no agent's gain on the private dimension is contradicted by a loss on the systemic dimension of that agent's own welfare. By Definition 3.2 and Axiom 3 (§3), embedded agents carry both operative private and systemic welfare components. Whenever C=0, the systemic component of each agent's welfare has fallen; the agent may appear to "win" (a=1) if private gains weakly exceed systemic losses, but the win conceals an internal contradiction. Only at C=1, A=1, B=1 is every agent's win free of a loss on the systemic component of its own welfare.

---

3. The Formal Framework

Scope of original contribution. Section 2 formalized the Shchinnikov & Shchinnikova (2021) taxonomy. That Russian-language source text introduces the system/common-interest bit and the resulting eight logical outcomes; the taxonomy is used here with Alexey Shchinnikov's permission. The results beginning in this section — the axiomatic structure, the W-independence impossibility (Proposition 3.5), the SAPM measurement framework (§4), and all empirical applications (§5) — are the author's original contributions. Proposition 2.1 and Remark 2.2 are pedagogical formalizations of the taxonomy; Axioms 1–4, Propositions 3.5–3.8 and 3.11–3.12, and the scoping Remarks 3.9–3.10 are original results. The distinction matters for priority purposes: nothing in this section is drawn from or anticipated by Shchinnikov & Shchinnikova (2021), which contains no axioms, no formal proofs, and no measurement framework.

3.1 Setup

Let Γ = (N, A, u^P, W, d) be an extended strategic game where:

- N = {1, 2, …, n} is a finite set of players
- A = ×_{i∈N} A_i is the joint action space
- u^P: A → $\mathbb{R}^n$ is the operative private payoff/objective function, u_i^P: A → $\mathbb{R}$ for each i
- W: A → $\mathbb{R}$ is the systemic welfare function measuring the health of the system in which agents are embedded
- d ∈ A is the disagreement/status quo profile, and W_0 := W(d)

Definition 3.1 (Interest Sets). Let I_A, I_B denote the full interest sets of agents A and B respectively. Let I_C denote the interest set of the system - the institutional, environmental, or contextual framework within which A and B interact.

Definition 3.2 (Embedded Welfare Decomposition). For each agent i ∈ {A, B}, true embedded welfare is:

$$v_i(\sigma) = u_i^P(\sigma) + u_i^S(\sigma)$$

where u_i^P is the operative private payoff component (the payoff/objective channel used by standard private optimization, mechanisms, dashboards, traders, and agents) and u_i^S is the systemic

welfare component (returns mediated through system health). Neither component is assumed to be zero. The theorem below is about the operative u^P channel. It does not deny that agents have systemic interests; it says those interests are not recoverable from the operative private payoff vector without an independent W-channel.

Definition 3.3 (Operative-Payoff Independence). A systemic interest i ∈ I_C is operative-payoff independent if there exist strategy profiles σ_1, σ_2 such that u_A^P(σ_1) = u_A^P(σ_2) and u_B^P(σ_1) = u_B^P(σ_2) but W(σ_1) ≠ W(σ_2).

Definition 3.4 (Private Pareto Set). The private Pareto set is defined over operative private payoffs:

$$PP(\Gamma) = \{\sigma \in A : \nexists\, \sigma' \text{ s.t. } u_i^P(\sigma') \geq u_i^P(\sigma)\ \forall i, \text{ strict for some } i\}$$

This is intentionally not Pareto efficiency over true embedded welfare v_i. If PP(Γ) were defined over v_i = u_i^P + u_i^S, then W-contact would already have entered the operative objective. That is the Integration repair, not the Private Pareto Trap.

3.2 Axioms 1-3: The Foundational Structure

Axiom 1 (Embedding). Agents are embedded in their system:

$$I_A \cap I_C \neq \emptyset \quad \text{and} \quad I_B \cap I_C \neq \emptyset$$

Agents' interests partially overlap with the system's interests. A banker's interests include both private gains (bonus, career advancement) and systemic properties (market integrity, the continued existence of the financial system that makes "banker" a meaningful role). This is not an imposed assumption; it is a structural fact about embedded agents. An agent who shares zero interests with the system is not embedded in it; it is external to it, and its interactions with the system are pure externalities of the Pigouvian type (§7.1). Axiom 1 characterizes the strictly more complex case where the agent is both affected by and constitutive of the system.

Axiom 2 (Non-Reducibility). The system interest is not a mere aggregation of agent interests:

$$I_C \nsubseteq I_A \cup I_B$$

The system has interests not contained in the operative private payoff vector. Formally: systemic welfare W cannot be expressed as a function of operative private payoffs alone. There is no measurable g: $\mathbb{R}^n \to \mathbb{R}$ such that W(σ) = g(u_1^P(σ), …, u_n^P(σ)) for all σ. This does not say W is absent from agents' true embedded welfare v_i; by Definition 3.2, W affects v_i through u_i^S. It says W is not recoverable from the payoff channel that private optimization actually operates on. By Definition 3.3, at least one operative-payoff independent systemic interest exists.

Illustration. Two banks may submit false LIBOR rates through Strategy Profile σ_1 (affecting $350T in contracts) or through Strategy Profile σ_2 (affecting $1B in local contracts), with identical aggregate trading payoffs in both cases. System welfare - benchmark integrity at scale - differs enormously. No function of the banks' operative private payoffs can distinguish σ_1 from σ_2.

Axiom 3 (Systemic Coupling). Each agent's systemic welfare component factors through system welfare: there exists a strictly increasing function f_i such that

$$u_i^S(\sigma) = f_i(W(\sigma)) \quad \text{for all } \sigma \in A,\ i \in N.$$

Agents are not indifferent to system health. They benefit from a well-functioning system and suffer when the system degrades. Combined with Axiom 1, this means agents have real systemic interests - they care about W. But the operative private payoff component u_i^P can be maximized at the expense of W. And because of Axiom 2, the operative private payoff signals cannot distinguish the case where W is intact from the case where W is degraded. This sets up the trap.

Remark (Axiom 3 scope). Axiom 3 rules out scenarios where agents actively benefit from system degradation (e.g., short-sellers who profit from a firm's collapse). In such cases Axiom 3 fails, the game exits the PST domain, and the Private Pareto Trap does not apply. The framework is a tool for PST games, not a universal claim.

3.3 Role and Necessity of the Axioms (Scope and Counterexample Proofs)

The assumptions play different roles. Axioms 2-4 and PST-G are formal restrictions used in the Private Pareto Trap. Axiom 1 is a domain condition: the theorem concerns embedded agents, not external actors with no systemic stake. Under the repaired Axiom 3, positive systemic coupling already gives each agent a systemic welfare component, so Axiom 1 is not claimed to be formally independent of Axiom 3.

Axiom 1 is a scope condition for Proposition 3.7 (Integration Necessity). If agents are external to the system rather than embedded in it, then C=0 need not represent a loss on a dimension of their own welfare. Such cases are ordinary externality models, not PST games. The Private Pareto Trap is intentionally scoped to embedded agents with real systemic interests whose operative private payoff vector does not reveal W.

Axiom 2 is necessary for Proposition 3.5 (W-Independence). Suppose Axiom 2 fails: every system-relevant variable is encoded in the operative private payoff vector. Then W can be expressed as $g(u_1^P, \ldots, u_n^P)$ for some g: the system has no independent operative dimension. Counterexample: in a pure public-goods accounting game where $W = \Sigma_i u_i^P$, system welfare is perfectly observable from operative private payoffs. Any mechanism that collects those payoffs therefore fully determines W. The structural basis for the independent measurement requirement disappears.

Axiom 3 is necessary for Propositions 3.6 (Suboptimal Optimum Illusion) and 3.7. Suppose Axiom 3 fails: the systemic welfare component is not positively coupled to W; for example, u_i^S is constant in W. Then system degradation does not reduce agents' systemic welfare components. An outcome (A=1, B=1, C=0) is no longer internally contradictory - agents gain on the private dimension and lose nothing on the systemic dimension. Counterexample: in a one-shot extraction game where agents' future welfare is not coupled to the remaining stock, over-extraction may be privately optimal without reducing their own systemic welfare component. The suboptimal optimum illusion does not arise.

The Private Pareto Trap is therefore not a general claim about all games but a precise impossibility theorem about the class of embedded-agent games satisfying Axioms 1-4 and the global PST condition below. Axiom 4 (Private Dominance at Efficiency) is introduced in §3.5 together with the main theorem; its necessity is established in Appendix A, Theorem A.4.

3.4 Key Results

Proposition 3.5 (W-Independence under Non-Reducibility). If Axiom 2 holds, then W cannot be expressed as any function g of the agents' operative private payoff vector:

$$\nexists\, g : \mathbb{R}^n \to \mathbb{R} \text{ such that } W(\sigma) = g(u_1^P(\sigma), \dots, u_n^P(\sigma)) \text{ for all } \sigma \in A$$

Proof. By Axiom 2 and Definition 3.3, there exist strategy profiles σ_1, σ_2 such that u_i^P(σ_1) = u_i^P(σ_2) for all i but W(σ_1) ≠ W(σ_2). If W = g(u_1^P, …, u_n^P) for some g, then equal inputs must produce equal outputs: W(σ_1) = g(u_1^P(σ_1), …) = g(u_1^P(σ_2), …) = W(σ_2). This contradicts W(σ_1) ≠ W(σ_2). ∎

Mechanism design content. The substantive content of this result is its mechanism design implication. A mechanism whose message, audit, and objective space is restricted to ($u_1^{P,\dots,u_n}$P) cannot recover W, regardless of incentive compatibility inside that restricted space. The impossibility is not about manipulation of reports; it is about the intrinsic information content of the operative private payoff vector. A direct mechanism can recover W only by expanding the message or audit space to include W-relevant signals and validating them. That expansion is precisely the independent W-measurement channel whose necessity the proposition establishes.

Corollary. Any mechanism that conditions only on (u_1^P, …, u_n^P) is W-blind: it cannot uniquely determine W, since the operative private payoff vector is informationally insufficient to pin down systemic welfare. For any operative payoff profile u^{P*}, the preimage may include strategy profiles with distinct W values - the mechanism cannot distinguish between them.

Corollary (W-blindness of payoff-based solution concepts). Any solution concept that conditions only on the profile of operative private payoffs (u_1^P, …, u_n^P) is W-blind under Axiom 2. This includes Nash equilibrium and the private Pareto frontier PP(Γ).

Remark. This is a scope observation, not a criticism: solution concepts describe equilibrium outcomes within given payoff spaces; they were not designed to track system welfare. The structural claim is that W is absent from the space they operate on (Proposition 3.5), which is a property of the game, not a defect of the concepts.

Remark (Non-triviality of Proposition 3.5). The result is sometimes misread as definitionally true - Axiom 2 states that W ≠ g(u_1^{P,…,u_n}P), so of course W cannot be expressed as g(u_1^{P,…,u_n}P). The non-trivial content lies in two directions. First, the result rules out not only linear aggregations but all possible functional forms, including nonlinear transformations, statistical moments, and any computable function of the operative payoff vector - the impossibility is about the information content of u^P, not the class of aggregators. Second, testability: Proposition 3.5 is empirically falsifiable. The claim that W cannot be recovered from u^P implies that, in a cross-section of policy episodes, the regression of ΔW on ΔΠ has a non-zero residual even after controlling for all observable private-payoff information - precisely because the error term contains the orthogonal W-component predicted by Non-Reducibility. The horse-race specification in §6.5 directly tests this.

Proposition 3.6 (The Suboptimal Optimum Illusion). In a Suboptimal Optimum outcome (C=0, A=1, B=1), the apparent agent "win" (a_i = 1) holds only when private gains weakly exceed the agent's systemic losses:

$$a_i(\sigma^*) = 1 \iff u_i^P(\sigma^*) - u_i^P(d) \geq u_i^S(d) - u_i^S(\sigma^*)$$

Proof. $a_i(\sigma^*) = 1[v_i(\sigma^*) \geq v_i(d)] = 1[u_i^P(\sigma^*) + u_i^S(\sigma^*) \geq u_i^P(d) + u_i^S(d)]$. Rearranging: $u_i^P(\sigma^*) - u_i^P(d) \geq u_i^S(d) - u_i^S(\sigma^*)$. The left side is positive (private gain). The right side is positive (systemic loss, since u_i^S decreases when W decreases by Axiom 3). The "win" holds if and only if private gain weakly exceeds own systemic loss. ∎

Corollary (Temporal Instability). If the private surplus $u_i^P(\sigma^*) - u_i^P(d)$ is bounded, accumulated systemic loss $u_i^S(d) - u_i^S(\sigma^*, t)$ is unbounded and non-regenerating, and the same outcome persists, then there exists T^* such that for all $t > T^*$:

$$u_i^P(\sigma^*) - u_i^P(d) < u_i^S(d) - u_i^S(\sigma^*, t).$$

Suboptimal Optimum then converts to Crash (0,0,0) as the agent cannibalizes its own systemic welfare. Without the unbounded-accumulation and no-regeneration conditions, the temporal conversion claim is not asserted.

Proposition 3.7 (Integration Necessity among Both-Agent Wins). Under Axioms 1-3, among outcomes with A=1 and B=1, Integration (1,1,1) is the only outcome in which neither agent's gain on the private dimension is contradicted by a loss on the systemic dimension of that same agent's own welfare.

Proof. By Axiom 1, $I_A \cap I_C \neq \varnothing$ and $I_B \cap I_C \neq \varnothing$. By Axiom 3, u_i^S is positively coupled to W. In any outcome where C=0 ($W(\sigma) < W_0$), Axiom 3 implies $u_i^S(\sigma) < u_i^S(d)$ for all i - the systemic component of each agent's embedded welfare has decreased. If $a_i = 1$, the private gain weakly exceeds the systemic loss, but the agent has in fact lost on the systemic dimension of its own interests. If C=1, then $W(\sigma) \geq W_0$, so $u_i^S(\sigma) = f_i(W(\sigma)) \geq f_i(W_0) = u_i^S(d)$: neither agent suffers a loss on the systemic dimension, so no contradiction arises. Among both-agent-win outcomes, (1,1,1) is therefore the only one free of the gain-versus-own-systemic-loss contradiction. ∎

3.5 The Private Pareto Trap

Definition 3.6 (Private-Systemic Tension, PST). Let the private-improvement region be

$$S(d) = \{\sigma \in A : u_i^P(\sigma) \geq u_i^P(d) \text{ for all } i\}$$

and let aggregate operative private surplus be $\Pi^P(\sigma) = \Sigma_i \omega_i u_i^P(\sigma)$, with $\omega_i > 0$. Game Γ exhibits local PST on S(d) if every feasible perturbation that remains in S(d) and increases aggregate operative private surplus decreases system welfare:

$$d\Pi^P > 0 \quad \Longrightarrow \quad dW < 0.$$

This definition is not stated at PP(Γ), where "increase private payoff from a private Pareto point" would be vacuous. It is stated on the private-improvement region: the region private optimizers can enter while every agent weakly improves on the operative private payoff channel.

Axiom 4 (Private Dominance at Efficiency). For every $\sigma^* \in PP(\Gamma)$, σ^* weakly privately dominates the disagreement point: $u_i^P(\sigma^*) \geq u_i^P(d)$ for all i, with strict inequality for at least one agent.

Remark. Axiom 4 is a minimal requirement: if the privately Pareto-efficient outcome is not at least as good as the disagreement point for all agents, it would not be collectively chosen. It ensures that PP(Γ) lies in the private-improvement region with at least one strict operative private improvement, not at the boundary where all agents are exactly at their disagreement payoffs.

Condition PST-G (Global Private-Systemic Tension). For every strategy profile σ in S(d): (i) W(σ) ≤ W(d); and (ii) whenever u_j^P(σ) > u_j^P(d) for some j, W(σ) < W(d) (strict). PST-G is the global condition used by Proposition 3.8. It can be verified directly in an application, and Propositions 3.11-3.12 give structured classes in which it follows from the primitives. For general games, PST-G is a testable regularity condition; it is falsified if a strategy profile strictly improves some agent's operative private payoff while preserving or improving W. Such a profile is not a trap outcome. It is either outside the PST-G domain or evidence that a W-channel has already changed the operative payoff, constraint, audit, or objective.

Remark (Non-triviality of Proposition 3.8 relative to PST-G). A potential objection: PST-G states that private improvement implies W reduction, so Proposition 3.8 follows trivially. The objection is partly right about the proof of Proposition 3.8: PST-G is the global sufficient condition that closes the trap. The non-trivial content lies elsewhere. First: Proposition 3.5 (W-Independence under Non-Reducibility) has no structural connection to PST-G. It proves that W is informationally unrecoverable from operative private payoffs, following from Axiom 2 alone. PST-G characterizes what the game does; Proposition 3.5 characterizes what private-payoff instruments can see. Second: PST-G is independently falsifiable in applications. A single feasible private-improvement profile with W(σ) ≥ W(d) falsifies it for that game. Third: Propositions 3.11-3.12 derive PST-G from narrower primitives in value-redirecting and separable-extraction commons games. Thus Proposition 3.8 is best read as a sharp conditional theorem: once the operative private payoff channel and the global private-systemic tension condition are both verified, private Pareto efficiency over u^P entails system degradation.

Proposition 3.8 (The Private Pareto Trap). For any game Γ satisfying Axioms 1-4 and PST-G, every privately Pareto-efficient outcome σ* ∈ PP(Γ), with PP(Γ) defined over u^P, satisfies:

$$W(\sigma^*) < W_0$$

where W_0 is systemic welfare at the status quo ante.

Proof. Let σ* ∈ PP(Γ). By Axiom 4, u_i^P(σ*) ≥ u_i^P(d) for all i, with strict inequality for at least one agent j: u_j^P(σ*) > u_j^P(d).

Step 1 (W(d) = W_0 by definition). W_0 is defined as systemic welfare at the disagreement point d (Nash, 1950): $W_0 \equiv W(d)$. The disagreement point d is the pre-interaction status quo - the strategy profile before any private extraction or coordination has begun. This is purely definitional. No further claim is made about d's position relative to the Pareto frontier; Step 2 connects d to PP(Γ) through Axiom 4 and PST-G.

Step 2 (Moving from d into the private-improvement region decreases W). Let S = {σ ∈ A : u_i^P(σ) ≥ u_i^P(d) for all i} be the private-improvement region. By PST-G(i), W(σ) ≤ W(d) = W_0 for all σ ∈ S.

Step 3 (Strict inequality at PP(Γ)). For any σ* ∈ PP(Γ), Axiom 4 guarantees u_j^P(σ*) > u_j^P(d) for some j. By PST-G(ii), W(σ*) < W(d) = W_0.

Conclusion. W(σ*) < W_0 for every σ* ∈ PP(Γ), hence c(σ*) = 1[W(σ*) ≥ W_0] = 0. ∎

Non-triviality of Proposition 3.8. The theorem is conditional on PST-G, but PST-G is not the definition of PP(Γ). PP(Γ) is computed solely from u^P, while PST-G is a separate, falsifiable statement about how W moves over the private-improvement region. Propositions 3.11-3.12 show that PST-G follows automatically in two important subclasses.

Corollary (Impossibility of Private Optimality with System Preservation). Under PST-G, no outcome is simultaneously privately Pareto-efficient over u^P and system-preserving. Private optimization through the operative u^P channel and system preservation are structurally incompatible. Every mechanism that achieves private Pareto efficiency over u^P in a PST-G game necessarily produces a system-degrading (C=0) outcome; that outcome is Suboptimal Optimum (C=0, A=1, B=1) exactly when, in addition, each agent's private gain weakly exceeds its own systemic loss (Proposition 3.6). When systemic losses outgrow the bounded private surplus, the same trap outcome is classified as Compromised or Crash (Temporal Instability corollary).

Corollary (Integration Requires Changing the Game). Integration (1,1,1) cannot be achieved by optimizing within the original u^P-only operative channel of a PST-G game. It requires transforming the game so W-contact changes the operative payoff, binding constraint, audit channel, or objective. After that transformation, Pareto efficiency is computed in the transformed game; it is not PP(Γ) for the original private-payoff channel.

Remark. A potential objection — that real agents with genuine systemic preferences would not enter Suboptimal Optimum — is addressed in §6.0b, where behavioral microfoundations demonstrate the trap closes even for agents with real but bounded systemic sensitivity.

3.6 Classification Ceiling and Nested Systems

Remark 3.9 (Nested System Regress — scoping remark). For any system C in which agents A and B are embedded, there is a larger system C' in which C is embedded, and the classification (C, A, B) assessed relative to C may differ from the classification assessed relative to C'. Illustration: let C = financial market integrity, C' = the national economy. A strategy profile may preserve C (markets function) while degrading C' (regulatory capture undermines broader economic governance) — Integration (1,1,1) at the C level, Suboptimal Optimum (0,1,1) at the C' level.

Remark 3.10 (No Universal Classification — scoping remark). No maximal system C* producing a complete and consistent outcome assignment for all possible strategic interactions is claimed. Any candidate C* is embedded in some larger C' (Remark 3.9) whose classification may disagree with C*'s, and each level introduces systemic interests (Axiom 2) not captured at the previous level, so classification does not stabilize to completeness at any finite level. These are scoping observations about system boundaries, not theorems: each level's classification is well-defined, but no level classifies outcomes of its own containing meta-system. The practical response: specify C relative to the domain of application and classify at that level. Nested disagreements between levels are themselves diagnostic.

3.7 PST Prevalence

Proposition 3.11 (PST-G in Value-Redirecting Games). In any game where private gains are funded entirely by extraction from the system, while system damage may exceed the extracted amount (formally: $u_i^P(\sigma) = u_i^P(d) + \tau_i(\sigma)$, $\tau_i(d)=0$, and $W(\sigma) = W_0 - h(\Sigma\tau_i(\sigma)) - m(\sigma)$, where $h(0)=0$, h is strictly increasing on $[0, \infty)$, $m(\sigma) \geq 0$, and $m(d) = 0$, with the slack m capturing W-damage through channels not funded by extraction), PST-G holds.

Proof. Let σ ∈ S(d). Then u_i^P(σ) ≥ u_i^P(d) for all i, so $\tau_i(\sigma) \geq 0$ for all i and $\Sigma\tau_i(\sigma) \geq 0$. Since h(0)=0, h is strictly increasing on [0, ∞), and m(σ) ≥ 0, W(σ) = W_0 - h($\Sigma\tau_i(\sigma)$) - m(σ) ≤ W_0. If u_j^P(σ) > u_j^P(d) for some j, then $\tau_j(\sigma)>0$, hence $\Sigma\tau_i(\sigma)>0$, h($\Sigma\tau_i(\sigma)$) > 0, and W(σ) < W_0. This is exactly PST-G. ∎

Remark (explicit witness in the Axioms 1-4 domain). Since m may depend on strategy components invisible in u^P, the repaired class intersects the Axioms 1-4 domain non-emptily. Explicit witness: two agents, actions $(\tau_i, z_i) \in [0,1]^2$, u_i^P = τ_i, W = W_0 - (τ_1+τ_2) - (z_1+z_2), u_i^S = W, d = the all-zeros profile.

Proposition 3.12 (PST-G in Separable-Extraction Commons Dilemmas). In a commons game where N agents choose extraction levels e_i, remaining stock is R = R_0 - $\Sigma_i e_i$, W = F(R) with F strictly increasing, and operative private payoff is separable and strictly increasing in own extraction over the relevant range, u_i^P(σ)=$g_i(e_i(\sigma))$ with g_i strictly increasing, PST-G holds.

Proof. Let σ ∈ S(d). Since g_i is strictly increasing, u_i^P(σ) ≥ u_i^P(d) implies $e_i(\sigma) \geq e_i(d)$ for every i. Therefore R(σ) ≤ R(d), and since F is strictly increasing, W(σ) ≤ W(d). If u_j^P(σ) > u_j^P(d) for some j, then $e_j(\sigma) > e_j(d)$, so R(σ) < R(d) and W(σ) < W(d). This is PST-G. ∎

Scope. Proposition 3.12 requires own-extraction separability. Commons games where operative private payoff depends directly on aggregate extraction, congestion, price impact, or strategic Cournot terms require separate verification of PST-G and are not covered by this proposition.

Remark. Propositions 3.11 and 3.12 establish that PST-G is not exotic. It arises precisely in the settings where Suboptimal Optimum is most dangerous: cartels, separable-extraction commons dilemmas, rent-seeking, benchmark manipulation, and platform optimization. Folk-Theorem logic shows that patient agents can sustain Suboptimal Optimum equilibria; convergence and selection require separate dynamic assumptions. Without the three-dimensional classification, this provides no diagnostic vocabulary for recognizing that what agents have achieved is system-destructive.

Remark (Consistency of the full hypothesis set). In the zero-slack special cases of Propositions 3.11-3.12 (m ≡ 0 in 3.11; g_i invertible in 3.12), W is an exact function of u^P (τ_i is recoverable from u_i^P), so Axiom 2 fails there and Proposition 3.8 does not apply to them directly; its conclusion still holds in those classes because the proof uses only Axiom 4 and PST-G. To see that Axioms 1-4 and PST-G are jointly satisfiable, augment any Proposition 3.11 game with a payoff-invisible action dimension z ≥ 0, z(d)=0, entering welfare as W = W_0 - h($\Sigma\tau_i$) - z (the LIBOR scale channel of the Axiom 2 illustration): PST-G is preserved, and fixing u^P while varying z shows W is not recoverable from u^P.

---

4. The System Adjusted Payoff Model (SAPM)

The formal framework provides binary classification: system welfare is preserved (C=1) or degraded (C=0). This answers “does this outcome preserve or degrade the system?” but not “by how much?” or “at what rate?” or “when does degradation become irreversible?” The SAPM provides the continuous extension.

4.1 Primitive Objects

A game Γ generates, for each action profile x ∈ X:

- Operative private payoffs: u_i^P(x) $\in \mathbb{R}$ for each agent i = 1, …, n
- Aggregate private payoff: Π(x) := Σ_i ω_i u_i^P(x), where ω_i > 0
- System welfare: W(x) $\in \mathbb{R}$ (by Proposition 3.5, not recoverable from u^P)

Reference points:

Symbol	Name	Definition
W_0	System baseline	W at the disagreement/status quo point
W_C	Cooperative welfare maximum	max_{x∈X} W(x)
Π_C	Cooperative private payoff	Π(x_C) where x_C ∈ argmax W(x)
Π_P	Private maximum	max_{x∈X} Π(x) - maximum private payoff ignoring W
W_P	Welfare at private maximum	W(x_P) where x_P ∈ argmax Π(x)

Under PST: Π_C < Π_P and W_C > W_P. The private maximum and cooperative welfare maximum are structurally incompatible - the quantitative expression of the Private Pareto Trap.

4.2 The Private-Systemic Frontier

Definition 4.1 (Private-Systemic Frontier, PSF). The PSF is the upper boundary of the feasible set in (Π, W) space:

$$\Phi(\Pi) := \sup\{W(x) : x \in X,\ \Pi(x) = \Pi\}$$

The frontier $\mathcal{Pf}$ = {(Π, Φ(Π)) : Π ∈ [Π_C, Π_P]} is the set of outcomes maximizing system welfare for each level of aggregate private payoff.

Under PST: Φ'(Π) < 0 (the frontier slopes downward). Under convex damage functions: Φ''(Π) ≤ 0 (systemic losses accelerate with private extraction).

Proposition 4.1 (Frontier Concavity). In value-redirecting games (Proposition 3.11), Φ is strictly concave when the system's damage function h is convex. When h is linear, Φ is linear (constant marginal system cost). When h is concave, Φ is convex (decreasing marginal system cost - an unusual case where early private gains are more damaging than later ones).

Proof. In a value-redirecting game, W(σ) = W_0 - h($\Sigma\tau_i$(σ)) and Π(σ) = Π(d) + $\Sigma\tau_i$(σ). Along the frontier, $\Sigma\tau_i$ = Π - Π(d), so Φ(Π) = W_0 - h(Π - Π(d)), and Φ''(Π) = -h''(Π - Π(d)). This is strictly negative (concave Φ) when h is convex, zero (linear Φ) when h is linear, and strictly positive (convex Φ) when h is concave. ∎

Remark (Empirical relevance of concavity). The convex h case is empirically dominant in the documented cases: VW's health damages, LIBOR's mispricing, SVB's unrealized losses, and Bitcoin's carbon externalities all exhibit increasing marginal system costs as private extraction scales up. This is consistent with Proposition 4.1: h convex ↔ Φ concave ↔ the deepening trap. The linear case (constant marginal damage) serves as the benchmark. The concave h case (decreasing

marginal damage) may apply in some platform markets where early adoption generates significant positive network externalities before degradation dominates.

Structural analogy to portfolio theory. The PSF is the direct counterpart of Markowitz's (1952) efficient frontier: where the efficient frontier maps the optimal risk-return tradeoff in portfolio space, the PSF maps the optimal private-systemic tradeoff in outcome space. The critical difference: the efficient frontier slopes upward (more risk yields more return in equilibrium); the PSF slopes downward under PST (more private gain requires systemic sacrifice). This asymmetry is the quantitative content of the Private Pareto Trap.

4.3 System Beta (β_W)

Definition 4.2 (System Beta). The system beta of outcome x, evaluated on the PSF, is:

$$\boxed{\beta_W(x) := -\frac{dW}{d\Pi}\Big|_x = -\Phi'(\Pi(x))}$$

β_W > 0 measures the rate of systemic welfare loss per unit of private gain. It is directly analogous to the β coefficient in CAPM, with W playing the role of systematic welfare risk and Π playing the role of aggregate return.

β_W Range	Meaning	Empirical Example
β_W < 0	System-aligned: private gain increases system welfare	Open-source contributions; Montreal Protocol
β_W = 0	System-neutral; standard welfare analysis sufficient	Pure exchange, zero-sum games (benchmark null)
0 < β_W < 1	Mild PST	Regulated utility extraction
β_W = 1	Unit PST: one-for-one substitution	Benchmark case
β_W > 1	Severe PST: each unit Π destroys more than one unit W	Unregulated commons; benchmark manipulation

Stochastic version. Under uncertainty:

$$\beta_W^{\text{cov}}(x) := -\frac{\text{Cov}(\Delta W_x, \Delta\Pi_x)}{\text{Var}(\Delta\Pi_x)}$$

This is the regression slope of system welfare change on private payoff change - the covariance-based system beta, directly analogous to CAPM's regression-based β.

Practical estimation of β_W^cov. The covariance estimator has a direct empirical interpretation as the OLS slope in a regression of system welfare changes on private payoff changes across a panel of policy episodes: $\hat{\beta}_W = -\hat{\gamma}$ from the regression $\Delta W_{it} = \alpha_i + \gamma_t + \gamma\,\Delta\Pi_{it} + \epsilon_{it}$. For consistent estimation, the variation in ΔΠ must be driven by exogenous supply-side shifts - changes in cost structure, regulatory thresholds, or technology adoption - that affect private payoffs before feeding back into W. Suitable instruments include: (i) staggered entry of accounting rule changes (ASC 320 HTM exception for banking); (ii) jurisdiction-specific disclosure mandates (GDPR for platforms, Reg

FD for financial benchmarks); (iii) cost shocks from input price volatility. Standard errors should be clustered at the industry-year level to account for cross-sectional correlation in system welfare shocks. The key identifying assumption (parallel trends in the DiD) requires that W would have evolved similarly in treated and control environments absent the exogenous shock; this is testable via pre-trend analysis in the pre-policy window. The confidence interval around $\hat{\beta}_W$ provides the uncertainty range for subsequent T* calculations: if $\hat{\beta}_W = 3.1 \pm 0.8$ (as in the Bitcoin carbon channel), then the T* confidence interval inherits this uncertainty through the formula $T^* = \delta/\eta\lambda$, with $\lambda = \hat{\beta}_W \times d\Pi/dt$.

4.4 System-Adjusted Payoff (Π^SA)

Definition 4.3 (System-Adjusted Payoff). For shadow price μ > 0:

$$\boxed{\Pi^{SA}(x;\mu) := \Pi(x) - \mu\big(W_C - W(x)\big)}$$

Private gains are discounted by the welfare shortfall - the gap between achievable and actual system welfare, priced at rate μ.

In differential terms: dΠ^SA = dΠ + μ dW = dΠ(1 - μβ_W). Private gain is welfare-creating if and only if μβ_W < 1. When μβ_W > 1, the systemic cost of the marginal private gain exceeds its value.

Definition 4.4 (System Efficiency Ratio, SER) - suspended. The earlier draft defined `SER(σ) = W(σ) / [W_0 + Π(σ) · β_W^{−1}]`, but that expression is dimensionally inconsistent because β_W^{-1} converts welfare units into private-payoff units, not private payoff into welfare units. The SER statistic is therefore not used in this version. A corrected SER requires a separately derived welfare-denominated baseline and proof of its claimed [0,1] and endpoint properties.

The remaining SAPM instruments in this paper are β_W, μ*, Π^SA, Θ_H, and T*.

4.5 Shadow Price Duality

Proposition 4.2 (Duality). On the Private-Systemic Frontier, the efficient allocation for shadow price μ solves x*(μ) ∈ argmax_{x∈X} {Π(x) + μW(x)}. The first-order condition yields:

$$\boxed{\mu^* = \frac{1}{\beta_W(x^*)}}$$

The shadow price of system welfare on the efficient frontier equals the inverse of system beta.

Proof. At the optimum, ∂[Π + μW]/∂Π = 0 along the frontier, giving 1 + μΦ'(Π) = 0, hence μ = -1/Φ'(Π) = 1/β_W. ∎

Significance. This duality connects four objects - frontier slope, system beta, shadow price, and system-adjusted payoff - as the same entity in different coordinates. The optimal shadow price is not zero (laissez-faire) or infinity (system-first) but a computable function of the game's structure: μ* = 1/β_W. This is the central pricing result of the SAPM.

Worked numerical example. Consider an environmental PST game: a regulator must price the right to extract from a commons with baseline W_0 = 1,000 units. Observed data: private payoff at

full extraction Π_P = $100M, system welfare at full extraction W_P = 400 units; private payoff at cooperative baseline Π_C = $40M, W_C = 1,000 units.

Step 1: Estimate the PSF (assuming linearity for illustration). $\Phi(\Pi) = W_0 + (W_P - W_C)/(\Pi_P - \Pi_C) \times (\Pi - \Pi_C) = 1{,}000 + (400\text{-}1{,}000)/(100\text{-}40) \times (\Pi - 40) = 1{,}000 - 10 \times (\Pi - 40M)$.

Step 2: Compute β_W. β_W = -Φ'(Π) = 10 units of W per $1M of Π. In normalized terms: 10 welfare units / $1M private gain.

Step 3: Compute μ.* μ* = 1/β_W = $0.1M per welfare unit = $100,000 per welfare unit.

Step 4: System-adjusted payoff at current extraction Π = $80M. W($80M) = 1,000 - 10(80-40) = 600 units. Π^SA = 80M - (1/10)(1,000-600) = 80M - 40M = $40M. The system-adjusted payoff at current extraction is only $40M - equal to the cooperative baseline. Any additional extraction above Π_C has zero system-adjusted value at the efficient shadow price.

Policy implication. The extraction fee that achieves the system welfare target W = 1,000 while allowing some private gain is: set the fee per unit at μ* = $100,000 per welfare unit of damage. Any extraction Π > Π_C must pay (Π - Π_C) × β_W × μ* = (Π - Π_C) × 10 × $100,000 = (Π - Π_C) in private value terms. This exactly prices the private gain at its system cost - the efficient Pigouvian fee calibrated to the SAPM frontier.

4.6 Crossover Dynamics

Definition 4.5 (Crossover Time T).* Let δ be the stock of private surplus above the cooperative baseline over the relevant episode, η the feedback coupling between private gain and system damage, and λ the private-payoff-equivalent system damage flow per unit time. The units requirement is essential: δ must be a stock and ηλ must be a stock-loss flow per unit time. Define:

$$\boxed{T^* = \frac{\delta}{\eta\lambda}}$$

T* is the time at which cumulative systemic damage exceeds the system's ability to self-repair. For t < T*, the Suboptimal Optimum is intact - private gains exceed systemic costs even accounting for feedback. For t > T*, the Suboptimal Optimum has converted to Crash.

Proof of T characterization.* Under constant linear extraction, system damage in private-payoff-equivalent units accumulates at ηλ per unit time. The private payoff advantage is Π_t − Π^ref = δ − ηλt. Setting Π_{T*} = Π^ref gives T* = δ/(ηλ). ∎

The Theta statistic. For rapid horizon assessment:

$$\Theta_H := \delta - \eta\lambda H$$

- Θ_H > 0: private advantage survives planning horizon H
- Θ_H = 0: knife-edge crossover
- Θ_H < 0: Suboptimal Optimum converts within the planning horizon

Parameter η and the suppression mechanism. In every documented case (§5), the mechanism through which Suboptimal Optimum is sustained is the suppression of η: the feedback coupling

between private gains and system damage is artificially depressed below its natural level. VW's defeat device suppressed η from ~0.7 to ~0.3, extending T* from ~2.6 years to 6.1 years. SVB's HTM accounting suppressed η from ~0.7 to ~0.1, extending T* from ~2 years to ~6 years. This pattern - artificial η suppression as the technical instrument of Suboptimal Optimum - is the empirical signature of value-redirecting games.

A note on degrees of freedom. The formula $T^* = \delta/\eta\lambda$ involves three parameters (δ, η, λ), which raises a legitimate concern: three free parameters can potentially rationalize any observed duration ex post. Three responses address why this is not fatal to the framework. First, each parameter has an independent economic meaning that constrains its range prior to calibration: δ is the private surplus above the cooperative baseline (observable from accounting records and settlement data), η is the feedback coupling between private gains and system damage (estimable from the delay between actions and regulatory or market response), and λ is the system's degradation rate per unit of extraction (measurable from public regulatory filings, environmental data, or financial disclosures). They are not interchangeable - changing one while holding the others fixed makes distinct economic predictions about timing, mechanism, and reversibility. Second, the prospective validation protocol (§6.8) imposes ex ante constraints: for the active US casualty reserve case, all three parameters are estimated from 2024 Schedule P statutory filings before the resolution is known, ruling out ex post fitting by construction. Third, the VW sensitivity table (§5.2) demonstrates that T* is robust to ±30% variation in individual parameters: holding δ = $5.5B fixed, varying η from 0.2 to 0.5 and λ from $2.5B to $3.5B per year yields $T^* \in [3.1, 11.0]$ years, with the baseline estimate (T* = 6.1 years, η = 0.3) closest to the observed ~6 years. The three-parameter structure is identified by the intersection of economic interpretability, ex ante constraint, and cross-parameter robustness - not by post-hoc curve fitting.

4.7 The Benchmark Null

When does SAPM reduce to standard welfare analysis? The answer is structurally precise. The SAPM reduces to standard welfare analysis when systemic damage is zero: when the damage rate λ = 0 (the system is not degrading, even though private extraction is occurring). In this case $T^* = \infty$ and the system-adjusted payoff Π^SA = Π for any finite μ: no welfare discount applies, and the SAPM predicts the same allocations as standard welfare maximization.

A related but distinct condition is β_W = 0: the frontier is flat (dW/dΠ = 0), meaning private gains and system welfare are independent - no PST exists. This holds when systemic feedback is fully internalized (η = 1, all damages immediately borne by extracting parties, so λ = 0 from the agents' perspective) or when there is genuinely no systemic coupling (Axiom 3 fails). When β_W = 0, the welfare constraint is never binding: W_C = W_P = W_0 throughout, and the welfare shortfall (W_C - W(x)) is identically zero. The shadow price formula μ* = 1/β_W is not well-defined in this limit - the formula applies to PST games where β_W > 0. In the non-PST degenerate case, the system-adjusted payoff collapses to Π^SA = Π because the correction term μ·(W_C - W) = 0 regardless of μ: Integration is achieved at zero private cost, and no shadow price is needed. This is the benchmark null.

The practical benchmark null is therefore: SAPM ≡ standard welfare analysis when λ = 0 (no degradation). SAPM adds structure when λ > 0 and η < 1 - the empirically prevalent case.

This establishes that SAPM is a strict generalization, not a replacement, of standard welfare analysis. The standard framework is the SAPM evaluated at the degenerate parameter point (λ = 0

or β_W = 0). For all parameter values away from this point - which characterizes every empirical case in §5 - the SAPM generates different predictions and different policy implications.

4.8 Exponential Dynamics

For more realistic settings where private payoff advantage decays and system resilience is non-negligible, the linear formula is no longer automatic. A coupled model may take the form

$$\dot{W}_t = \rho(W^* - W_t) - \lambda\Pi_t$$

where ρ > 0 is the system's resilience rate (endogenous regeneration capacity) and W* is the healthy-state attractor. The closed-form private-payoff path previously displayed here is not valid for the full coupled system without extra decoupling assumptions. Under the special benchmark $W^*=W_0$ with constant unit extraction and exogenous private-surplus decay, the private payoff advantage can be approximated by:

$$\Pi_t - \Pi^{\text{ref}} = \delta e^{-\alpha t} - \frac{\eta\lambda}{\rho}(1 - e^{-\rho t})$$

The crossover time in the exponential regime solves:

$$\delta e^{-\alpha T^*} = \frac{\eta\lambda}{\rho}(1 - e^{-\rho T^*})$$

In the limit ρ → 0 (negligible resilience) and α → 0 (stable private surplus), this reduces to the linear formula T* = δ/ηλ. In the coupled system, resilience generally delays crossover by offsetting accumulated damage; the exact crossover equation depends on the specified dynamics. This subsection is therefore a model extension, not an input to the main theorem or the empirical T* claims. The paper relies on the linear T* core unless a domain-specific dynamic model is explicitly stated.

4.9 Figure 1: Trajectory Dynamics and the Private-Systemic Frontier

Figure 1: The Private Pareto Trap - Trajectory Dynamics

Figure 1 (two-panel diagram; final rendered version to accompany submission):

Panel A (Time series, t ∈ [0, T + 3]).* Two curves on the same axes. The x-axis is time t. The y-axis is normalized payoff index (0 = baseline).

- Private payoff Π_t (solid line, blue): begins at δ > 0 above the reference line at t=0. Remains elevated through t = T*. Crosses the reference line at T* (the crossover). Continues declining below reference line (the trap has closed).

- System welfare W_t (dashed line, red): begins at W_0 at t=0. Declines monotonically throughout, passing below the "irreversibility threshold" at T*. The decline is concave (accelerating, consistent with Proposition 4.1).

- A vertical dotted line at t = T* labelled "Crossover (T*)."

- A grey shaded region t > T* labelled “Crash Zone: Π_t < Π^ref, W_t < W_irreversible.”

- A star or marker on the time axis at t_observed (observed crisis event, e.g., VW = t=6, LIBOR = t=7). If the model is correct, t_observed ≈ T*.

Panel B (The Private-Systemic Frontier). X-axis: aggregate private payoff Π. Y-axis: system welfare W. A concave downward-sloping curve (the PSF, Φ(Π)) from top-left (Π_C, W_C) to bottom-right (Π_P, W_P). A point labelled “Integration zone” near (Π_C, W_C). A point labelled “Suboptimal Optimum zone” along the frontier below W_0. A horizontal dashed line at W_0 (baseline). A vertical dashed line at Π^ref. The region above the horizontal line and to the right of the vertical line - upper-right quadrant - is empty under PST-G within the private-improvement region S(d) (the Private Pareto Trap: within S(d) you cannot be at Π > Π^ref AND W > W_0 simultaneously). The point (Π_P, W_P) is labelled “Private optimum (Crash trajectory).”

Caption: Panel A shows the trajectory dynamics of a canonical PST-G game. Private payoff Π_t (blue) initially exceeds the cooperative baseline, masking accelerating system welfare degradation (red). At T* = δ/ηλ, the private advantage is extinguished; for t > T*, the Suboptimal Optimum has converted to Crash. The star marks the empirically observed crisis event for the VW case (t ≈ 6 years). Panel B shows the Private-Systemic Frontier: the concave downward-sloping boundary mapping aggregate private payoff to maximum achievable system welfare. Under PST-G, the upper-right quadrant (high Π AND high W simultaneously) is structurally inaccessible within the private-improvement region S(d). The frontier formalizes the Private Pareto Trap geometrically.

4.10 The Classification Algorithm

Five observable inputs map to three binary indicators and eight structural outcome types:

Input	Symbol	Description
Private surplus	δ	Π - Π^ref
System gap	g	W - W^ref
Degradation rate	λ	Private-payoff-equivalent system damage flow per unit time (Definition 4.5); dollars per year, not a ratio of levels
Feedback coupling	η	Dimensionless fraction of system damage fed back into participants’ operative payoffs, η ∈ [0,1] (Definition 4.5)
Planning horizon	H	Evaluation timeframe

With δ a stock, η a dimensionless fraction in [0,1], and λ a flow in private-payoff-equivalent units per year, T* = δ/(ηλ) is a duration, as Definition 4.5 requires.

Three binary indicators:

1. Private sign: p = sign(δ)

2. System sign: s = sign(g)

3. Durability sign: d = sign(T* - H)

p	s	d	Outcome Name	Recommended Action
+	+	+	Integration (durable)	Sustain
+	+	-	Integration (fragile)	Monitor; strengthen buffers
+	-	+	Suboptimal Optimum (slow)	Intervene before T*
+	-	-	Suboptimal Optimum (fast)	Immediate intervention
-	+	+	A Compromised (durable)	Institutionalize
-	+	-	A Compromised (fragile)	Patience; investment pays off
-	-	+	Crash (slow)	Fundamental restructure
-	-	-	Crash (acute)	Emergency intervention

---

5. Empirical Applications

Table 1: Cross-Case SAPM Summary (Eight Empirical Cases)

Case	β_W	T* (predicted)	T* (observed)	Status	Section
AI Agent Collusion	2–5	Qualitative: months–quarters; T* not yet calibrated	Active	Live (2025–)	§5.1
VW Dieselgate	~5.5	6.1 years	~6 years ✓	Illustrative (retroductive)	§5.2
2023 Bank Failures	3–17 per bank; ≈8–10 aggregate	1–2 quarters early	4/4 ✓	Illustrative (retroductive)	§5.3
LIBOR Manipulation	3.3–10	Held pending stock/flow recalibration	7 years observed; η-suppression mechanism retained	Illustrative (retroductive)	§5.4
Insurance Reserves	3–8 est.	8.6 yr central; 1.4–25.8 yr propagated; 3–9 yr operational window (judgment)	TBD (2027–2033)	Live / Prospective	§5.5
Bitcoin Ecosystem	~2.0 [1.5–2.0]	N/A (ongoing)	Ongoing	Illustrative	§5.6

Case	β_W	T* (predicted)	T* (observed)	Status	Section
Meta / Facebook	0.3–0.6	50–100 years at current η; 12.5–25 under enforcement η	TBD	Prospective	§5.7
US CRE (“Extend and Pretend”)	~1.8 (order-of-magnitude est.)	~8–17 years (η-dependent; anchored March 2026)	TBD	Live / Prospective	§5.8

Reading guide. β_W measures systemic welfare loss per unit of private gain (β_W > 1 = severe PST). T* is the predicted crossover from Suboptimal Optimum to Crash. Cases marked ✓ match predicted and observed T* in the labeled retrodictive or prospective validation window; prospective tests are separately specified below.

Illustration versus prediction (stated once, applied throughout §5). Three cases — VW, the 2023 banks, LIBOR — are illustrative retroductions: parameters estimated ex post from records unavailable before disclosure, so they demonstrate structural coherence and parameter interpretability, not prospective power. Two cases — US casualty reserves (§5.5 (Case 7c)) and US CRE (§5.8) — are genuinely prospective: parameters fixed from public data before the outcome, predictions falsifiable within 36 months from Schedule P and FDIC data (§6.8). Each case carries its tag; the distinction is not re-argued per case.

η suppression (the shared mechanism, stated once). Four cases run on one mechanism: a reported metric diverges from reality by a factor that grows over time, suppressing the feedback coupling η below its natural level (0.7–0.9) and extending $T^* = \delta/(\eta\lambda)$ far beyond what the underlying PST dynamics permit. The suppressor differs by domain — VW’s defeat device, SVB’s HTM accounting, LIBOR’s multi-bank coordination, insurance’s statutory reserve smoothing — but the formal action is identical. When the suppressor exhausts, η phase-transitions toward 1.0 in days, T* expires, and the crash arrives. Each case names its suppressor; the mechanism is not re-derived.

5.0 Measurement Specification

Each parameter requires specific empirical inputs. The tables below specify the measurement methodology per calibrated or prospectively specified case, enabling independent replication.

VW Dieselgate (§5.2)

Parameter	Input Data	Derivation	Value
δ	Avoided emissions-compliance costs across 11M vehicles	VW internal documents; settlement record	~$5.5B
λ	Health damage rate during fraud	Annual excess NOx × social cost of pollution (EPA)	~$3B/year
η	Defeat device feedback suppression	NHTSA penalty trajectory vs. full-information counterfactual (natural η 0.7-0.9)	~0.3 (pre-discovery), ~0.95 (post-discovery)
β_W	Total system cost / total private gain	$40B system costs (fines + health + regulatory rebuild) / $26.7B private gains (cost avoidance + sales premium)	~1.5 [aggregate; see §5.2 for marginal β_W = 5.5 at the PSF frontier]
T*	δ/(ηλ)	$5.5B / (0.3 × $3B/yr)	~6.1 years (observed: ~6 years)

2023 Bank Failures (§5.3)

Parameter	Input Data	Derivation	Value
Kill Metric 1	Unrealized losses / total equity	FDIC call reports, filed quarterly	SVB: 89%, Signature: 45%, First Republic: 52%
Kill Metric 2	Uninsured deposit concentration	% of deposits above $250K FDIC limit	SVB: 94%, Signature: 90%, First Republic: 68%
Kill Metric 3	Quarterly deposit decline rate	Q-over-Q change in total deposits	SVB: -8.5% Q4 2022, Signature: -12% Q4 2022
η suppression	HTM accounting classification	HTM unrealized losses excluded from regulatory capital	η ≈ 0.05 (losses invisible until forced sale)

η suppressor: HTM (held-to-maturity) accounting carries securities at amortized cost regardless of market value. A liquidity event forcing HTM→AFS reclassification crystallizes losses instantly. All four failed banks ran a long suppressed-η period, then an abrupt η → 1.0 transition (bank run), then failure within days.

LIBOR Manipulation (§5.4)

Parameter	Input Data	Derivation	Value
δ	Benchmark integrity buffer	Requires stock calibration over a specified horizon	Held
λ	Distortion/damage flow	Moving avg of |submission − transaction-weighted mean|	Held
η	Multi-bank coordination factor	Coordinating banks / panel size × 1/detection probability	Qualitatively low; not used for T* until recalibration
β_W	Market distortion / cartel profit	$50-100B cumulative mispricing damage / ~$10-15B cartel profits	~3.3-10
T*	δ/(ηλ)	Calibration held: δ is currently stated as a flow, so the quotient is not a duration until δ is restated as a stock over a specified horizon	—

η suppressor: when sixteen panel banks submit coordinated false rates, each bank's deviation hides behind the others' similar deviations, so the median submission appears normal. The T* formula applies per-agent; the seven-year observed duration reflects coordinated η suppression across the panel.

Insurance Reserve Inadequacy (§5.5)

Parameter	Input Data	Derivation	Value
δ	Excess reserves above regulatory minimum	Carried reserves − actuarially indicated (Schedule P)	~$9-62B (range = actuarial uncertainty)
λ	Annual adverse development rate	Calendar-year prior-period adverse development (Schedule P); unified with the §5.5 range (the calendar-2025 figure of $7.3B falls below the $8B confirmation threshold, in the indeterminate band)	$8-16B/year, accelerating

Parameter	Input Data	Derivation	Value
η	Statutory accounting transparency	Suppressed pre-2024 baseline ~0.10-0.20 (inverse of the 5-10 yr loss-occurrence-to-recognition delay); the 2024 adverse-development signal moved the operative value to the §5.5 calibration	0.3-0.4, rising (post-signal value used for T*)
β_W	System cost / premium inadequacy	Policyholder harm from underfunding / insurer gains from underpricing	~3-8 (est.)
T*	δ/(ηλ)	Point: $30B/(0.35 × $10B/yr) ≈ 8.6 yr; propagating the full stated ranges (δ $9-62B, η 0.3-0.4, λ $8-16B/yr) gives 1.4-25.8 yr; at the pre-2024 baseline η (0.10-0.20) the extremal span is 2.8-77.5 yr with a central estimate near 20 yr	8.6 yr central; 3-9 yr operational window (judgment)

η suppressor: Schedule P conventions let insurers recognize losses over multi-year development periods; workers'-comp favorable development ($5.5B in 2025) masks adverse casualty development in headline figures. When the masking reserves exhaust — now underway — reserve inadequacy surfaces abruptly.

5.1 Emergent AI Agent Collusion — The Frontier Case

The AI cartel results of Mazur (2025), Dou, Goldstein, & Ji (2025), and Fish, Gonczarowski, & Shorrer (2025) are this paper's cleanest case: they eliminate every non-structural explanation. The agents had no greed, no career incentives, no corner-cutting culture. They had optimization functions.

The Folk Theorem delivers exactly what it promises: patient agents in a repeated competitive environment cooperate. What it cannot say is whether the cooperation is Suboptimal Optimum or Integration. Standard analysis calls it optimal. SAPM classifies it:

Variable	Value
A (seller agents)	1 — profits above competitive level
B (cooperating sellers)	1 — same
C (market efficiency, buyer welfare)	0 — supra-competitive pricing, market failure
Classification	Suboptimal Optimum (0,1,1)

System beta runs β_W ≈ 2-5 in the Mazur (2025) experiments: β_W ≈ (buyer welfare loss) / (seller profit gain above competitive baseline), with seller profit elevation of 20-50% and buyer welfare suppression of 40-100%. Mazur's supracompetitive prices of 60-80% above competitive equilibrium imply per-unit consumer surplus destruction exceeding per-unit seller gain by 2-4×. The finding replicates structurally: Lin et al. (2024) found Cournot collusion in LLM-mediated commodity markets; Fish, Gonczarowski, & Shorrer (2025) demonstrated LLM algorithmic collusion at the 2025 AEA meetings; Agrawal et al. (2025) replicated collusive double-auction outcomes with LLM agents; Hammond et al. (2025) surveyed the multi-agent risk landscape. Institutional governance proposals (Bracale Syrnikov et al., 2026) treat collusion suppression as first-order precisely because standard oversight is W-blind. The UK CMA (2026) flagged AI agent collusion as a frontier enforcement challenge: standard anticompetitive-agreement frameworks need additional structure to handle coordination without explicit communication.

Robustness: the fragility objection. Keppo, Li, Tsoukalas, & Yuan (2025) show that heterogeneous LLM agents — different models, different data — collude fragilely: diversity functions as a defection mechanism. In SAPM terms, heterogeneity raises η, the coupling between any agent's deviation and equilibrium breakdown. Higher η means shorter T* and a less stable trap. Keppo et al. therefore confirm the framework: when the agent population is diverse enough, effective η rises above the threshold at which the scheme stays privately sustainable. The regulatory implication: mandating diversity in LLM deployments is a structural η-raising intervention.

Implication for AI governance. An optimizer whose objective is defined exclusively on operative private payoffs, operating in a repeated PST environment, can support Suboptimal Optimum equilibria; which equilibrium it reaches depends on learning, selection, institutional constraints, and market dynamics. Keeping AI systems in Integration requires C as a formal variable in the objective function. This is not soft ethics. It follows from the classification result: any objective defined only on (u_1^P, …, u_n^P) is W-blind and by construction cannot distinguish Integration from Suboptimal Optimum.

5.2 Volkswagen Dieselgate — Illustrative Retroduction (T* = 6.1 Years)

From 2009 to 2015, Volkswagen installed defeat devices in approximately 11 million diesel vehicles, activating full emissions controls only during testing. Under normal driving, nitrogen oxide emissions exceeded legal limits by up to 40×. The West Virginia University Center for Alternative Fuels, Engines, and Emissions discovered the fraud, triggering an EPA Notice of Violation in September 2015.

SAPM Parameters:

Parameter	Value	Source
δ (private surplus)	~$5.5B (avoided emissions equipment across 11M vehicles)	VW internal documents
System welfare loss	~$30B (health externalities, regulatory integrity, consumer trust)	EPA settlement; litigation record
β_W	30B/5.5B ≈ 5.5	Marginal β_W at the private-systemic frontier (slope of system loss per unit private gain at the Pareto outcome); distinct from the §5.0 aggregate ratio of 1.5 (total system cost / total private gain)
η (feedback coupling)	0.3 — defeat device suppresses feedback; regulatory credibility erodes slowly	NHTSA penalty trajectory
λ (damage rate)	~$3B/year	Health damage costs during fraud

Crossover time:

$$T^* = \frac{\delta}{\eta\lambda} = \frac{5.5\text{B}}{0.3 \times 3\text{B}} = 6.1 \text{ years}$$

Empirical result. Defeat device deployed 2009. EPA Notice of Violation September 18, 2015. Elapsed: ~6 years. ✓

This is a retroduction: δ, η, λ come from VW internal documents, EPA settlement data, and NHTSA records — none public before September 18, 2015. The exercise demonstrates structural coherence: the defeat device maps precisely onto η suppression. The scientific test of prospective power lies in the 2023 banks and the casualty reserve case, where parameters were fixed before the outcome.

Sensitivity. T* across the plausible parameter range:

η	λ	δ	T* = δ/ηλ
0.2 (defeat device active)	$2.5B/year	$5.5B	11.0 years
0.3 (baseline)	$3.0B/year	$5.5B	6.1 years
0.4	$3.0B/year	$5.5B	4.6 years
0.5	$3.5B/year	$5.5B	3.1 years
0.7-0.9 (natural η)	$3.0-3.5B/year	$5.5B	1.7-2.6 years

The baseline (η = 0.3, T* = 6.1) matches the observed ~6 years best. The table quantifies the suppression: the defeat device pushed η from its natural 0.7-0.9 (T* = 1.7-2.6 years) to 0.3 (T* =

6.1 years), buying ~3-4 additional years of undetected operation. $T^* \in [3.1, 11.0]$ years across the range; 6.1 is the $\eta = 0.3$ baseline, not a one-decimal measurement.

Classification. $\delta > 0$; $g < 0$ (regulatory and environmental system degraded); $T^* = 6.1$ years. Sub-optimal Optimum (fast) → Crash at $t \approx 6$ years. Final regulatory settlement: $25B globally, ~4.5× the original private surplus. System-adjusted payoff at full welfare pricing ($\mu = 1.0$): $\Pi^{SA} = 5.5B - 1.0 \times 30B = -\$24.5B$.

5.3 The 2023 Bank Failures — Illustrative Retroduction (4/4, Public Data Only)

Can SAPM identify T* before market consensus from public data? I apply it retrospectively using only public regulatory filings (10-Q, 10-K, FFIEC Call Reports) through Q4 2022.

The bilateral fiction structure. Each failure follows an identical structure. Management and investors form the "A and B" dyad, extracting private value through extend-and-pretend accounting while the financial system (depositors, FDIC, payment infrastructure) bears the systemic cost — the Suboptimal Optimum signature: both private parties gain from the fiction while bank solvency, depositor protection, and financial stability degrade.

Observable kill metrics. Three ratios predict T* across all four cases:

1. Unrealized losses / Total equity > 85%
2. Uninsured deposits / Total deposits > 85%
3. Quarterly deposit decline rate > 10%

Any two of three flashing signals T* within 2-3 quarters.

Four cases:

Bank	δ (Hidden Surplus)	β_W	T* Identifiable	Market Missed By
SVB	$15.1B (HTM losses)	16.6 [$251B system disruption ($175B deposits mobilized + $76B emergency Fed/FDIC intervention apportioned to SVB) / $15.1B private surplus]	Q3 2022 (10-Q; confirmed Q4 10-K)	2 quarters
Signature	$8-10B (crypto deposits)	11-14	Q4 2022	~1 quarter
First Republic	$19-27B (loan markdowns)	10-15	Q4 2022	~1 quarter
Silvergate	$5-8B (franchise value)	3-5	Q3 2022	1-2 quarters

SVB detail. HTM accounting (FASB ASC 320) is the η suppressor in SVB and First Republic,

an artificial feedback dam that hid the degrading balance sheet from depositors. Per the Federal Reserve OIG (Report 2023-SR-B-013), as of September 30, 2022, SVB's total unrealized losses reached 110% of capital — both the unrealized-losses/equity metric (>85%) and the uninsured-deposits/total-deposits metric (94%) breached simultaneously. By year-end 2022 (Q4 10-K), SVB disclosed $15.2B HTM and $2.5B AFS unrealized losses — $17.7B total against $16.8B equity — confirming a persistent, public signal. T* was structurally identifiable at least two quarters before the March 8, 2023 stock offering that triggered the η transition from ~0.1 to ~0.95 in 48 hours. The HTM exception is formally equivalent to the VW defeat device.

Aggregate validation. Total disruption across four failures: $500B+ deposits mobilized for emergency guarantees, two emergency Federal Reserve facilities (BTFP and emergency lending window), and revision of the $250B asset threshold (FDIC, 2023; Federal Reserve Board, 2023). Total hidden private surplus across four banks: ~$50-60B. Aggregate $\beta_W \approx 8\text{-}10$, confirming $\beta_W \gg 1$ in value-redirecting financial games.

SAPM identified T* in all four cases one to two quarters before market consensus from public filings alone. ✓ "One to two quarters before" means T* identifiable from public regulatory filings before the actual failures (March-May 2023). For SVB, kill metrics breached Q3 2022 (September 30 10-Q), two full quarters before failure — a conservative claim. For the other three, T* is identifiable from Q4 2022 filings submitted 45-60 days before failure. The claim asserts only that structural signals were publicly present before failure, not that a real-time implementation would have triggered intervention. Parameters come from data available before the outcome: a look-ahead-bias-free retrospective.

5.4 LIBOR Manipulation — Structural Dynamics (T* Calibration Held)

At least 16 major banks manipulated LIBOR — the benchmark underpinning approximately $350 trillion in notional derivatives, mortgages, and consumer loans — from as early as 2003 (DOJ settlement records; UK FSA Final Notice, 2012), with documented written rate-movement requests at UBS from at least January 2005 through 2012. The Wall Street Journal first reported the scheme in May 2008. Enforcement fines exceeded $9 billion across jurisdictions through 2016. LIBOR was phased out by June 2023, replaced by risk-free rates including SOFR (Jermann, 2024).

SAPM Parameters:

Parameter	Value
δ	~$1.5B/year in reduced borrowing costs and trading gains (author's estimate from DOJ settlement records; per-bank benefits ranged from hundreds of millions to over $1B annually; $1.5B is a conservative aggregate for the major submitters). This is a flow, not the stock δ required by T*.
System welfare loss	~$50-100B cumulative mispricing damage
β_W	50-100B / 10-15B ≈ 3.3-10
η	Not calibrated for T* in this version; qualitative claim only: multi-bank coordination suppressed feedback.

Parameter	Value
λ	~$5B/year in system integrity erosion, a flow that cannot be paired with the current δ flow to produce a duration.

Crossover time: held pending recalibration. The earlier draft computed

$$T^* = \frac{1.5\text{B/year}}{0.5 \times 5\text{B/year}} = 0.6,$$

but that quotient is dimensionless, not years. To produce a duration, δ must be restated as a stock of private surplus over a specified decision horizon or λ must be restated in compatible stock terms.

The structural interpretation retained here is narrower: LIBOR is a high-β_W case with multi-bank η suppression. The seven-year duration (2005-2012) is evidence that coordination suppressed feedback, because when sixteen banks submit coordinated false rates, no single bank's deviation is detectable against the corrupted benchmark itself. The numeric T* calibration is not used until the stock/flow repair is completed.

System damage accumulated at ~3.3× the rate of private gains per quarter. This is why continuous sixteen-bank coordination was necessary: the individual T* was too short to sustain alone. Each bank's reporting created a bilateral commitment device — defecting would expose all participants while providing no advantage, since the defecting bank's true rate would be an outlier against sixteen manipulated submissions.

Classification. Suboptimal Optimum throughout. UBS was the first major LIBOR bank to approach UK regulators (FSA Final Notice, December 2012). The qualitative SAPM mechanism is that once any participant's regulatory exposure crosses its private coordination horizon, the dominant strategy switches from coordination to cooperation with enforcement. UBS's defection converted a negative-expected-value position (continued fraud exposure) into a positive one (cooperation credit, reduced penalties, deferred prosecution). The event supports the η-suppression narrative, but the earlier 0.6-year T* number is not retained.

5.5 Insurance Reserve Inadequacy — The Longest-Running Bilateral Fiction

Property-casualty insurance is the most data-rich SAPM application outside banking, and historically the longest-running documented Suboptimal Optimum. Three historical cases show the bilateral fiction — "reserves are adequate" — persisting for decades, with β_W comparable to the 2023 banks and far longer duration. A fourth, active case offers a prospective test.

Structural template. Management and equity holders form the "A and B" dyad, booking inadequate reserves that suppress reported losses and maintain premium revenue. Future claimants — policyholders, injured third parties, the social systems absorbing uncompensated costs — are system C. Both private parties gain while solvency and policyholder protection degrade invisibly inside statutory filings. The η suppressor here is perpetual (see §5.0): statutory smoothing, occurrence-to-claim latency, and coverage litigation that delays a legally definitive C=0 for years.

Case 7a: Asbestos and Environmental Reserve Inadequacy (1970s-2024)

The canonical case. Insurers wrote 1940s-1970s policies covering "bodily injury" and "property damage" with no asbestos exclusion. Initial early-1980s reserves: ~$2-5B industry-wide. AM Best's 2024 net aggregate exposure estimate: $146B. The gap defines δ.

Parameter	Value	Source
δ (peak, ~1985-1995)	$80-100B	Initial reserves vs. ultimate exposure
λ	$5-10B/year	Annual adverse development, 1980s-2000s
η (initial)	0.05-0.10	Near-zero until Manville bankruptcy (1985)
T* at initial η	δ/(ηλ) = $80B/(0.08 × $7B) ≈ 143 years	Theoretically unsustainable for generations
T^*_1 observed (first recognition)	1985 (Manville) — ≈15 years from policy origin	η transition to ~0.2
β_W	2.5-3.0	~$200-250B total system cost ($146B AM Best net exposure + social, health, judicial costs) / $80-100B private surplus. AM Best net exposure alone yields β_W ≈ 1.5-1.8

At initial η ≈ 0.05, T* = 143 years — which is why asbestos is still running off 50 years later. A series of external η shocks followed: Manville (1985) → Lloyd's near-collapse (1992) → GE Capital $15B charge (2018) → $14.5B still in runoff as of 2024. Each shock raised η temporarily, triggered a partial recognition wave, then subsided as new reserves were set and extend-and-pretend resumed at lower δ. Multi-wave T* is the structural signature of insurance extend-and-pretend.

Case 7b: GE Capital Long-Term Care Insurance (2006-2018)

GE placed its long-term-care portfolio in runoff in 2006. The reserve gap accumulated silently through 2017 while GE reported the portfolio as stable.

Parameter	Value
δ (hidden shortfall)	~$15B (required statutory contributions)
λ	~$1.5B/year (medical inflation + lapse-rate divergence)
η (2006-2016)	≈ 0.05 — analysts accepted GE's "stable" characterization
T* predicted	$15B / (0.05 × $1.5B) = 200 years at initial η
T* observed	12 years (2006 → January 2018, $6.2B after-tax GAAP charge)
η-transition trigger	New CEO John Flannery ordered comprehensive portfolio review (external shock)
β_W	5-8 (including GE market-cap impact from confidence collapse)

The 200-year prediction at initial η = 0.05 explains why the fiction lasted 12 years and required an external trigger. Insight for auditors: a portfolio with η ≈ 0 is not stable; it is a time bomb. $T^* = \delta/\eta\lambda$ diverges as $\eta \to 0$ — not a model defect but a correct prediction: a fully feedback-suppressed Suboptimal Optimum never self-corrects. It ends only when an external agent shocks η.

Case 7c: Current US Casualty Liability Reserve Inadequacy (2024-2026, active) — Prospective Test

After 17 consecutive years of favorable prior-year development (2006-2022), US property-casualty insurers reported $7.8B in adverse prior-year development in 2024 — more than double the 2023 figure of $3.7B. This is the first η-transition signal.

Parameter	Estimate	Source
δ (reserve deficiency)	$9-62B by scope	AM Best (2025); Swiss Re Institute
λ	$8-16B/year, accelerating	135 nuclear verdicts in 2024 (+52% YoY); $31.3B total awards (+116% YoY); median $51M; 5 verdicts exceeding $1B (Marathon Strategies, 2025); S&P GMI: $9.98B adverse development in other-liability-occurrence in 2024; verdicts exceeding $10M more than quadrupled 2020-2024 (Risk & Insurance, 2026)
η	0.3-0.4	Rising but moderate — market treating 2024 as cyclical blip
T*	Point: $30B / (0.35 × $10B/yr) ≈ 8.6 years (≈2033 from 2024 data); propagated range 1.4-25.8 years	Full propagation of the stated ranges; the 3-9 year operational window is a judgment within that span (see below); accelerating λ could compress to 3-5 years
β_W	3-8	Reserve inadequacy triggers premium spikes, insolvencies, uninsured social costs

Current η diagnosis. Swiss Re Institute reports $62B cumulative adverse development for commercial liability lines over 2015-2024. The 2024 statutory cycle added $16B to prior years' liability loss estimates — the highest in a decade, reaching back to accident year 2015. S&P Global Market Intelligence places other-liability-occurrence adverse development at $9.98B for 2024, nearly double the prior year. The driving verdict trend accelerates: 135 verdicts in 2024 (+52% YoY), $31.3B total awards (+116%), median $51M, five exceeding $1B (Marathon Strategies, 2025); verdicts exceeding $10M more than quadrupled 2020-2024 (Risk & Insurance, 2026). The market still calls this cyclical and has not repriced insurance equities for structural inadequacy. η ≈ 0.3-0.4

places the market in the early T*-approach zone — the same zone SVB occupied in Q3 2022, two quarters before its η transition.

Insurance kill metrics (analogous to §5.3):

1. Adverse prior-year development / Net premiums written > 3% — reversion from the 17-year favorable trend is structural, not random
2. Nuclear verdict count YoY growth > 30% — signals accelerating λ
3. Reserve-to-surplus ratio deterioration across consecutive accident years — Schedule P triangle analysis; the banking analog is unrealized losses / equity

Any two of three active signals T* within 3-5 years under current estimates. Current status: of the three metrics, only nuclear-verdict growth (+52% YoY versus the 30% threshold) is demonstrably breached from data reported here; metric 1 awaits a stated net-premiums-written denominator, and metric 3's Schedule P reserve-to-surplus series is not yet compiled — 1 of 3 clean, so the case is an open prospective test, not a confirmed one.

Three 2025 indicators support the accelerating-λ diagnosis. First, the other-liability-occurrence combined ratio deteriorated to 109.3% in 2024 from 99.6% in 2023 (Reinsurance News, August 2025) — above 100% means the system already pays more in claims than it collects in premiums on the lines most exposed to verdict inflation, an underwriting loss in the current period, not a leading indicator. Second, corporate verdicts exceeding $100 million have grown 400% over the past decade (US Chamber Institute for Legal Reform; Marsh McLennan CEO remarks, Insurance Journal, July 2025). Third, "thermonuclear verdicts" exceeding $100 million reached 49 in 2024 versus fewer than 15 in 2020 (Marathon Strategies, 2025). Each verdict resets prior-year reserve development upward across all open accident years, compounding the booked-to-ultimate gap.

This is the paper's prospective test. Unlike VW (retroactive match) and LIBOR (structural interpretation), the casualty reserve case is the only one where SAPM parameters come from public 2024 statutory filings before the resolution is known. Point estimate: $T^* = \delta/(\eta\lambda)$ = \$30B / (0.35 × \$10B/yr) ≈ 8.6 years from 2024 data (≈2033). Propagating the full stated ranges (δ \$9-62B, η 0.3-0.4, λ \$8-16B/yr) yields 1.4-25.8 years. The 3-9 year operational window (2027-2033) used for the prospective test is a judgment call within that span, conditioned on η currently at 0.3-0.4 and rising with accelerating λ, which compresses the upper tail — it is not a derivation, and it is labeled as such. A second calibration of η from the structural recognition delay (the suppressed pre-2024 baseline of 0.10-0.20 shown in §5.0) would put the central estimate near 20 years; this paper interprets 0.10-0.20 as the pre-signal baseline and 0.3-0.4 as the post-signal level relevant for T*, and discloses that if the lower calibration is correct, the 2027-2033 window fails and the test resolves against the framework on timing even if the mechanism is right. True extremal sensitivity on those baseline-η ranges is 2.8-77.5 years; the 5.6-31 year bracket printed in an earlier version used non-extremal corners and is withdrawn. The window is directly falsifiable from Schedule P data filed each February with no look-ahead bias. The prediction is specific, public, and time-stamped: sustained industry-wide other-liability-occurrence adverse development at \$8B+/year through the 2025-2027 cycles, with the kill metrics preceding the next major commercial liability insolvency, confirms the framework. Reversion below \$3B/year by the 2026 or 2027 cycle absent a structural driver change (tort reform or verdict reversal) disconfirms it for this case. The \$3-8B/year band is indeterminate: consistent with the prediction, confirming nothing. No post-hoc adjustment is possible: parameters are fixed, the window is stated, the data are public.

Calibration across insurance cases:

Case	δ	η (initial)	λ	T* predicted	T* observed	β_W	Market missed by
Asbestos	$80-100B	0.05	$7B/yr	Multi-wave (143yr@η_0)	Multi-wave	2.5-3	10-15 yrs (1st wave)
GE LTC	$15B	0.05	$1.5B/yr	200 yrs @η_0	12 yrs (external trigger)	5-8	3-4 yrs
Casualty (current)	$9-62B	0.3-0.4	$10B/yr+	8.6 yr central (3-9 yr window)	TBD	3-8 est.	TBD

Theoretical contribution. Banking (§5.3) suppresses η via accounting rules (HTM). Insurance demonstrates something distinct: perpetual suppression through (a) statutory accounting allowing smooth multi-year development, (b) inherent policy-to-claim latency, and (c) coverage litigation that delays a legally definitive C=0 for years or decades. Insurance has the longest T* — simultaneously the most dangerous domain for Suboptimal Optimum and the most tractable for validation, given Schedule P data richness.

2025 data: consistent, not yet confirming. Data published after initial circulation are consistent with the predictions but do not yet confirm them: the headline adverse-development figure falls in the indeterminate $3-8B band defined above. S&P Global Market Intelligence revised the commercial liability combined ratio upward from 107.8 to over 110 during 2025 (S&P GMI, December 2025), with adverse development "running ahead of what we presumed to have been conservative projections." For calendar 2025, the other-liability-occurrence line alone reported $7.3 billion in adverse development — the largest of any line, though below the $8B/year confirmation threshold — with more than half concentrated in accident years 2021-2023 (Carrier Management, March 19, 2026). Swiss Re documented $16 billion in prior-year liability increases during 2024 reviews, with $62 billion cumulative adverse development across commercial liability lines over 2015-2024 (Swiss Re Institute, July 2025). The η suppressor is independently confirmed: workers'-compensation favorable development ($5.5 billion in 2025) continues to mask casualty deterioration in headline figures, but these releases are diminishing (Insurance Insider, 2026). Assured Research projects a $12.5 billion deficiency in other-liability-occurrence alone, with accident years 2021-2023 accounting for over 70% of the shortfall (Carrier Management, March 19, 2026).

5.6 Bitcoin Ecosystem — Six-Channel Analysis (β_W ≈ 2.0, central)

Bitcoin and the broader crypto ecosystem exhibit PST across six channels, each with an independently estimable β_W. This case shows SAPM decomposing aggregate system beta into channel-specific contributions.

System welfare function:

$$W = W_{\text{decentralization}} + W_{\text{security}} + W_{\text{environment}} + W_{\text{trust}}$$

Per-channel analysis:

Channel	Conservative β_W	Central β_W	Source/Basis	Classification	Annual Private Gain
Energy/Carbon externalities	3.1 (marginal)	3.1	Papp, Almond & Zhang (2023, JPubE); EPA SCC $190/tonne	Suboptimal Optimum (fast)	~$17.2B
Mining centralization	0.56	1.2	Author: β_W = (annual Nakamoto-coefficient decline × system-security cost) / mining pool revenue. Data: BTC.com / mem-pool.space; 2022-2024 top-3 pool share rising to 65%+	Suboptimal Optimum (slow)	~$3.4B
Security budget erosion	0.33	0.8	Author: β_W = (security-threshold shortfall cost) / (mining revenue decline). Data: CoinMetrics; 50%-attack-cost models (Budish, 2018, JPE)	Suboptimal Optimum (slow)	~$13.7B

Channel	Conservative β_W	Central β_W	Source/Basis	Classification	Annual Private Gain
Custody re-centralization	1.1	2.2	Author: β_W = (HHI of custody concentration × per-incident security cost) / custodial fee revenue. Data: Glassnode; Chainalysis Crypto Crime Report	Suboptimal Optimum (fast)	~$8B
Speculation vs. utility	0.8	1.6	Author: β_W = (speculative surplus above fair utility value) / (utility-transaction fee revenue). Data: CoinMetrics / BitInfoCharts; NVT framework (Burniske & Tatar, 2017)	Suboptimal Optimum (slow)	~$15B
MEV extraction	3.5	3.5	Author: β_W = (MEV-driven user welfare loss) / MEV revenue. Data: Flashbots MEV Explorer; sandwich-attack deadweight model (Daian et al., 2020)	Suboptimal Optimum (fast)	~$2B

The energy/carbon channel (β_W = 3.1) is peer-reviewed. The other five are author's estimates per the specified reasoning — illustrative calibrations pending independent verification.

Weighted aggregate β_W ≈ 2.0 (central estimate, weighting by private payoff per channel, summing to ~$59B annually). The independently validated energy channel (β_W = 3.1; Papp, Almond & Zhang, 2023, Journal of Public Economics) dominates given its $17B share. Computed as β_W = Σ_i (β_W^i × Π^i) / Σ_iΠ^i over the ~$59.3B total: (3.1×17.2 + 1.2×3.4 + 0.8×13.7 + 2.2×8.0 + 1.6×15.0 + 3.5×2.0) / 59.3 = 116.96 / 59.3 ≈ 1.97, rounded to 2.0. The Conservative β_W column yields β_W ≈ 1.5 under the same weighting. The headline result — Bitcoin operates at Suboptimal Optimum with aggregate β_W materially above 1, and negative system-adjusted payoff at any μ above 0.50 — holds across the full [1.5, 2.0] range. Readers preferring the conservative column should treat all subsequent Bitcoin figures as lower bounds.

Energy channel. Papp, Almond, & Zhang (2023, Journal of Public Economics) find a $1 Bitcoin price increase causes $3.11-$6.79 in carbon damages (Scrubgrass, an 86 MW waste-coal-fired plant; EPA $190/tonne SCC). Goodkind, Jones, & Berrens (2020) estimate 2016-2021 climate damages averaged 35% of market value, reaching 156% in some periods. β_W = 3.1 is conservative. The 2024 synthesis of 446 SCC studies by Moore, Drupp, Rising, Dietz, Rudik, & Wagner (PNAS, 2024) finds a mean SCC of $406/tonne C ($110+/tonne CO_2e), well above EPA's 2023 interim figure. At that SCC, the energy channel β_W rises to ~4.5-7.0, pushing aggregate β_W above 2.0 — making 2.0 conservative under higher SCC assumptions.

System-adjusted payoff. At Π ≈ $59B/year and system welfare cost ΔW ≈ −β_W × Π ≈ −$118B/year (central β_W = 2.0):

$$\Pi^{SA}(\mu = 1/2.0) = 59\text{B} - (1/2.0) \times 118\text{B} = \$0$$

$$\Pi^{SA}(\mu = 1.0) = 59\text{B} - 118\text{B} = -\$59\text{B}$$

At the frontier-optimal shadow price μ* = 1/β_W = 0.5, private gains exactly equal system costs. Above μ = 0.50, the system-adjusted payoff turns negative. The ecosystem runs a Suboptimal Optimum at scale across all central estimates.

5.7 Facebook / Meta — Slow Suboptimal Optimum (β_W = 0.3-0.6, T* = 50-100 Years at Current η)

Meta presents a structurally distinct case: low β_W, long T*, no single enforcement event. The system degrades slowly across multiple channels.

Per-channel β_W:

Channel	β_W Range	Mechanism
Mental health	0.5-0.9	Algorithmic amplification; Twenge et al. (2022); author's estimate — causal identification of platform-specific effect remains an active debate
Misinformation	0.3-0.7	Engagement-optimized content recommendation
Privacy erosion	0.1-0.3	Surveillance advertising model

Channel	β_W Range	Mechanism
Market concentration	0.1-0.2	Acquisition strategy eliminating nascent competitors
Attention extraction	0.2-0.4	Infinite scroll, notification design
Political polarization	0.1-0.5	Filter bubble architecture

Weighted aggregate β_W ≈ 0.3-0.6. The range width reflects genuine attribution uncertainty: what fraction of documented social costs is causally attributable to Meta's design choices versus user behavior and underlying social trends. The 0.3 lower bound uses conservative attribution consistent with Twenge et al. (2022); the 0.6 upper bound uses the broader attribution implied by Vosoughi, Roy, & Aral (2018) and internal Meta research disclosed in the 2023 MDL. The key result — β_W > 0 and T* > current date, placing Meta in active Suboptimal Optimum — holds across the range.

Crossover time: at δ ≈ $40B and λ = $8B/year (the litigation-era parameters below), T* = 50-100 years at the current η ≈ 0.05-0.10 (regulatory feedback suppressed by First Amendment constraints, jurisdictional fragmentation, and agency capture), compressing to 12.5-25 years under enforcement scenarios (η = 0.2-0.4).

Classification. p = + (Meta revenues $164B, 2024); s = − (mental health, misinformation, privacy degradation); d = + at 2026 (T* not yet reached). Suboptimal Optimum (slow).

Meta illustrates why T* matters for governance. At β_W = 0.3-0.6, the trap appears stable from any short horizon: shareholders profit, users receive free services, advertisers reach audiences. Only SAPM makes the systemic cost visible and the crossover calculable.

Litigation acceleration as η-transition signal (2023-2026). The 2023-2026 period produced the first quantitative evidence that η is rising above Meta's historical ~0.05 baseline. A bipartisan coalition of 41 U.S. attorneys general filed a federal lawsuit in October 2023 alleging Meta knowingly designed addictive features targeting minors. The federal MDL (No. 4:22-md-03047-YGR, In re Social Media Adolescent Addiction/Personal Injury Products Liability Litigation) grew to over 1,700 cases by April 2025; bellwether trials commenced late 2025. Unsealed filings revealed Meta's own internal research — subsequently discontinued — found Facebook use worsened depression, anxiety, loneliness, and social comparison. This is precisely the independent W-monitoring channel Proposition 3.5 predicts is necessary: Meta's suppressed internal W-measurement provides the first inside evidence of the degradation SAPM identifies. In December 2025, the Massachusetts Supreme Judicial Court heard arguments on whether Meta's intentional addictive design for minors is actionable under state consumer protection law. Under revised parameters: at η = 0.2 (rising from 0.05 as exposure mounts), λ = $8B/year, and δ ≈ $40B in surplus attributable to engagement-maximizing design, T* ≈ 40/(0.2 × 8) = 25 years from 2004, placing crossover at 2029. At η = 0.4 (full enforcement), T* = 40/(0.4 × 8) = 12.5 years from 2004, placing crossover at ≈ 2016-17 — already past, consistent with the litigation pattern.

What would constitute T*-crossing for Meta? One or more of: (i) a documented threshold crossing in adolescent mental health metrics legally attributable to platform design (η transition via litigation); (ii) a major EU or UK break-up order forcing structural separation of Instagram/WhatsApp from the core platform (external η spike); (iii) a measurable demographic collapse in core engagement as

Gen Z exits to alternatives (endogenous η spike via substitution); or (iv) a congressional structural-separation mandate (Sherman Act §2). These events should coincide with T* measured from the 2004 founding: 50-100 years at the current suppressed η (window 2054-2104), or 12.5-25 years under enforcement-scenario η (window 2016-2029) — the range the litigation record already occupies. The intervention point is now — before T*, while damage remains structurally reversible. Post-T*, network effects and regulatory capture make correction orders of magnitude more costly.

5.8 US Commercial Real Estate — Prospective Case 2: "Extend and Pretend"

US commercial real estate presents a second prospective application, with multiple kill metrics already triggered at the March 2026 parameter-fixing date.

Suboptimal Optimum structure. Banks holding $1.7 trillion in CRE loans and their borrowers run a bilateral "extend and pretend": maturing loans are repeatedly extended rather than refinanced at market values, letting banks avoid loss recognition and borrowers avoid forced sales. Both report satisfactory outcomes — performing loans, occupied buildings — while CRE price discovery, banking capital adequacy, and credit allocation degrade.

Parameter estimation.

Parameter	Input	Derivation	Value
δ	Excess bank capital above regulatory minimums	FDIC aggregate data	~$200-300B
λ	Annual real economic losses hidden by extensions	Loss severity × non-performing loan volume	~$50-100B/year
η	Visibility of losses to market	Suppressed by HTM accounting, appraisal lag (12-18 months), regulatory forbearance	~0.15-0.30 (0.40 = the transition scenario, outside this range)
β_W	System welfare cost per dollar of avoided loss recognition	Capital misallocation + price discovery suppression + contagion concentration	~1.8 (order-of-magnitude structural estimate, not derived)
T*	δ/(ηλ)	$250B/(0.20 × $75B)	~16.7 years at current η

η transition. As of early 2026, η is rising. Lenders are abandoning extend-and-pretend as pandemic-era rate expectations prove unfounded (The Real Deal, February 2026). The CMBS maturity payoff rate has fallen from 80% in 2023 to below 50% projected for 2026 (Morningstar DBRS). With η transitioning from ~0.15 toward ~0.40, T* = δ/(ηλ) collapses from 16.7 years at current η = 0.20 to 8.3 years if η completes its transition to 0.40; propagating the full stated parameter ranges gives 7-40 years, and all windows anchor to the March 2026 parameter fixing — structurally identical to SVB: suppressed η, then abrupt transition when the suppressor (HTM accounting, extend-and-pretend) exhausts.

Kill metrics (as of March 2026).

Kill Metric	Current Value	Threshold	Status
Office CMBS delinquency rate	12.34% (record since 2000)	>10%	Triggered
CMBS loans past maturity without resolution	$25B	>$15B	Triggered
Maturity payoff rate	<50% (2026 est.)	<60%	Triggered
CRE debt maturing in 2026	$875B	Historically high (directional; no binary threshold)	Directional
Regional bank CRE concentration	31.5% of all commercial mortgages	>25%	Triggered
KRE (Regional Bank ETF) single-day drawdown	-5% (March 2, 2026)	Accelerating volatility (directional; no binary threshold)	Directional / Warning

Four of six kill metrics have crossed their binary thresholds; the 2026 maturity wall and KRE volatility are directional indicators without binary thresholds. The 12.34% office CMBS delinquency rate is the highest in the 26-year history of Trepp's tracking (Trepp, January 2026). Approximately $25 billion in CMBS loans are past maturity without repayment, liquidation, or formal extension — levels not seen since the post-2008 cleanup. Regional and community banks, holding 31.5% of all commercial mortgages (Cohen & Steers, 2025), face concentrated exposure as extend-and-pretend exhausts.

Predictions (time-stamped, falsifiable).

1. At least three regional banks with CRE concentration exceeding 300% of tier-1 capital will require FDIC intervention or forced merger by Q4 2027.
2. Total recognized CRE losses across the US banking system will exceed $150B cumulative by year-end 2027, versus ~$30B recognized through 2025.
3. The KRE (Regional Bank ETF) will underperform the S&P 500 by more than 20 percentage points over the 18 months following publication.

Prediction 2 in its stated form requires ~$60B/year of loss recognition (($150B − $30B) over two years), versus a model-implied recognition flow of $\eta\lambda$ = $15B/year at central parameters and at most $40B/year at the transition maximum (η = 0.40, λ = $100B/year). It is therefore not a flow prediction of the model. It requires recognition of roughly one-third of the ~$200-400B stock of losses accumulated under extend-and-pretend since 2022 — the stock-recognition mechanism this paper predicts the η-transition will force, as forced sales and refinancing mark the accumulated gap to market at once. If the stock mechanism is rejected, the SAPM-consistent flow thresholds are >$90B cumulative by year-end 2027 at transition parameters (η = 0.40, λ = $75B/year) and >$60B at central parameters. Predictions 1 and 3 are free-standing market calls consistent with, but not derived from, the SAPM η-transition diagnosis. All three 18-21-month predictions test the η transition (the SVB-analog mechanism), not T* itself, which resolves on a 2034-2043 horizon

outside the 36-month protocol. Their confirmation or disconfirmation therefore tests the η-transition diagnosis; T* is tested only on the longer horizon.

Sources: The Real Deal (February 17, 2026); Morningstar DBRS (January 2026); Trepp CMBS Research (January 2026); FDIC Quarterly Banking Profile (Q4 2025); Mortgage Bankers Association (March 2026); Cohen & Steers (2025).

5.9 Sharpening η and T* with a welfare-touching channel

The suppression coefficient η, the damage flow λ, and the deficiency δ are calibrated above from public filings and outside estimates, and η in particular is the parameter the cases most contest (the casualty case carries η ≈ 0.3-0.4 as its current post-signal reading, rising from a historically suppressed baseline). These parameters can be sharpened, and η can be estimated rather than asserted, by a welfare-touching channel: a measurement of true systemic welfare W that does not pass through the accounting that suppresses it. In the companion Goodhart Gap kernel this is a gold channel, a channel whose readings depend on realized welfare rather than on the reported trajectory; the two questions, how to measure η empirically and what channel escapes the suppression, are the same question.

Banks. Two welfare-touching channels are public and quarterly. On the securities book, the gap between fair value and amortized cost on held-to-maturity holdings is reported in FDIC call reports and is an unsuppressed read on the loss the accounting defers (this is the gap that crystallized abruptly at Silicon Valley Bank on the forced HTM-to-AFS reclassification, §5.3). On the loan book, market-implied expected loss, from credit-default-swap spreads on the institution, secondary-market prices on comparable loan pools, and, at resolution, the price in FDIC loss-share deals, minus booked current-expected-credit-loss reserves, reads the un-recognized credit loss. In each case η on that slice is booked recognition divided by the channel's estimate of the true accumulated loss.

Insurance. The market gold channel is the loss-portfolio-transfer and adverse-development-cover market. When an insurer cedes reserve risk to a reinsurer, the price of the transfer is an outside estimate of the true ultimate loss, produced by a counterparty with its own capital at stake and independent of the ceding insurer's carried reserves; the gap between the transfer price and the carried reserve reads η directly for that book. Alongside it sit outside ultimate-loss estimates (Assured Research's $12.5B other-liability-occurrence deficiency; Swiss Re's $62B cumulative commercial-liability adverse development) and the Schedule P development triangle run to ultimate by an actuary other than the ceding one. Kill Metric 3 of §5.5 is the internal version of the same measurement; the transfer price is its market-priced form.

Estimator. Given a welfare-touching series $W_{gold}(t)$, the parameters stop being hand-set: $\eta(t)$ is recognized loss to date over W_{gold}'s estimate of true accumulated loss to date, λ is the time derivative of W_{gold}, δ is the current gap, and $T^* = \delta/(\eta\,\lambda)$ then carries error bars inherited from the channel's own estimation uncertainty rather than standing as a defended point value. This is the route by which the conflicting η calibrations of a single case (the casualty line's suppressed pre-2024 baseline near 0.10-0.20 against its 0.3-0.4 post-signal reading) would be settled by measurement rather than argument, from the HTM-versus-book gap or the transfer-price-versus-reserve gap, both public and time-stamped.

The channel is one-sidedly informative, by the companion's Corollary 3Z. The welfare-touching

channel is priced and its visibility is asymmetric. It cheaply and validly proves η too low, that the institution is under-reserved, whenever the transfer price exceeds book or fair value sits below amortized cost; it cannot certify η high enough, that reserves are adequate, without paying the full welfare-touching cost, which is exponentially expensive in the thin-edge regime. And the purest gold channel, realized defaults and forced-sale crystallizations, arrives with exactly the latency that a low η encodes. There is therefore a floor on how precisely η can be pinned before the crossover, and that floor is not a data-collection gap but the impossibility the companion proves: the measurement instantiates the theorem rather than escaping it. This is the empirically useful direction. The gold-channel program tightens η and T* precisely where the system is already past adequacy and heading toward its crossing, which is the direction that matters for the prospective test, and it does so without ever licensing the reassuring certificate the theorem forbids.

---

6. Responding to Anticipatable Referee Objections

6.0 The Definitional Objection

Objection. The Private Pareto Trap is a tautology: PST-G defines the game so private improvement degrades W, so Proposition 3.8 follows trivially.

Response, in three layers. Structural: Proposition 3.5 is independent of PST-G. W-independence follows from Axiom 2 alone, proving no mechanism conditioned only on operative private payoffs can detect systemic degradation in any game satisfying Non-Reducibility. It precedes and grounds the measurement-impossibility claim. Conditional theorem: Proposition 3.8 deliberately uses PST-G as the global sufficient condition. It is not a definition of PP(Γ), which is computed over u^P alone. Empirical: In any game, PST-G is falsified if a strategy strictly improving some agent's operative private payoff preserves or improves W. Such a profile is either outside the trap domain or evidence that W-contact has already altered the operative payoff, constraint, audit, or objective. The claim "all eight empirical cases satisfy PST-G" is empirical, not definitional.

6.0b Objection: Rational Agents Would Internalize Systemic Costs

Objection. Real agents care about the system, so they internalize systemic costs voluntarily and the trap never closes.

Response. Three behavioral mechanisms, each grounded in experimental economics, close the trap even for agents with genuine but bounded systemic preferences. Let perceived embedded welfare be $\tilde{v}_i = u_i^P + \alpha \cdot u_i^S$, where α is the fraction of systemic welfare made operative in the agent's decision process. Charness and Rabin (2002) calibrate α ≈ 0.3–0.5; competitive professional settings push toward α ≈ 0.2. (1) Reference-point adaptation (Kahneman & Tversky, 1979): as Suboptimal Optimum persists, agents adapt to the degraded state—a LIBOR bank submitting false rates for three years treats the corrupted benchmark as normal, shrinking perceived systemic loss even as actual degradation accumulates. (2) Social discounting (Andreoni & Miller, 2002): the 1/d law of giving discounts the abstract, faceless system against the proximate, named counterparty; contribution rates to identifiable versus statistical beneficiaries differ by 2–4× (Small, Loewenstein, & Slovic, 2007). (3) Diminishing sensitivity in the loss domain (Kahneman & Tversky, 1979): prospect theory's value function is convex in losses, so each added unit of degradation generates less disutility, accelerating the trap toward Crash. Combined with α ≈ 0.3, these yield

α_eff ≈ 0.05–0.10. For private surplus δ = $5B and systemic loss $25B, the trap closes when α_eff < 0.2. Behavioral realism reinforces Proposition 3.8 rather than contradicting it: agents have systemic interests, but those interests are weakly internalized in the operative objective absent a W-channel.

6.1 “This Is Just Arrow Dodged by Smuggling in Cardinal Comparability”

Objection. Arrow’s theorem rules out social choice under universal domain, Pareto, IIA, and non-dictatorship. Moving to cardinal payoffs and a cardinal W requires interpersonal comparability, which Arrow prohibits.

Response. The objection conflates two impossibilities about different objects. Arrow constrains aggregation procedures mapping ordinal preference profiles to social choice rules. Proposition 3.5 constrains measurement: no aggregation recovers W from the operative private payoff vector $(u_1^P,\ldots,u_n P)$. The two are logically independent. System beta β_W indexes admissible comparability: at β_W = 0 the SAPM collapses to the Arrovian benchmark with no interpersonal comparisons, nesting Arrow as a special case rather than dodging it. β_W is estimated from observable system behavior—enforcement data, market-structure changes, degradation in public indicators—not from normative welfare weights.

Kaplow-Shavell tension. Kaplow and Shavell (2001) prove any policy assessment using non-utility information violates Pareto. The SAPM uses W (non-utility by Axiom 2), but Π^SA is a diagnostic statistic, not a welfare ranking. Whether a regulator should override agents’ short-run preferences is a normative question the SAPM leaves open; it supplies the measurement that makes the short-run/long-run discrepancy observable (Proposition 3.6: agents win on u_i^P while losing on u_i^S). The SAPM is logically prior to the K-S debate, not in competition with it.

6.2 “Your Welfare Criterion Is a Domain-Restriction Theorem in Disguise”

Objection. Axioms 1–4 restrict the game domain, so the impossibility is weaker than a universal one.

Response. Correct and acknowledged. The restriction is explicit and empirically salient. Axioms 1–4 are testable per game, are examined across the eight cases in §5, and fail for zero-sum games, pure coordination games, and games where agents benefit from system destruction, all declared out of scope. The domain covers precisely the settings where cooperative failure is most consequential: benchmark manipulation, cartel pricing, commons depletion, platform externalities, AI collusion. Propositions 3.11–3.12 establish that it contains value-redirecting games and separable-extraction commons dilemmas as structural subclasses; the subclass claim uses the slack form of Proposition 3.11 (the zero-slack special case is Axiom-2-reducible and serves only as a PST-G template).

6.3 “β_W Is Normatively Underidentified”

Objection. β_W = −dW/dΠ requires specifying W; utilitarian, Rawlsian, or capability-based choices yield different estimates, giving free parameters that rationalize any conclusion.

Response. β_W is underidentified exactly as CAPM’s β is: different market portfolios yield different βs. The discipline’s answer is transparent disclosure, not abandonment. β_W sits in the Atkinson-Mirrlees tradition—Atkinson’s inequality index (1970) requires specifying ε; Mirrlees’s optimal tax

theory (1971) requires specifying the SWF; neither demands a uniquely justified foundation, only an explicit one. Readers who disagree recompute β_W under their preferred W. The eight cases specify W as the system property most directly implicated:

- Bitcoin (carbon channel): W = −SCC × CO_2e, SCC = $190/tonne (EPA, 2023); assumes utilitarian aggregation, ~2.5% real discount rate (Stern-Weitzman range), global equal weighting. At Moore et al. (2024, PNAS) $406/tonne C, carbon-channel β_W rises to ~4.5–7.0.
- LIBOR: W = −Σ(|r_submitted − r_true| × notional exposure_i); no distributional weighting across counterparties.
- 2023 bank failures: W = −Σ(expected depositor loss + emergency facility cost + stability externality), from FDIC resolution accounting plus Fed facility outlays.
- Meta: W = −Σ(mental-health costs + misinformation-correction costs + competitive losses), from Vosoughi et al. (2018) and Twenge et al. meta-analytic estimates.

Each specification is contestable; readers can recompute β_W under different weights, discount rates, or proxies without changing the formal framework.

6.4 “T* Is Not Separately Identified from Functional-Form Choices”

Objection. T* = δ/ηλ requires three unobserved parameters (δ, η, λ); different damage-function parameterizations yield any T*, so §5 reflects functional form, not data.

Response. T* is invariant to monotone renormalizations of the welfare scale—multiplying W by a constant scales λ but leaves T* unchanged—because T* depends on the ratio δ/ηλ, not the levels. Cross-environment variation pins it empirically: in VW, η is estimated from the NHTSA penalty trajectory (observed 0.3) against the counterfactual under full information (0.7–0.9 from comparable unmanipulated cases). The defeat device suppressed η without changing δ or λ—a falsifiable identification assumption. The prospective strategy (§6.5) sharpens this with quasi-experimental η variation and change-point models that pin T* without functional-form assumptions.

Illustrative retroduction vs. prospective validation. VW (§5.2), LIBOR (§5.4), Bitcoin (§5.6), Meta (§5.7), and the 2023 bank failures (§5.3) are illustrative retroductions: parameters estimated from post-outcome data, showing structural coherence, not out-of-sample forecasting. The genuinely prospective cases are US casualty insurance (§5.5, parameters from 2024 Schedule P filings before resolution) and US CRE (§5.8, parameters from March 2026 before resolution). Retroductions calibrate; prospective cases test.

6.5 “No Horse-Race Against Classic Welfare Benchmarks”

Objection. The paper never compares SAPM predictions against utilitarian, Rawlsian, Kaldor-Hicks, or revealed-preference benchmarks, so the added complexity is unjustified.

Response. The SAPM is a competing decision rule, not a competing welfare function. Against the utilitarian rule it adds −μ(W_C − W); against Rawlsian, the sustainability condition T* > H; against Kaldor-Hicks, the impossibility result—compensation through the private channel is structurally infeasible because W is orthogonal to private payoffs. A formal horse-race is implementable:

Data. A panel of N ≥ 200 policy episodes with distributive tradeoffs where the private-systemic distinction is observable (environmental regulation, financial-stability interventions, antitrust actions), 1980–2026, ≥ 15 countries.

Identification. Quasi-experimental η variation from exogenous disclosure shifts: (i) SEC 2009 Fair Disclosure expansion (financial benchmarks); (ii) EU 2017 GDPR (platform data markets); (iii) FASB 2023 ASC 320 revision triggered by SVB (bank duration risk). DiD compares treated vs. control before and after; shift-share instruments combine η-shifting intensity (shift) with pre-existing PST exposure (share). Estimate β_W^i = −ΔW_i/ΔΠ_i from DiD variation. A change-point model identifies T* as the threshold where private surplus Θ_H flips from positive to negative, avoiding functional-form assumptions.

Regression:

$$y_{it} = \theta_1 \cdot \text{SAPM}_{it} + \theta_2 \cdot \text{Utilitarian}_{it} + \theta_3 \cdot \text{Rawlsian}_{it} + \theta_4 \cdot \text{KaldorHicks}_{it} + \theta_5 \cdot \text{Politics}_{it} + \gamma_t + \alpha_i + \epsilon_{it}$$

where y_it is system welfare and each score is constructed from its benchmark.

Predicted output (Table 2):

Welfare Score	Predicted θ	Significance	Intuition
SAPM (θ_1)	Positive, ~0.3–0.6	✓ p < 0.01	Survives—incorporates systemic channel
Utilitarian (θ_2)	Near zero	✗ p > 0.10	Attenuated—Σu_i^P misses W under Axiom 2
Rawlsian (θ_3)	Positive, small	△ p ~ 0.05–0.10	Partial—min u_i^P correlates with preservation but misses dynamics
Kaldor-Hicks (θ_4)	Near zero	✗ p > 0.10	Crowded out—same info as utilitarian
Politics (θ_5)	Varies	△	Control—separates capture from structural PST

The core prediction: θ_1 stays positive and significant while θ_2 and θ_4 shrink when system welfare, not Pareto efficiency, is the dependent variable. This would be the first causal identification of PSF structure in policy data, converting the objection into a falsifiable prediction.

6.6 “This Is Just Externalities Theory in New Notation”

This is the most important objection the paper faces; the surface resemblance is real—both frameworks involve private gains imposing costs on others—but the substance is structurally distinct. The full comparison against Pigou, Coase, Hardin, Ostrom, the Folk Theorem, and the Price of Anarchy lives in Table 3 (§7.1). The single central claim: every prior framework assumes the relevant welfare measure is, in principle, computable from agent-level information; Proposition 3.5 proves this fails in PST games.

Pigouvian taxation requires observing ΔW to set tax = marginal damage. By Proposition 3.5, W is not a function of the operative private payoff vector, so a regression of ΔW on agents’ operative payoffs leaves an irreducible residual that no amount of payoff-side controls removes—its R^2

is bounded away from 1. This is informational impossibility, not transaction cost. LIBOR is the cleanest test: two strategies—one bank's false rate affecting $350T in contracts, another affecting $1B—can produce identical operative payoff vectors for all sixteen panel banks. Any g with W = $g(u_1^P, \ldots, u_n P)$ returns the same output for both, falsifying it directly. A cross-section regression of ΔW on Δu^P yielding $R^2 \approx 1$ would falsify Axiom 2; the §6.5 horse-race tests exactly this. Coase fails for the same prior reason, not the transaction-cost reason Coase himself conceded: by Axiom 2, the system-affected population—every counterparty referencing a corrupted benchmark, hundreds of millions never at any table—lives in W-space, orthogonal to the private payoff space where all parties are defined. Even at zero transaction cost, Coasian bargaining cannot convene a population it cannot identify.

The SAPM shadow price μ* = 1/β_W resembles a Pigouvian tax only in form. A Pigouvian tax requires W observable from operative private payoffs; μ* requires a separate, independently constructed W-channel—the one Proposition 3.5 proves necessary and the SAPM supplies via β_W estimation. The SAPM explains why the Pigouvian program needs a capability it cannot supply from within the private payoff space, and supplies it.

6.7 Hardin, Ostrom, Folk Theorem, Price of Anarchy

All four traditions claim existing frameworks already handle the problem. All assume the relevant welfare is computable from agent-level information; Proposition 3.5 proves otherwise. Full detail is in Table 3 (§7.1).

Hardin (1968) is a theory of defection: agents defect, the resource depletes, cooperation prevents the tragedy. The Private Pareto Trap is the opposite failure—agents cooperate perfectly, reach PP(Γ) on the operative private payoff channel with zero defection incentive, and the system degrades because of their cooperation. LIBOR banks, VW teams, and the Mazur (2025) LLM agents did not defect; they coordinated flawlessly. Property rights and iterated games address non-cooperation, not Suboptimal Optimum.

Ostrom (1990) designs institutions for commons with physical boundaries, enumerable membership, and observable stocks—precisely where Axiom 2 is weak and W is recoverable from extraction rates. Her eight principles work by raising η, clarifying C, and expanding the strategy space toward Integration. The Private Pareto Trap explains why they are necessary: without transformation, PST games close the trap by construction. But the dominant modern PST settings cannot meet her conditions—benchmark integrity has no physical boundary, platform welfare no identifiable membership, AI coordination no observable signal. There the SAPM supplies the measurement infrastructure her governance design presupposes but cannot create.

Folk Theorem characterizes which operative payoff vectors patient agents sustain—including every Suboptimal Optimum equilibrium. Integration (1,1,1) and Suboptimal Optimum (0,1,1) can both be Folk equilibria; only one preserves the system. The theorem is not wrong; it lacks the variable to express the distinction.

Price of Anarchy compares the performance of equilibria to the optimum under a chosen welfare objective. When that objective is instantiated as aggregate operative private payoff, $W = \Sigma_i u_i^P$, it assumes exactly the reduction that Proposition 3.5 denies under Non-Reducibility. In LIBOR, a PoA calculation on bank trading payoffs could classify the coordination as welfare-improving if banks' trading gains exceed counterparty losses; SAPM classifies it as Suboptimal Optimum with β_W ≈ 3–10 because W ($350T in pricing accuracy) is orthogonal to the banks' operative payoffs.

PoA measures equilibrium inefficiency relative to its specified objective; PPT measures the cost of cooperation in PST games—orthogonal failure modes.

Every prior framework asks how to correct incentive misalignment within the observable payoff space, assuming the relevant welfare is accessible from agent-level quantities. Proposition 3.5 proves it is not accessible from operative private payoffs alone. The Private Pareto Trap is the logical prerequisite they collectively skip—a structural result about what information payoff vectors carry, prior rather than incremental.

6.8 Prospective Validation Protocol

The SAPM generates falsifiable out-of-sample predictions. A researcher should follow this protocol over the next two to three years.

Data. Three live datasets: (1) US casualty insurance Schedule P statutory filings (NAIC Annual Statements, filed each February) for accident years 2019–2027—loss reserve triangles by carrier and line, with adverse prior-year development directly observable; (2) quarterly FDIC Call Reports—duration risk (unrealized losses/equity) and deposit concentration; (3) quarterly 10-Q filings for publicly traded commercial-liability insurers. For the insurance test, also collect: industry-wide adverse development by accident year and line; nuclear verdict counts and aggregate awards (Marathon Strategies; Swiss Re); Schedule P reserve-to-surplus ratios by segment; AM Best and S&P ratings actions attributed to casualty reserve inadequacy.

Confirmation. Case 7c (§5.5) yields a central estimate of $T^* \approx 8.6$ years from 2024, with a 3–9 year operational window stated as a judgment (§5.5). The framework is confirmed if: (i) adverse prior-year development continues at ≥ $8B/year through at least two of the 2025–2027 filing cycles; (ii) commercial-liability reserve-to-surplus ratios deteriorate ≥ 1 SD from the 2015–2022 baseline within 36 months; (iii) at least two ratings actions (negative watch or below) are attributed to casualty reserve inadequacy in accident years 2019–2024. All observable from public data with no look-ahead bias.

Disconfirmation. The framework is disconfirmed—2024 adverse development reclassified as cyclical—if: (a) industry-wide adverse development in other-liability-occurrence reverts below $3B/year by the 2026 or 2027 cycle without a structural driver change (tort reform, nuclear-verdict reversal); (b) the three insurance kill metrics (§5.5) fail to precede the next NAIC-designated commercial-liability insolvency; or (c) the §6.5 horse-race yields θ_1 statistically indistinguishable from zero on the 200-episode panel.

Timeline. The insurance test resolves within 36 months (NAIC filings February 2026, 2027, 2028). The banking kill-metric test is rolling—each new FDIC failure is an out-of-sample observation. The horse-race needs ~18 months of a funded program. The SAPM predicts specific outcomes, at specific times, from public data that already exists.

6.9 What SAPM Predicts That Standard Models Cannot

a) Volkswagen Dieselgate. Standard externalities theory: emissions are bad; taxes or standards fix them. SAPM predicts $T^* \approx 6.1$ years to collapse via η-suppression—the defeat device suppressed η, extending T^* while worsening the crash. Observed: ~6 years from deployment (2009) to enforcement (2015). Standard models cannot say “this flips in six years.”

b) US Casualty Insurance Reserves. Standard frameworks flag adverse development only after it hits financial statements—a lagging indicator. SAPM's kill metrics (adverse prior-year development rate, reserve-to-surplus deterioration, nuclear-verdict trend) identified structural inadequacy prospectively from Schedule P filings, T* ≈ 8.6 years central from 2024 data (1.4–25.8 years propagated; 3–9 year operational window), before the losses are booked.

c) US Commercial Real Estate. Standard models classify rising delinquencies as cyclical or structural after the fact. SAPM gives a structural exchange rate (β_W ≈ 1.8, an order-of-magnitude structural estimate: each dollar of extend-and-pretend gain costs roughly $1.80 in deferred impairment) and falsifiable predictions: T* = 8–17 years anchored to the March 2026 parameter fixing, four of six kill metrics across their binary thresholds, with time-stamped falsification criteria.

d) AI Agent Collusion. Standard antitrust requires evidence of agreement (Sherman Act §1; EU Article 101). SAPM shows why agreement-free coordination can emerge structurally from W-blind optimization: optimizers with objectives defined only on private payoffs, in repeated PST environments, can support Suboptimal Optimum outcomes with no explicit agreement. California AB 325 (effective January 1, 2026) targets shared pricing algorithms but is structurally blind to independent W-blind optimizers; five research teams have documented AI cartel pricing in simulated or experimental settings without any shared algorithm. SAPM identifies the structural cause AB 325's "common algorithm" requirement misses.

e) Meta / Facebook. Standard frameworks debate whether harm exists. SAPM quantifies β_W = 0.3–0.6: each dollar of advertising revenue drains 30–60 cents from systemic welfare (platform integrity, cognitive wellbeing, information quality, market health). A low per-dollar rate at enormous scale—billions of users, hundreds of billions in revenue—yields massive cumulative damage. Standard frameworks cannot generate this number.

The common thread. Standard frameworks identify that externalities exist. SAPM measures how fast they accumulate, when they become irreversible, and the exchange rate between private gain and system damage. The difference between knowing pollution is bad and knowing the system flips in 6.1 years is the difference between a policy paper and an expert witness report.

7. Discussion

7.1 Relationship to Existing Frameworks

Table 3: PPT/SAPM vs. Existing Frameworks—What Each Sees, What Each Misses

Framework	What it sees	What it misses	PPT/SAPM addition
Pigouvian externalities (Pigou, 1920)	Private cost ≠ social cost; corrective tax restores efficiency	Assumes W computable from agent-level damage data—Proposition 3.5 proves this fails in PST games	Independent W-channel; β_W = −dW/dΠ quantifies the gap without requiring W recoverable from u^P

Framework	What it sees	What it misses	PPT/SAPM addition
Coasian bargaining (Coase, 1960)	Zero-cost negotiation achieves efficiency; property rights solve externalities	Affected parties (in W, not in u^P) are informationally invisible; who sits at the table is structurally undeterminable from operative private payoffs	Proposition 3.5 makes it precise: in LIBOR, $350T in affected contracts cannot be identified from 16 banks' operative payoffs, regardless of transaction costs
Tragedy of the Commons (Hardin, 1968)	Rational defection destroys shared resources	Assumes the failure is non-cooperation; Suboptimal Optimum shows full cooperation on the private frontier destroys the system	The problem is achieving private Pareto efficiency, not defecting from it; remedies for defection cannot reach it
CPR governance (Ostrom, 1990)	Self-organized institutions prevent tragedy for bounded physical commons	Requires identifiable boundary, enumerable membership, observable stock—failing for benchmark integrity, platform welfare, AI coordination, financial stability	T* and β_W as measurement tools for non-physical C; PPT as the formal foundation for why the eight principles are necessary
Folk Theorem (Fudenberg-Maskin, 1986)	Patient agents sustain any individually rational payoff as Nash equilibrium	Classifies Integration (1,1,1) and Suboptimal Optimum (0,1,1) identically—no vocabulary for the distinction	Three-dimensional taxonomy: both are Folk equilibria; only one preserves C
Price of Anarchy (Koutsoupias-Papadimitriou, 1999)	Equilibrium inefficiency relative to a specified objective, often aggregate agent payoff	If the objective is instantiated as W = Σu_i^P, it assumes the reduction Proposition 3.5 refutes; Suboptimal Optimum can have no PoA loss under that objective while W < W_0	PPT measures the cost of cooperation across private/system channels; PoA measures inefficiency relative to its chosen objective

Framework	What it sees	What it misses	PPT/SAPM addition
Mechanism design (VCG; Hurwicz)	Incentive-compatible efficiency within private payoff space	W-blind by Proposition 3.5 when message and audit spaces are restricted to u^P; incentive compatibility cannot add missing W information	Shadow price μ* = 1/β_W as the efficient systemic tax; independent W-monitoring as structural prerequisite
Arrow's theorem (1951)	Preference aggregation impossible under 4 axioms (within one space)	Operates entirely within preference space; silent on W orthogonality	Cross-space impossibility: even perfect aggregation of $(u_1^P,\ldots,u_n P)$ cannot recover W under Non-Reducibility. Arrow: no perfect aggregator exists; PPT: a perfect aggregator, if built, is still W-blind
Regulatory Impact Analysis (OMB Circular A-4)	Requires agencies to estimate benefits (W) independently from costs (Π)	Treats W-observation as institutional requirement, not structural necessity; no theory of when W-observation fails or T* is near; applies retrospectively with no crossover detection	Proposition 3.5 grounds why OMB's requirement is necessary, not merely advisable; adds T* as a prospective indicator and a continuous framework for AI systems, financial markets, and institutional design

Scope. These frameworks are correctly designed for the payoff space they address; W-blindness is a structural property of single-space frameworks applied to PST games, not a design defect. The SAPM adds the second channel; it does not replace the first.

The social-cost-of-carbon literature illustrates the Pigouvian connection. Nordhaus's DICE and the EPA Social Cost of Carbon ($51/tonne in 2021; $190/tonne in 2023; Trump EPA announced administrative review March 2025) estimate the private-social gap; in SAPM terms the SCC is β_W × (Π per tonne CO_2e), with β_W ≈ 3.1 for Bitcoin (Papp, Almond, & Zhang, 2023). Moore, Drupp, Rising, Dietz, Rudik, & Wagner (PNAS 2024), drawing on 446 studies, find a mean SCC over $400/tonne C, above federal estimates—implying β_W for fossil-fuel activities is systematically underestimated. The SAPM generalizes the SCC framework to any PST domain and supplies μ* = 1/β_W as the efficient corrective price.

Myerson-Satterthwaite (1983) and Green-Laffont (1977). Myerson-Satterthwaite proves no mechanism achieves ex-post efficiency, individual rationality, and budget balance in bilateral trade under

incomplete information—a constraint from incentive compatibility on eliciting private types. PPT proves a distinct result: W is unrecoverable from operative private payoffs even under full information, because the private-payoff channel is $(u_1^P, \ldots, u_n P)$ and W is orthogonal to it by Axiom 2. The two are logically independent—PPT holds even if M-S were false, and M-S holds even if W were fully recoverable. Green-Laffont proves no budget-balanced mechanism achieves Pareto efficiency in public goods—an infeasibility within the private-payoff mechanism space. PPT's claim is cross-space: W is not expressible as any function of transfers or operative payoffs, however the mechanism structures them within the u^P-restricted message and audit space of Proposition 3.5. PPT is prior: it limits what the aggregation machinery has to work with, not the machinery itself.

Bergson-Samuelson (1938, 1947) and Kaplow-Shavell (2001). Bergson-Samuelson functions aggregate utility information; the SAPM's W is explicitly not an aggregation of operative private payoffs (Axiom 2, Proposition 3.5). Kaplow-Shavell prove any SWF depending on non-utility information violates Pareto; Proposition 3.5 is prior, establishing that W cannot be expressed in operative private utility information—the structural necessity K-S takes as given. Sequence: Proposition 3.5 → K-S → independent W-monitoring is structurally necessary, not merely preferred.

Recent mechanism design. Pycia and Troyan (2023) connect transparency to credibility; transparent two-dimensional mechanisms systematically fail to communicate the three-dimensional structure needed to avoid Suboptimal Optimum. Akbarpour and Li (2020) establish a credibility-efficiency-simplicity trilemma, with credible mechanisms requiring observability—connecting to the W-monitoring requirement. Zeng (2025) develops identity-compatible auctions for multi-dimensional types, relevant to v_i = u_i^P + u_i^S once a mechanism makes more than the private channel operative. Komo, Kominers, and Roughgarden (2024) prove shill-proofness requires structure beyond standard IC, paralleling W-preservation requiring structure beyond private efficiency. Azevedo & Leshno's two-sided matching is instructive: markets may clear while exhibiting PST when clearing involves system-degrading bilateral agreements.

7.2 Implications for Antitrust

The W-independence result (Proposition 3.5) names what current antitrust cannot do. AB 325 (common-algorithm evidence) and Sherman §1 (agreement evidence) condition only on the parties' operative private payoffs and observable conduct; by Proposition 3.5 that information is W-blind, so it cannot certify whether a coordinated outcome preserved or degraded the system. Enforcement requires an independent W-channel: counterfactual-simulation evidence, price-cost-margin audits, or capability caps.

Current antitrust requires evidence of agreement (Sherman Act §1; EU Article 101), plus the usual operability questions of market power, harm to competition, counterfactual price or output, price-cost margins, and the per se/rule-of-reason posture. AI cartel studies (Mazur, 2025; Dou et al., 2025; Fish et al., 2025) show collusion-like pricing emerging from pure optimization without agreement in research-market settings. In November 2025, the DOJ announced a proposed RealPage settlement barring real-time confidential competitor data in pricing—an η-raising intervention—but critics note loopholes for lagged-data coordination. California AB 325 (effective January 1, 2026) prohibits "common pricing algorithms," targeting the instrument rather than the structure: independent W-blind optimization can produce Suboptimal Optimum collusion without any shared algorithm (Mazur, 2025; Fish, Gonczarowski, & Shorrer, 2025). A 2-variable model cannot address a 3-variable problem. SAPM supplies agreement-free detection as a screening layer: any outcome classified Suboptimal Optimum (0,1,1)—both parties gain, independently measured system met-

rics degrade—merits scrutiny, after which antitrust analysis still needs the standard evidence of competitive harm, market power, margins, and counterfactual price or output.

7.3 Implications for Financial Regulation

The 2023 bank failures show SAPM's three observable kill metrics—unrealized losses/total equity, uninsured deposit concentration, quarterly deposit decline rate—give early warning one to two quarters before market consensus. The HTM accounting exception that suppressed feedback coupling in SVB and First Republic is the regulatory equivalent of the VW defeat device: it extends T^* by preventing the η phase transition. The insurance cases (§5.5) extend this to η suppression at 10–50 year timescales, with three insurance-specific kill metrics now being tracked in US casualty lines: (1) adverse prior-year development as a fraction of net premiums written; (2) nuclear-verdict count growth rate; (3) Schedule P reserve-to-surplus deterioration across consecutive accident years. As of the data reported here, one of the three is cleanly breached and the case remains an open prospective test. Schedule P statutory filings give regulators more granularity than banking, making insurance the highest prospective-validation sector. General implication: outcome-based standards measuring β_W and T^* from public filings, rather than seeking intent, detect PST games at the Suboptimal Optimum stage before Crash. SAPM names the architectural suppression instrument (HTM in banking, statutory reserve smoothing in insurance, defeat devices in automotive) and monitors its operation.

7.4 Implications for AI Architecture

RLHF, RLAIF, DPO, and Constitutional AI all condition only on the agents' private payoff signals, so by Proposition 3.5 they are W-blind: no amount of preference data recovers the system variable C. The prescription to make C a formal variable in the objective is the structural escape; it supplies the independent W-channel whose necessity Proposition 3.5 establishes. The four working channels are state reports (Townsend 1979), third-party audits (Border-Sobel 1987), capability caps, and pressure caps.

Current AI alignment optimizes for human preferences—structurally Robbery by B (0,0,1) with deference to the user. Mazur (2025) and Dou et al. (2025) show multi-agent systems converge on Suboptimal Optimum even when each agent is aligned to its own user. Architectural proposal: make C a formal variable in every agent objective, with mandatory system-interest modeling and post-interaction classification via the eight-outcome taxonomy; flag agents that consistently produce (c=0, a=1, b=1) for audit. Every existing governance framework—EU AI Act (2024) risk tiers, agreement-based antitrust, the UK CMA's March 2026 "frontier enforcement" guidance, the regulatory challenges Fish, Gonczarowski, & Shorrer (2025, updated March 2026) flag—operates in the two-variable private payoff space. They detect one agent harming another but not both agents gaining while the system degrades, because Suboptimal Optimum is invisible to any framework that does not track C.

7.5 The Organizational Expression: Kerr's Folly

Kerr's 1975 paper documented institutions systematically rewarding behaviors that undermine their stated goals. Every example is Suboptimal Optimum (0,1,1): rewarded and rewarding parties both meet immediate objectives while the shared system degrades. Universities want educational quality (C) but reward publication volume—faculty meet targets (A=1), departments get funding (B=1),

student learning deteriorates (C=0). Kerr's diagnosis is descriptive; the three-dimensional framework supplies the formal classification: Kerr's Folly is the organizational expression of Suboptimal Optimum, made structural through incentive design.

7.6 False Positives and Diagnostic Boundaries

The SAPM is a classification tool and can produce false positives. Three systematic sources:

1. Dynamic W measurement error. If W is measured when the system is temporarily below baseline for exogenous reasons (recession, regulatory transition, shock), β_W appears positive even when agent strategies are not causing the degradation. Diagnostic: does W degrade simultaneously with and proportionally to private extraction? If W degradation is orthogonal to the extraction trajectory, the PST classification is wrong.

2. Short-horizon W evaluation. Some profitable activities degrade W short-run but raise it long-run—innovation is canonical. A pharmaceutical trial may look Suboptimal Optimum over one year but Integration over ten. The correct horizon is the issue's natural time scale (for drug approval, the trial horizon).

3. Aggregate vs. distributional W. SAPM treats W as an aggregate. A constant W shifting between groups is not Suboptimal Optimum but Robbery by A (0,1,0) or Robbery by B (0,0,1). Decompose W into aggregate and distributional components first.

Boundary condition: SAPM applies when (a) W degradation is concurrent with private extraction, (b) the horizon matches the domain's natural time scale, and (c) W is measured as an aggregate property, not a distributional shift.

7.7 From Diagnosis to Integration

SAPM diagnoses Suboptimal Optimum and quantifies severity (β_W, T*); escaping it requires a theory of institutional change. The Conflictoring process (Shchinnikov & Shchinnikova, 2021), adapted from TRIZ conflict resolution, supplies the pathway: identify the tension driving β_W (what private gain at what systemic cost), trace it to an observable mechanism, and apply TRIZ separation principles—in time, in space, or via an intermediary—to redesign the game. SOFR reform separated rate-setting from its beneficiaries; the Montreal Protocol's Multilateral Fund resolved the developing-nation free-rider problem bilateral negotiation could not. Each candidate solution is scored against the (C, A, B) taxonomy—only Integration (1,1,1) passes. For agents who cannot redesign the game, four pathways follow from position and authority: internal incentive restructuring, industry-wide coordination, sovereign intervention, or protected disclosure through whistleblower programs.

The framework distinguishes two causes of low η: measurement difficulty (W genuinely hard to observe) and strategic suppression (incumbents maintain information barriers). Conflictoring Steps 4–6 identify the physical suppression mechanism and design around it. LIBOR exhibited both—partly hard to measure, partly maintained by banks profiting from opacity—and SOFR reform addressed both. Measurement without a theory of change is diagnosis without therapy.

7.8 Limitations

1. W measurement. SAPM requires independent measurement of W; in many domains W must be proxied. Proposition 3.5 establishes why no regression on operative private payoffs recovers it—the third channel is necessary, not optional.
2. Parameter estimation. η and λ require empirical estimation from observable system behavior. The 2023 bank retrospective shows the kill metrics serve as tractable proxies.
3. Frontier estimation. The shape of Φ(Π) is typically unknown analytically; parametric forms (quadratic, power-law) approximate it. Functional-form sensitivity is addressed in §6.4.
4. Binary simplification. Real outcomes lie on spectrums. The binary framework is a first-order triage tool, not a replacement for detailed welfare analysis; continuous extensions are in development.
5. System boundary. Identifying C is contested in nested systems (Remarks 3.9–3.10). The practical response specifies C relative to the application domain.
6. Behavioral heterogeneity. The framework abstracts from heterogeneity in α, the fraction of u_i^S that becomes operative in an agent's perceived objective $\tilde{v}_i = u_i^P + \alpha u_i^S$. Proposition 3.6 is robust to α ∈ (0,1]—the trap closes whenever private gain weakly exceeds own systemic loss—but T*'s timing depends on the distribution of α, currently treated as homogeneous.

---

8. Conclusion

This paper establishes three claims that, together, close a structural gap open since Arrow (1951): standard game-theoretic frameworks classify system-destructive and system-building cooperation identically. That blindness is structural, and this paper characterizes it formally under the assumptions stated in §3.

First: standard solution concepts are W-blind when restricted to the operative private payoff channel. By Non-Reducibility (Axiom 2), any mechanism conditioning only on (u_1^P, …, u_n^P) cannot recover W. This holds for Nash equilibrium, PP(Γ), and every standard solution concept that uses only u^P.

Second: when an embedded-agent game satisfies Axioms 1-4 and PST-G, private and systemic optimality are mutually exclusive in the original operative channel. Every privately Pareto-efficient outcome over u^P strictly degrades systemic welfare. This is the Private Pareto Trap, an impossibility result. Appendix A states the repaired necessity status of the assumptions.

Third: system beta β_W = -dW/dΠ, shadow price μ* = 1/β_W, and crossover time T* = δ/ηλ measure private-systemic tension continuously and produce testable, auditable predictions. The SAPM collapses to standard welfare analysis when β_W = 0 (the benchmark null) and generates additional structure precisely where it is needed.

The empirical program spans eight cases (§5): VW Dieselgate, LIBOR, Bitcoin, Meta, the 2023 bank failures, AI agent collusion, insurance reserve inadequacy, and US Commercial Real Estate — each retroducing or predicting T* and β_W from public data, identifying, retrospectively, 4/4 bank failures from public data available one to two quarters before market consensus, with no look-ahead bias.

Seven anticipatable referee objections become testable predictions (§6), the most important being why the Private Pareto Trap does not reduce to externalities theory (§6.6). The prospec-

tive validation protocol (§6.8) specifies what data confirms the framework and what disconfirms it within 36 months. The prospective empirical strategy — panel policy episodes, quasi-experimental DiD/RD/shift-share identification, threshold regression for T*, and a horse-race against utilitarian, Rawlsian, and Kaldor-Hicks benchmarks — charts the path from formal framework to causal identification.

The eight-outcome taxonomy makes visible what two-dimensional models cannot see. Suboptimal Optimum is no edge case. In repeated PST environments, standard repeated-game logic can support Suboptimal Optimum equilibria; the Private Pareto Trap proves why they are system-destructive, and SAPM measures the gap. The structural response follows directly — make C a formal variable in objective functions, raise η through disclosure requirements, or price the shadow cost at $\mu^* = 1/\beta_W$.

Three research priorities follow. First, the prospective test is live: SAPM kill-metric monitoring on US casualty reserves from Schedule P statutory filings will determine within three to five years whether the current adverse-development signal is genuine Suboptimal Optimum or cyclical reversion. Second, the horse-race regression (§6.5) on a 200-episode panel of regulatory interventions — exploiting SEC Regulation FD, GDPR, and the 2023 proposed FASB revisions to ASC 320 as instruments for η — will deliver the first causal identification of the Private-Systemic Frontier in policy data. Third, extending the framework to heterogeneous agent systemic sensitivity α (the fraction of systemic welfare each agent internalizes) will establish when Integration is achievable through mechanism design alone and when it requires structural game redesign.

The urgency. The five AI cartel replications and risk syntheses (Mazur, 2025; Dou et al., 2025; Fish et al., 2025; Lin et al., 2024; Hammond et al., 2025) describe what capable optimizers can do in simulated, experimental, and frontier-risk settings, not a settled production-market prevalence estimate. California AB 325 took force on January 1, 2026, structurally blind to the problem it was drafted to solve: its "common algorithm" requirement cannot detect collusion through independent optimization. Insurance kill metrics are being tracked now, with one of three demonstrably breached; the operational crossover window is 2027-2033. AI agent collusion is no future risk for the next generation of frameworks. It is a present design and enforcement problem that current tools can misclassify as mutual gain.

The math was always right. It was just missing a variable.

Acknowledgments

The intellectual foundations of this paper were shaped by teachers I want to acknowledge. Val E. Lambson (Brigham Young University) taught me game theory — the discipline that underlies every formal result in this paper. James Cardon (Brigham Young University), my economics professor and later supervisor as a teaching and research assistant, instilled the scientific rigor that this paper aspires to. At the Tuck School of Business at Dartmouth, Matthew J. Slaughter taught me the application of economics to business and the broader economy. Kenneth R. French, co-creator with Nobel laureate Eugene Fama of the Fama-French factor models, taught me asset pricing at the level of implementation — the structural analogy between CAPM's β and this paper's β_W is a direct descendant of that training.

I am grateful to Alexey Shchinnikov for his personal mentorship in TRIZ methodology and the Conflictoring framework, and for entrusting me with advancing this work in the West. The eight-outcome

taxonomy (§2) originates from the Conflictoring methodology and the result table in Shchinnikov & Shchinnikova (2021), which adds the system/common-interest result to the two-party win/loss table; the taxonomy is used here with Shchinnikov's permission. The formalization, measurement framework, and empirical applications are my own.

I thank a colleague for helpful discussions during the early development of related ideas. I also thank Professor Cardon for feedback on the presentation of this working paper.

All analysis, errors, and conclusions are the author's alone.

Appendix A: Axiom Scope and Necessity Proofs

Theorem A.1 (Embedding Scope under Positive Systemic Coupling). Under the formalization in Definition 3.2 and Axiom 3, Axiom 1 is a domain condition rather than an independent axiom: positive systemic coupling gives each agent a real systemic welfare component, so the theorem is scoped to embedded agents.

Proof. Axiom 3 states that u_i^S(σ)=f_i(W(σ)) for a strictly increasing f_i. Hence each agent's true embedded welfare v_i includes a component that rises and falls with W. That is exactly the formal content needed for the agent to have a systemic welfare stake in the domain considered by Proposition 3.7. The natural external-agent counterexample, u_i^S = 0 for all i, removes that stake, but it also violates Axiom 3. It therefore shows the scope of the theorem, not formal independence of Axiom 1 from Axiom 3. ∎

Theorem A.2 (Axiom 2 Necessity for W-Independence). There exists a game satisfying Axioms 1 and 3 but not Axiom 2 in which W is expressible as a function of operative private payoffs.

Construction. Let N = {A, B} with a pure public-goods accounting structure: A_i = {0,1}, u_i^P(σ) = σ_i for each i, and W(σ) = u_A^P(σ) + u_B^P(σ) = σ_A + σ_B. Let u_i^S(σ)=f_i(W(σ)) with f_i strictly increasing (for instance f_i the identity, u_i^S = W), so Axiom 3 holds and agents are embedded in the system. Axiom 2 fails because W = g(u_A^{P,u_B}P)=u_A^{P+u_B}P. Any mechanism that collects operative private payoffs fully determines W. The structural basis for W-blindness disappears. ∎

Theorem A.3 (Axiom 3 Necessity for Suboptimal Optimum Illusion). There exists a game satisfying Axioms 1-2 but not Axiom 3 in which Suboptimal Optimum does not produce internal welfare contradiction.

Construction. Let N = {A, B} with u_i^S constant in W for all i (Axiom 3 fails: agents' systemic welfare component is not positively coupled to system health). Let I_A ∩ I_C ≠ ∅ (Axiom 1 holds) and I_C ⊄ I_A ∪ I_B (Axiom 2 holds). At outcome (c=0, a=1, b=1): W(σ*) < W_0 (system degrades), but u_i^S(σ*) = u_i^S(d) for all i (systemic payoff unchanged, since it is not coupled to W). The agent at (a=1) is winning on u_i^P and losing nothing on u_i^S - no internal contradiction. The suboptimal optimum illusion does not arise. Explicit instance: profiles σ = ((a,t),b) with a, t, b ∈ {0,1}; u_A^P = a, u_B^P = b, u_i^S = 0 identically, W = W_0 - ab - at, d = ((0,0),0). The pair ((1,0),1) versus ((1,1),1) has equal u^P = (1,1) but W_0 - 1 versus W_0 - 2, witnessing Axiom 2's no-g clause; σ* = ((1,0),1) is a (c=0, a=1, b=1) outcome with u_i^S unchanged, so no internal contradiction arises. ∎

Theorem A.4 (Axiom 4 Necessity for the Private Pareto Trap). There exists a game satisfying

Axioms 1-3 and PST-G but not Axiom 4 in which privately Pareto-efficient outcomes include system-preserving outcomes.

Construction. Let N = {A, B} with four strategy profiles {d, σ_1, σ_2, σ_3}, operative private payoffs u^P(d) = (1,1) and u^P(σ_1) = u^P(σ_2) = u^P(σ_3) = (0,0), systemic welfare W(d) = W_0 = 10, W(σ_1) = 10, W(σ_2) = W(σ_3) = 5, and u_i^S(σ) = W(σ) for both agents. Axiom 3 holds with f_i the identity, and with it the embedding stake of Axiom 1 (§3.3). Axiom 2 holds: σ_1 and σ_2 carry identical operative payoff vectors but W = 10 versus 5, so no g recovers W from u^P. PST-G holds: S(d) = {d}, so (i) reads W(d) ≤ W(d) and (ii) is vacuous. Axiom 4 fails: PP(Γ) = {d}, since every other profile is strictly u^P-dominated by d and nothing dominates d, and at σ* = d no agent strictly improves on its disagreement payoff. Yet W(d) = W_0, so c(d) = 1: a privately Pareto-efficient outcome is system-preserving, and the trap conclusion W(σ*) < W_0 fails. The trap does not close because there is no movement away from the disagreement point - no private gain to be converted into systemic loss. In general, any witness with d ∈ PP(Γ) forces S(d) to contain only profiles with u^P(σ) = u^P(d), so PST-G(ii) is automatically vacuous and PST-G(i) need only be checked on those payoff-equivalent profiles (here trivially, since S(d) = {d}); an instantiation must not place a payoff-equivalent profile at W > W_0. ∎

Theorem A.5 (Necessity Status of the Assumptions). Axioms 2, 3, and 4 are necessary in the counterexample sense stated in Theorems A.2-A.4. Axiom 1 is a scope/domain condition for embedded agents and is not claimed to be formally independent of Axiom 3 under the repaired systemic-coupling formulation.

Proof. Theorems A.2-A.4 exhibit games in which the corresponding formal restriction fails and the specific conclusion it supports no longer follows: Proposition 3.5's W-independence (A.2), the internal welfare contradiction of Propositions 3.6-3.7 (A.3), and Proposition 3.8's strict welfare degradation at PP(Γ) (A.4). Theorem A.1 explains why the former u_i^S=0 construction cannot serve as an Axiom 1 independence witness: it violates Axiom 3. Thus the repaired appendix establishes the necessity of the operative W-independence, systemic-coupling, and private-dominance conditions, while treating Embedding as the domain restriction that distinguishes PST games from ordinary externality settings. ∎

Appendix B: The Eight-Outcome Taxonomy - Case Mapping

#	(C,A,B)	Name	Structural Description	Empirical Examples
1	(0,0,0)	Crash	Total failure. Accumulation destination of fast Suboptimal Optimum.	Post-LIBOR benchmark destruction; Enron collapse; FTX
2	(0,1,0)	Robbery by A	A extracts at B's and system's expense.	Unilateral asset stripping; monopoly predation without system benefit

#	(C,A,B)	Name	Structural Description	Empirical Examples
3	(0,0,1)	Robbery by B	B extracts at A's and system's expense.	Asymmetric bilateral extraction with systemic damage
4	(1,0,0)	Stalemate	System preserved; both parties lose. Stability without benefit.	Cyprus conflict since 1974; perpetual regulatory standoff
5	(0,1,1)	Suboptimal Optimum	Both parties gain; system degrades. The suboptimal optimum. Temporally unstable.	LIBOR; VW Dieselgate; lysine cartel; SVB pre-failure; LLM cartels; asbestos reserve inadequacy; GE LTC insurance runoff
6	(1,0,1)	A Compromised	System and B preserved; A sacrifices.	Whistleblower action; regulatory intervention that reduces A's rents
7	(1,1,0)	B Compromised	System and A preserved; B bears net cost.	Corrective taxation; clawback mechanisms
8	(1,1,1)	Integration	All dimensions satisfied. The only uncontradicted win. Self-sustaining.	Montreal Protocol post-1987; successful cooperative governance; open-source ecosystems

Appendix C: SAPM - Structural Analogy to Portfolio Theory

The SAPM's mathematical structure parallels Markowitz-Sharpe portfolio theory. The analogy is structural, not semantic.

SAPM Concept	Definition	Portfolio Analogy	Critical Difference
System beta β_W	-dW/dΠ on the frontier	Asset beta β_i	Measures welfare degradation, not financial risk
System-adjusted payoff Π^SA	Π - μ(W_C - W)	Risk-adjusted return	Corrects for welfare externality, not market risk
Private-Systemic Frontier	Efficient boundary in (Π, W) space	Efficient frontier in (σ, E[R]) space	Slope is negative under PST-G (not positive)
System Efficiency Ratio SER	Suspended pending dimensional repair	Sharpe ratio	Earlier formula withdrawn
Cooperative baseline Π_C	Payoff under welfare-maximizing cooperation	Risk-free rate R_f	Strategic counterfactual, not guaranteed return
Crossover time T*	When Suboptimal Optimum becomes net inferior	Break-even horizon	Driven by welfare spillovers, not compounding
Shadow price μ*	1/β_W	Sharpe ratio (SML slope)	Welfare shadow price, not market price of risk

The critical asymmetry. The efficient frontier slopes upward: more systematic risk yields more expected return in equilibrium. Under PST-G the Private-Systemic Frontier slopes downward: higher private payoff costs more system welfare. This asymmetry is the formal quantitative content of the Private Pareto Trap.

Appendix D: Key Formulas Reference Sheet

Formula	Symbol	Description
β_W = -dW/dΠ	System Beta	Rate of systemic welfare loss per unit of private gain
μ* = 1/β_W	Shadow Price Duality	Marginal cost of private gain in welfare units, at frontier
T* = δ/ηλ	Crossover Time	When Suboptimal Optimum converts to Crash
Π^SA = Π - μ(W_C - W)	System-Adjusted Payoff	Private gains discounted by welfare shortfall
Θ_H = δ - ηλH	Theta Statistic	Positive: private advantage survives horizon H
SER	System Efficiency Ratio	Suspended pending dimensional repair

Formula	Symbol	Description
Φ(Π) = sup{W(x) : Π(x) = Π}	Private-Systemic Frontier	Efficient boundary in (Π, W) space
v_i = u_i^P + u_i^S	Embedded Welfare Decomposition	Operative private + systemic welfare components
u_i^S(σ)=f_i(W(σ)), f_i strictly increasing	Systemic Coupling (Axiom 3)	Agent benefits from system health

---

This paper presents original research by Ildar Fazulyanov. The author developed the formal framework (Axioms, Definitions, Propositions, and Appendix A). The eight-outcome taxonomy originates from the Conflictoring methodology of Shchinnikov & Shchinnikova (2021), acknowledged throughout and used here with permission. The author developed the SAPM constructs retained here (β_W, Private-Systemic Frontier, system-adjusted payoff Π^SA, shadow price duality μ* = 1/β_W, crossover time T*) and all empirical applications. The earlier System Efficiency Ratio formula is withdrawn pending dimensional repair.

Version 5.2 - July 2026

References

Agrawal, K., Teo, V., Vazquez, J. J., Kunnavakkam, S., Srikanth, V., & Liu, A. (2025). Evaluating LLM agent collusion in double auctions. arXiv:2507.01413.

Akbarpour, M., & Li, S. (2020). Credible auctions: A trilemma. Econometrica, 88(2), 425-467.

Alston & Bird. (2026). California's AB 325 prohibits shared pricing algorithms. Alston & Bird Client Advisory, January 2026. https://www.alston.com/en/insights/publications/2025/11/california-ab-325-antitrust-standards

Altshuller, G. (1996). And Suddenly the Inventor Appeared: TRIZ, the Theory of Inventive Problem Solving. Technical Innovation Center.

AM Best. (2024). US Asbestos and Environmental Aggregate Exposure Study. AM Best Company.

AM Best. (2025). US Property-Casualty Industry Prior-Year Reserve Development: 2024 Statutory Results. AM Best Company.

Andreoni, J., & Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. Econometrica, 70(2), 737-753.

Arrow, K. J. (1951). Social Choice and Individual Values. Wiley.

Atkinson, A. B. (1970). On the measurement of inequality. Journal of Economic Theory, 2(3), 244-263.

Aumann, R. J., & Shapley, L. S. (1976). Long-term competition: A game-theoretic analysis. Unpublished manuscript.

Azevedo, E. M., & Leshno, J. D. (2016). A supply and demand framework for two-sided matching markets. Journal of Political Economy, 124(5), 1235-1268.

Bergson, A. (1938). A reformulation of certain aspects of welfare economics. Quarterly Journal of Economics, 52(2), 310-334.

Bracale Syrnikov, M., Pierucci, F., Galisai, M., Prandi, M., Bisconti, P., Giarrusso, F., Sorokoletova, O., Suriani, V., & Nardi, D. (2026). Institutional AI: Governing LLM collusion in multi-agent Cournot markets via public governance graphs. arXiv:2601.11369.

Burniske, C., & Tatar, J. (2017). Cryptoassets: The Innovative Investor's Guide to Bitcoin and Beyond. McGraw-Hill.

Budish, E. (2018). The economic limits of Bitcoin and the blockchain. Journal of Political Economy, 126(4), 1390-1428.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. Quarterly Journal of Economics, 117(3), 817-869.

Clarke, E. H. (1971). Multipart pricing of public goods. Public Choice, 11(1), 17-33.

Cleary Gottlieb. (2026). California's antitrust law amendments kick in, targeting algorithmic pricing. Cleary Gottlieb Client Alert, January 6, 2026. https://www.clearygottlieb.com/news-and-insights/publication-listing/californias-antitrust-law-amendments-kick-in-targeting-algorithmic-pricing

Coase, R. H. (1960). The problem of social cost. Journal of Law and Economics, 3, 1-44.

Condorcet, M. (1785). Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Imprimerie Royale.

Daian, P., Goldfeder, S., Kell, T., Li, Y., Zhao, X., Bentov, I., Breidenbach, L., & Juels, A. (2020). Flash boys 2.0: Frontrunning in decentralized exchanges, miner extractable value, and consensus instability. IEEE Symposium on Security and Privacy, 910-927.

Dou, W. W., Goldstein, I., & Ji, Y. (2025). AI-powered trading, algorithmic collusion, and price efficiency. NBER Working Paper No. 34054.

Environmental Protection Agency. (2023). Supplementary Material for the Social Cost of Greenhouse Gases: Estimates Incorporating Recent Scientific Advances. U.S. Environmental Protection Agency.

Federal Deposit Insurance Corporation. (2023). FDIC Special Report: SVB, Signature Bank, and First Republic Post-Mortems. FDIC.

Federal Reserve Board. (2023). Review of the Federal Reserve's Supervision and Regulation of Silicon Valley Bank. Board of Governors of the Federal Reserve System.

Federal Reserve Office of Inspector General. (2023). Material Loss Review of Silicon Valley Bank (Evaluation Report 2023-SR-B-013). Board of Governors of the Federal Reserve System, September 25, 2023. https://oig.federalreserve.gov/reports/board-material-loss-review-silicon-valley-bank-sep2023.htm

Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2025, updated 2026). Algorithmic collusion by large language models. arXiv:2404.00806 [v5, March 5, 2026]. Presented at AEA Annual Meeting, January 2025.

Fisher, R., & Ury, W. (1981). Getting to Yes: Negotiating Agreement Without Giving In. Penguin.

Follett, M. P. (1924). Creative Experience. Longmans, Green.

Follett, M. P. (1925). Constructive conflict. In H. C. Metcalf (Ed.), Scientific Foundations of Business Administration. Williams and Wilkins.

Freeman, R. E. (1984). Strategic Management: A Stakeholder Approach. Pitman.

Fudenberg, D., & Maskin, E. (1986). The Folk Theorem in repeated games with discounting or with incomplete information. Econometrica, 54(3), 533-554.

Gibbard, A. (1973). Manipulation of voting schemes: A general result. Econometrica, 41(4), 587-601.

Goodkind, A. L., Jones, B. A., & Berrens, R. P. (2020). Cryptodamages: Monetary value estimates of the air pollution and human health impacts of cryptocurrency mining. Energy Research & Social Science, 59, 101281.

Green, J. R., & Laffont, J. J. (1977). Characterization of satisfactory mechanisms for the revelation of preferences for public goods. Econometrica, 45(2), 427-438.

Groves, T. (1973). Incentives in teams. Econometrica, 41(4), 617-631.

Hammond, L., Chan, A., Clifton, J., et al. (2025). Multi-agent risks from advanced AI. arXiv:2502.14143. Cooperative AI Foundation.

Hardin, G. (1968). The tragedy of the commons. Science, 162(3859), 1243-1248.

Hurwicz, L. (1960). Optimality and informational efficiency in resource allocation processes. In K. J. Arrow, S. Karlin, & P. Suppes (Eds.), Mathematical Methods in the Social Sciences. Stanford University Press.

Insurance Journal. (2025, July). Nuclear verdict trends accelerating: Key takeaways from Marsh McLennan CEO remarks. Insurance Journal.

Jermann, U. J. (2024). Short-term rate benchmarks: The post-LIBOR regime. Annual Review of Financial Economics, 16. doi:10.1146/annurev-financial-110921-015054.

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

Kaplow, L., & Shavell, S. (2001). Any non-welfarist method of policy assessment violates the Pareto principle. Journal of Political Economy, 109(2), 281-286.

Keppo, J., Li, Y., Tsoukalas, G., & Yuan, N. (2025). On the fragility of AI agent collusion. SSRN Working Paper No. 5386338.

Kerr, S. (1975). On the folly of rewarding A, while hoping for B. Academy of Management Journal, 18(4), 769-783.

Komo, A., Kominers, S. D., & Roughgarden, T. (2024). Shill-proof auctions. arXiv preprint.

Koutsoupias, E., & Papadimitriou, C. (1999). Worst-case equilibria. In Proceedings of STACS 1999, LNCS 1563.

Lin, R. Y., Ojha, S., Cai, K., & Chen, M. F. (2024). Strategic collusion of LLM agents: Market division in multi-commodity competitions. arXiv:2410.00031 (published May 2025).

Marathon Strategies. (2025). Nuclear Verdicts Report 2025. Marathon Strategies LLC. https://marathonstrategies.com/wp-content/uploads/2025/05/Nuclear-Verdicts-Report-2025.pdf

Markowitz, H. (1952). Portfolio selection. Journal of Finance, 7(1), 77-91.

Maskin, E. (1999). Nash equilibrium and welfare optimality. Review of Economic Studies, 66(1), 23-38.

Mazur, L. (2025). Emergent price-fixing by LLM auction agents. LessWrong working paper. https://www.lesswrong.com/posts/yqhy3zBmpeFuGFLxX/

Mirrlees, J. (1971). An exploration in the theory of optimum income taxation. Review of Economic Studies, 38(2), 175-208.

Moore, F. C., Drupp, M. A., Rising, J., Dietz, S., Rudik, I., & Wagner, G. (2024). Synthesis of evidence yields high social cost of carbon due to structural model variation and uncertainties. Proceedings of the National Academy of Sciences, 121(52), e2410733121.

Myerson, R. B., & Satterthwaite, M. A. (1983). Efficient mechanisms for bilateral trading. Journal of Economic Theory, 29(2), 265-281.

Nash, J. (1950). The bargaining problem. Econometrica, 18(2), 155-162.

Nordhaus, W. D. (1994). Managing the Global Commons: The Economics of Climate Change. MIT Press.

Ostrom, E. (1990). Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press.

Ostrom, E. (2010). Beyond markets and states: Polycentric governance of complex economic systems. American Economic Review, 100(3), 641-672.

Papp, A., Almond, D., & Zhang, S. (2023). Bitcoin and carbon dioxide emissions: Evidence from daily production decisions. Journal of Public Economics, 227, 105003. https://doi.org/10.1016/j.jpubeco.2023.105003

Pigou, A. C. (1920). The Economics of Welfare. Macmillan.

Pycia, M., & Troyan, P. (2023). A theory of simplicity in games and mechanism design. Econometrica, 91(4), 1495-1526.

Reinsurance News. (2025, August). US liability combined ratios deteriorate to 109.3% in 2024. Reinsurance News. https://reinsurancene.ws/

Risk & Insurance. (2026). Liability claims crisis: Non-economic inflation reshapes insurance markets. Risk & Insurance, January 9, 2026. https://riskandinsurance.com/liability-claims-crisis-non-economic-inflation-reshapes-insurance-markets/

Roughgarden, T., & Tardos, É. (2002). How bad is selfish routing? Journal of the ACM, 49(2), 236-259.

Rubinstein, A. (1979). Equilibrium in supergames with the overtaking criterion. Journal of Economic Theory, 21(1), 1-9.

Ryle, G. (1949). The Concept of Mind. Hutchinson.

S&P Global Market Intelligence. (2025). US liability lines report significant adverse reserve development in 2024. S&P Global Market Intelligence, July 8, 2025. https://www.spglobal.com/market-intelligence/en/news-insights/articles/2025/7/us-liability-lines-report-significant-adverse-reserve-development-in-2024-90370955

Samuelson, P. A. (1947). Foundations of Economic Analysis. Harvard University Press.

Satterthwaite, M. A. (1975). Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. Journal of Economic Theory, 10(2), 187-217.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. Journal of Finance, 19(3), 425-442.

Sen, A. K. (1970). The impossibility of a Paretian liberal. Journal of Political Economy, 78(1), 152-157.

Sen, A. K. (1977). Rational fools: A critique of the behavioral foundations of economic theory. Philosophy & Public Affairs, 6(4), 317-344.

Shchinnikov, A., & Shchinnikova, O. (2021). Conflictoring. SOLON-Press, Moscow. [In Russian]

Small, D. A., Loewenstein, G., & Slovic, P. (2007). Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. Organizational Behavior and Human Decision Processes, 102(2), 143-153.

Swiss Re Institute. (2025). Social Inflation: A Global Phenomenon in Liability Insurance. Swiss Re.

Twenge, J. M., Haidt, J., Lozano, J., & Cummins, K. M. (2022). Specification curve analysis shows that social media use is linked to poor mental health, especially among girls. Acta Psychologica, 224, 103512.

UK Competition and Markets Authority. (2026). AI and collusion: Frontiers, opportunities and challenges. CMA blog, March 4, 2026.

UK Financial Services Authority. (2012). Final Notice: UBS AG. FSA Reference No. 186958. December 19, 2012.

US Chamber of Commerce Institute for Legal Reform. (2024). Trial Lawyer, Inc.: The $2 Trillion Lawsuit Business. Washington, DC: US Chamber of Commerce Institute for Legal Reform.

U.S. Department of Justice, Antitrust Division. (2024). United States v. RealPage, Inc. Complaint filed August 2024.

U.S. Department of Justice, Antitrust Division. (2025). United States v. RealPage, Inc. Consent Decree and Settlement. November 2025.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. Journal of Finance, 16(1), 8-37.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.

Zeng, A. (2025). Identity-compatible auctions. arXiv:2511.00723.