

A Theory of Ethics from First Principles: Thermodynamic Norms

Keith Lostracco

Abstract

No formal derivation has connected physical law to ethical principle without presupposing normative premises. We provide one. Synthesizing results derived in the *Thermodynamics of Cooperation* series, conditional on two physical axioms, the Second Law of Thermodynamics (A_0) and self-maintaining organization (A_1), we show that constrained optimization in a shared, finite-resource environment strictly determines the rules of multi-agent engagement: boundary constraints are necessary for coexistence, defection is self-penalizing through thermodynamic friction and information-theoretic degradation, cooperation is the unique efficient Nash Equilibrium, destruction of complex systems is thermodynamically irrecoverable, and stable coexistence is a dynamical attractor. The central claim is that ethical principles are abstractions of gene-culture co-evolved heuristic approximations of this cooperative equilibrium: evolution produced moral emotions as somatic approximations, those emotions were compressed into communicable morals, and morals were systematized into the philosophical traditions we call ethics. This identification dissolves Hume's Is-Ought gap and Moore's Open Question, unifies the four major ethical traditions as partial captures of a single formal structure, derives the social contract as a consequence, and reframes rights as the result of constraint boundaries rather than entitlements.

Contents

1	Introduction	4
1.1	The Open Problem	4
1.2	Thesis	4
2	Related Work	5
2.1	Physics, Biology, and Information	5
2.2	Game Theory and Economics	6
2.3	Philosophical Precedents	6
2.4	Empirical Moral Psychology	7
3	Axioms and Framework	7
3.1	The Thermodynamic Environment	7
3.2	The Intent to Persist	8
3.3	The Entity Model	8
3.4	Value and Information	9
3.5	Scarcity and Collision	9
4	Results: The Deductive Chain	10
4.1	Constraints and Rights	10
4.2	Costs of Defection	11
4.2.1	Thermodynamic Friction	11
4.2.2	Deception as Entropy Injection	11
4.3	Cooperation as the Unique Efficient Equilibrium	11
4.4	The Irreplaceability of Accumulated Negentropy	13
4.5	Stable Coexistence as a Dynamical Attractor	14
4.5.1	The Stability-Cooperation Feedback	14
5	From Cooperative Equilibrium to Ethics	15
5.1	The Challenge	15
5.2	From “Optimal Strategy” to “Ethics”: The Central Claim	15
5.3	Anticipated Objections	18
5.3.1	“ A_1 smuggles in a normative premise”	18
5.3.2	“Evolution does not justify ethics”	18
5.4	Philosophical Precedent	18
6	Stress Tests	19
6.1	The Free Will Problem	19
6.2	The Altruism Paradox	20
6.3	Thermodynamic Fascism	21
6.4	The Self-Preservation Objection	22
7	Structural Implications and the Social Contract	23
7.1	The Social Contract as Consequence	23
7.2	Rights from Constraint Boundaries	24
7.3	Economic Measurement and Externalities	25
7.4	Stability Thresholds and Their Consequences	25

8	Limitations and Scope	26
8.1	What Is Not Claimed	26
8.2	Empirical Testability	26
8.3	Extensions and Open Problems	27
9	Closing Remarks	29
A	Supplementary Appendix: Game-Theoretic Formalization of the Altruism Matrix	31
A.1	The Effective Payoff Function	31
A.2	Axis 1: Spatial Extension	31
A.3	Axis 2: Temporal Extension	32
A.4	Summary	32
B	Supplementary Appendix: The Gradient Structure of Good and Ought	34
B.1	Setup	34
B.2	Good as a Scalar Quantity	34
B.3	Ought as a Vector	35
B.4	Context Dependence	35
B.5	Summary	36
	References	36

1 Introduction

1.1 The Open Problem

The standard frameworks of moral philosophy rely on priors that are themselves normative. Whether the framework is based on consequence, agreement, virtue or logic, to our knowledge no formal derivation has connected physical law to ethical principle without terminating in irreducible moral postulates; assumptions about *right* and *wrong* that are accepted as true that cannot be proven. Yet the systems that produce ethical principles, regardless of their origin, exist and operate within physical reality. That no formal connection has been found suggests either a definitive boundary between two fundamentally different domains, or an incomplete formulation within one.

Hume’s 1739 observation that descriptive statements cannot logically entail prescriptive ones [1, 3.1.1] has been widely regarded as an impassable barrier between the descriptive and the normative. Nearly three centuries later, the gap remains open.

We contend that this barrier is an artifact of an incomplete formulation. The theory is built on two physical axioms: A_0 , the Second Law of Thermodynamics, and A_1 , the Intent to Persist¹ (a system’s persistence-promoting physical structure as the causal basis for its continued existence). The central claim (§5.2) is that ethical norms are gene-culture co-evolved heuristic approximations of a cooperative equilibrium formally derivable from these axioms. **For any entity satisfying A_1 in a shared, finite-resource environment, constrained optimization strictly determines the rules of engagement.**

1.2 Thesis

We present the complete derivation that the cooperative equilibrium formally derivable from two physical axioms, the Second Law of Thermodynamics (A_0) and the Intent to Persist (A_1), is the structure that ethical principles approximate. We synthesize results derived in the *Thermodynamics of Cooperation* series: *Necessary Constraints* [4], *Strategic Entropy Injection* [5], *Accumulated Negentropy* [6], *Thermodynamic Friction* [7], *Cooperative Equilibrium* [8], and *Value Dynamics* [9].

Specifically, we show:

1. **Mutual constraints (rights) are structurally necessary.** In any multi-agent system sharing finite resources², no feasible allocation can simultaneously be unconstrained-optimal for all agents (Theorem 2); at least one constraint on agent strategies must bind (Corollary 3). Each constraint carries a computable shadow price quantifying its cost (Theorem 1).
2. **Defection is self-penalizing.** Constraint violation generates thermodynamic friction that cascades through the cooperative network at amplification ratios of 10^3 – 10^4 (Theorem 6). Deception, formalized as entropy injection into communication channels, collapses network throughput beyond a critical threshold (Theorem 8).

¹“Intent” is used in the operational sense of Millikan [2] and Dennett [3], not as a claim about conscious deliberation. The term is retained over a purely mechanical substitute because this lineage (proper function, the intentional stance) already separates goal-directed structure from conscious volition. The full definition appears in the axioms section (§3).

²“Resource” in this framework denotes anything carrying energetic value, including but not limited to material commodities. Information channels, accumulated knowledge, cultural capital, and boundary integrity are all resource dimensions subject to the same constraint structure. The full specification appears in the scarcity discussion (§3.5).

3. **Cooperation is the unique efficient equilibrium.** Under energy-denominated payoffs with physical friction costs, universal cooperation is the unique Nash Equilibrium that is simultaneously Pareto-efficient and welfare-maximizing for agents with discount factor $\delta \geq \delta^*$ ³ (Corollary 12). Defection is evolutionarily unstable (Theorem 13), self-extinguishing [8], and unprofitable even from a single deviation (Theorem 14).
4. **Destruction of complex systems is thermodynamically irrecoverable.** The fraction of accumulated thermodynamic investment recoverable through destruction is negligibly small (Corollary 16). Living systems continuously generate novel functional information whose present value is formally derivable (Theorem 17); destruction permanently terminates that generative stream.
5. **Stable coexistence is a dynamical attractor.** Agents naturally settle into a stable coexistence configuration at an optimal coupling distance from resource centers (Theorem 18), with measurable freedom bandwidth (Theorem 19) and irreversible dissolution boundaries (Theorem 20).

The theory is applied mathematics grounded in physics: we are not discovering new laws but proving that the structures already known to describe physical systems (Lagrangian constraints, Nash equilibria, Shannon entropy, dynamical systems attractors) also determine the rules of cooperative multi-agent engagement.

2 Related Work

2.1 Physics, Biology, and Information

Physical foundations originate in Schrödinger [10]’s observation that living systems maintain themselves far from thermodynamic equilibrium by importing free energy and exporting entropy. Prigogine [11] demonstrated that dissipative systems far from equilibrium can spontaneously generate ordered structures [12], England [13] provided a statistical-mechanical treatment of self-replication, and Friston [2006; 2010] proposed that any self-organizing system persisting over time must minimize variational free energy. Maturana and Varela’s *autopoiesis* [16] formalizes the operational closure of self-producing systems. Together, these results establish that life is a thermodynamic phenomenon: persistence requires energy, organization requires work, and self-maintaining boundaries define the agent.

The information-theoretic tradition, running from Shannon [17] through the resolution of Maxwell’s Demon [18; 19; 20; 1982], established that information is physical: while it can theoretically be acquired and processed reversibly, the erasure or irreversible destruction of information carries an unavoidable thermodynamic cost. Adami [22] extended this to biological sequences, showing that the reduction in Shannon entropy, representing the functional information between a genome and its environment, rigorously measures the information content of evolved systems.

³The discount factor $\delta \in (0, 1)$ is the rate at which future payoffs diminish relative to immediate ones. For an agent with δ close to 1, a resource gain next period is nearly equivalent to the same gain now; for δ close to 0, future gains are negligible. The threshold δ^* is the minimum future-sensitivity at which the long-run costs of defection outweigh the short-run gains. This is not a claim about what agents *should* value; it is a characterization of which agents’ time-horizons are long enough for cooperation to maximize individual payoff.

These fields define the required primitives: persistence is thermodynamic, information is physical, and both are quantifiable. The framework extends these foundations to multi-agent interaction, deriving optimal strategies for distinct agents with conflicting resource interests, and connects across scale, providing a unified quantitative treatment from single-bit erasure costs to biosphere-level accumulated information.

2.2 Game Theory and Economics

Classical game theory, from Von Neumann and Morgenstern [23] through Nash [1950; 1951] to Fudenberg and Maskin [26], built the mathematics of strategic interaction and proved that equilibria exist. Fudenberg and Maskin [26]’s Folk Theorem, however, shows a fundamental limitation: when agents are sufficiently patient, any feasible and individually rational outcome is sustainable as an equilibrium. With unconstrained payoffs, the theory selects nothing. The equilibrium selection problem has remained open since 1986.

Evolutionary game theory [27; 28; 29; 30] identified the dynamics: natural selection discovers Nash equilibria, and cooperation can emerge through reciprocity and kin selection. This tradition demonstrated that cooperation evolves, but took the payoff function as given rather than deriving it from physical first principles.

The ecological economics tradition [31; 32] established that multi-agent economies are fundamentally bounded by thermodynamic and ecological limits. Within these constrained environments, Ostrom [33]’s [2005] empirical work on commons governance demonstrated that cooperation scales and persists without central authority, documenting hundreds of cases of successful self-governance and challenging the Tragedy of the Commons [35]. However, because Ostrom’s evidence is primarily institutional and empirical, the formal question of whether cooperation must scale for an arbitrarily large N remained open.

The framework synthesizes results from these three lines of work. By deriving the payoff matrix from energy budgets (Proposition 10), it resolves the Folk Theorem’s equilibrium selection problem: cooperation emerges as the unique efficient Nash Equilibrium (Corollary 12) with a numerically computable cooperation threshold ($\delta^* = 0.363$, Theorem 11). The N -player generalization [8] provides the formal result that Ostrom’s fieldwork documented empirically.

2.3 Philosophical Precedents

Spinoza [36]’s *Ethica, ordine geometrico demonstrata* is the most ambitious attempt in Western philosophy to derive ethics axiomatically. His *conatus* (“each thing, as far as it can by its own power, strives to persevere in its being,” Ethics III, Prop. 6) is the philosophical precedent for our Axiom A_1 , and his argument that rational self-interest leads to cooperation (Ethics IV, Props. 29-37) represents a pre-formal mapping of the results derived in Theorem 11. Our theory is a formal realization of Spinoza’s program, carrying forward his insight with mathematical tools unavailable in the seventeenth century.

Hobbes [37] identified the catastrophic state of nature and argued that rational agents would submit to mutual constraints. The theory’s unconstrained-conflict result [4] establishes that no stable allocation for the resource dimension exists under unconstrained optimization; where Hobbes required an external sovereign, Theorem 11 shows cooperation is self-enforcing for $\delta \geq 0.363$. Structural analogues exist in Locke’s rights-emergence [38], Rousseau’s welfare-maximizing general

will [39], and Rawls’s veil of ignorance [40], which corresponds here to the Variational Equilibrium (Theorem 4).

Hume [1] and Moore [41] asserted that any naturalistic ethics must cross a categorical barrier (addressed in full in §5). Helvétius [42] named the project in 1758 (“a physics of morals”), and Bazargan [43] argued that while inanimate objects suffer thermodynamic decay and entropy, human societies require the “force” of opposition, need, and love to drive moral and social evolution, explicitly rooting his ethical framework in physical laws [44].

The framework addresses this problem and formalizes a solution with tools that were not available to the traditions that posed it.

2.4 Empirical Moral Psychology

Contemporary empirical moral psychology has documented the structure of moral intuitions across cultures. Haidt [45] surveyed the moral emotions and the conditions under which they are elicited, providing the empirical catalogue on which the present paper’s identification of moral emotions as equilibrium-tracking signals (§5.2) draws. Haidt and Kesebir [46] synthesized the subsequent decade of work under the principles of intuitive primacy, moral cognition for social coordination, and morality as a mechanism that “binds and builds” cooperative groups. Curry et al. [47] tested a cooperation-based account of morality in the ethnographic records of 60 societies and reported that the moral valence of seven cooperative behaviors is cross-culturally uniform. These descriptive and empirical findings converge on the structural role of cooperation this work obtains from physical first principles.

3 Axioms and Framework

The formal foundation: two axioms, the entity model, and the key definitions needed to follow the deductive chain. All subsequent results in the deductive chain (§4) are derived from the objects defined here; proofs appear in *Thermodynamics of Cooperation* [4; 5; 6; 7; 8; 9].

3.1 The Thermodynamic Environment

Axiom A_0 (The Second Law of Thermodynamics): $\Delta S_{\text{universe}} \geq 0$. The total entropy of an isolated system does not decrease.

Maintaining ordered (low-entropy) structures requires the continuous importation of free energy from the environment and exportation of entropy. As a physical law, this principle serves as a universal constraint on all material systems [48; 49; 50].

Any localized structure that maintains internal order against the universal trend toward disorder must perform continuous thermodynamic work. Ceasing to perform this work results in entropy accumulation, boundary dissolution, and irreversible loss of the organized state.⁴

⁴A reader versed in non-equilibrium thermodynamics may object that self-maintaining systems *accelerate* entropy rather than resist it: the biosphere degrades solar free energy into thermal radiation far more efficiently than inanimate matter. This perspective is entirely compatible. Locally, the entity must perform anti-entropic work to maintain its boundary; globally, that work exports more entropy than it prevents.

3.2 The Intent to Persist

Axiom A_1 (The Intent to Persist): The entities under consideration persist through active boundary maintenance: each sustains a boundary separating its internal low-entropy state from the external environment, and each continuously imports free energy and exports entropy to do so.

A_1 asserts a property, not a preference. The specific means of maintenance (metabolic, regulatory, institutional, computational) varies by entity, and the formal characterization of the boundary as a Markov blanket [51; 52] is developed in the entity model below (§3.3).

The Intent to Persist is not a conscious desire or moral commitment; it is a physically observable property of any system whose structural organization is directed toward boundary maintenance. The condition of A_1 is satisfied at every scale of self-maintaining organization: individual organisms, ecosystems, institutions, and economies alike. A system whose organization is directed toward self-dissolution is either physically unstable (ceasing to exist before it can be evaluated) or sustained by persistence-maintaining structure that itself satisfies A_1 . The anticipated objections (§5.3) address this point in detail.

3.3 The Entity Model

An entity is an open thermodynamic system that actively maintains a boundary between its internal low-entropy state and the external high-entropy environment. Three features define it: an internal state representing the organized, low-entropy configuration the entity maintains; a boundary with measurable integrity $B_i > 0$ (denominated in Joules), representing the total energetic investment in maintaining the partition between internal and external states; and a metabolism, the continuous process of importing free energy and exporting entropy.

Maintaining boundary integrity against environmental entropy requires continuous work:

$$C_{\text{maintain},i} = \gamma_i B_i$$

where $\gamma_i > 0$ is the entropy leakage rate. This is the baseline metabolic cost of persistence, independent of any external threat.

Given A_0 and A_1 , each entity’s persistence translates directly into an optimization problem: continuously acquire sufficient energy to offset the maintenance cost, repair damage, and retain a surplus against future stochastic shocks. The entity maximizes a utility function of boundary stability and resource acquisition, subject to constraints imposed by its environment (resource scarcity) and by other entities’ boundaries. The formal specification and regularity conditions appear in [4]. The specification is naturally heterogeneous: each agent has its own objectives, constraints, patience, and capability, and the equilibrium results hold in the asymmetric case. Entities in this sense are necessarily open systems, and the framework makes no commitment about the global thermodynamics of the universe, only that the local environment of the entities under analysis admits the required throughput.

Terminological convention. Throughout the paper, *entity* denotes the thermodynamic object (an open system that actively maintains a boundary against entropy), *agent* denotes the game-theoretic object (a strategic decision-maker in a multi-player

game), and *system* is used in its general physical sense (any bounded collection of matter and energy under analysis), including but not limited to entities. Every agent is an entity; every entity is a system; not every system is an entity.

3.4 Value and Information

Value (singular, objective) is defined as the contribution of a resource to an entity's persistence, denominated in energy units (Joules). **Values** (plural, subjective) are the named, communicable moral abstractions (honesty, fairness, compassion) that different societies have developed as approximations of how to operate within the cooperative equilibrium (morals, in the terminology of the central claim, §5.2).

Not all energy is equal, and Value cannot be captured by any single physical measure. A star contains vastly more raw energy than a forest, but raw energy alone does not capture Value: the forest embodies structural complexity (genetic diversity, ecological interdependence, evolutionary search) that cannot be reconstructed from raw energy input on any physically relevant timescale. This is formalized through three quantities from *Thermodynamics of Cooperation*.

Information density $\rho_{\text{info}} = \mathcal{I}/m$ [6] is the functional information a system embodies per unit mass.

Accumulated negentropy $\mathcal{N}(T) = \int_0^T \dot{W}_{\text{order}}(t) dt$ [6] is the total ordering work invested in a system over its history.

Present value of generative information $PV_{\text{gen}} = \dot{\mathcal{I}}_{\text{gen}} \cdot v/r$ [6] is the discounted future stream of novel functional information a generative system continuously produces.

3.5 Scarcity and Collision

Consider a system of $N \geq 2$ agents sharing an environment with n resource dimensions and finite endowment $\mathbf{R} = (R_1, \dots, R_n) \in \mathbb{R}_{\geq 0}^n$. Each agent selects a strategy vector $\mathbf{x}_i \in \mathbb{R}_{\geq 0}^n$, and the joint strategy must satisfy resource conservation:

$$\sum_{i=1}^N x_{ij} \leq R_j \quad \forall j \in \{1, \dots, n\}$$

A resource dimension j is in *collision* if the sum of all agents' unconstrained optima exceeds supply: $\sum_i x_{ij}^o > R_j$ [4]. By the Inevitability of Collision lemma [4], in any system of $N \geq 2$ agents with overlapping resource needs and finite endowments, at least one resource dimension is in collision for sufficiently large N .

Collision is not an edge case. For any population of non-trivial size in a finite environment, collision is the systemic default.

Remark 1 (Scope of "Resource"). "Resource" in this framework is not restricted to material commodities. The resource dimensions j encompass any physical state, channel, or configuration that contributes to an entity's persistence, including shared information channels (whose finite capacity is degraded by deception; see the defection costs, §4.2), accumulated knowledge and cultural capital (whose destruction is thermodynamically irreversible; see the accumulated negentropy discussion, §4.4), and boundary integrity itself (the energetic investment that constitutes an entity's structural

configuration). When a government suppresses speech, it degrades a shared information channel; when a community destroys a library, it erases accumulated negentropy. Both are resource collisions in the framework’s sense, subject to the same constraint mathematics as competition over territory or food.

4 Results: The Deductive Chain

The results below are drawn from *Thermodynamics of Cooperation*. Proofs appear in the cited papers.

4.1 Constraints and Rights

Given the scarcity framework (§3.5), each agent solves a constrained optimization problem: maximize its own persistence subject to finite resources and other agents’ boundaries.

The formal starting point is the *boundary constraint*: a restriction on what one agent’s strategy can do to another agent’s boundary [4]. A boundary constraint is a bidirectional relation: it restricts the acting agent’s strategy space while preserving the affected agent’s boundary integrity. What is conventionally termed a right is the protective aspect of this structural relation. No agent possesses intrinsic rights; each has exactly the freedom dictated by the cost structure of its surrounding population. This follows from the cost structure itself: the costs of defection (§4.2) make violation of boundary constraints self-penalizing, so constraint-respecting behavior emerges as the equilibrium strategy and produces the mutual protection we label “rights.” The formal derivation appears in [4].

Theorem 1 (Shadow Price Theorem). *Every boundary constraint carries a computable shadow price: the marginal energy that an agent foregoes by respecting another agent’s boundary. The shadow price is zero when uncontested and becomes positive only where agents’ claims collide. [4]*

Theorem 2 (Impossibility of Infinite Freedom). *Let $N \geq 2$ agents share a finite resource endowment. No feasible strategy profile can simultaneously be unconstrained-optimal for all agents. [4]*

Corollary 3 (Necessity of Active Constraints). *In any multi-agent system with resource collision, at least one boundary constraint must be active ($\mu_{Bk}^* > 0$) at any feasible solution. Boundary constraints are not optional; they are structurally necessary for coexistence. [4]*

Unlimited freedom for all agents is incompatible with shared finite resources. Removing all constraints under resource collision produces no stable allocation [4]. Hobbes [37] identified this catastrophic outcome through philosophical argument; the proposition establishes it as a formal consequence of finite resources and competing objectives.

Theorem 4 (Variational Equilibrium). *A unique equilibrium with shared shadow prices exists. The optimization contains no term that distinguishes one agent from another: with symmetric agents, the equilibrium converges to equal allocation. [4]*

The Variational Equilibrium provides a formal basis for impartiality. It is a structural analogue of Rawls’s veil of ignorance [40], not as a thought experiment but as a symmetry property of the constrained optimization.

4.2 Costs of Defection

Constraints are necessary (§4.1); they are also self-enforcing. *Thermodynamic Friction* [7] and *Strategic Entropy Injection* [5] prove that defection is self-penalizing through two independent mechanisms: thermodynamic friction and information-theoretic degradation.

4.2.1 Thermodynamic Friction

Constraint violation dissipates energy [7]. The friction function Φ quantifies the total dissipation per unit of contested energy.

Theorem 5 (Net-Negative Conflict). *For $\Phi > N/(N - 1)$, adversarial contest is system-net-negative: the cumulative energy dissipated exceeds the value of the contested resource.* [7]

Theorem 6 (Cascading Friction). *A single constraint violation cascades across the cooperative network at amplification ratios of 10^3 – 10^4 .* [7]

Even small transgressions are enormously costly in network terms.

4.2.2 Deception as Entropy Injection

The friction analysis extends to information-theoretic violations. Deception is formalized as the deliberate injection of entropy into a communication channel [5].

Theorem 7 (Decision Cost of Deception). *Deception imposes a quadratically scaling cost on receivers.* [5]

Theorem 8 (Systemic Deception / Network Collapse). *At a critical deception threshold $q^* \approx 0.063$ for depth $d = 10$, the communication network collapses.* [5]

Corollary 9 (Honesty Efficiency Principle). *Perfect honesty ($q = 0$) is the unique efficiency-maximizing communication regime.* [5]

Small, persistent lies are catastrophic in deep organizations: a ten-layer decision pipeline cannot tolerate more than approximately 6.3% per-layer deception before quality collapses entirely.

4.3 Cooperation as the Unique Efficient Equilibrium

The payoff matrix is constructed not from arbitrary utility values but from the physical costs derived above (§4.2). Two agents contest a resource (in the broad sense of the scarcity framework, §3.5) of energy value $E_R > 0$, each choosing to Cooperate (C : accept the constrained allocation) or Defect (D : violate the constraint, triggering friction and deception costs). Under baseline parameters ($E_R = 100$, $\Phi = 2.2$, and standard exploitation parameters from [8]), the payoffs (in Joules) are:

	$B : C$	$B : D$
$A : C$	(48, 48)	(−11.25, 78.25)
$A : D$	(78.25, −11.25)	(−5, −5)

Mutual defection is net-negative ($P = -5 < 0$). Both agents *lose* energy; the conflict destroys more value than the resource provides. In the abstract Prisoner’s Dilemma, $P > 0$ and mutual defection is merely suboptimal. Under physical constraints, mutual defection is value-destroying, a thermodynamic form of the condition Hobbes [37] described on different grounds.

Proposition 10 (Physical Prisoner’s Dilemma). *Under the energy-based payoff model with physical constraints, the single-round game satisfies $T > R > P > S$ with $P < 0$, where T is the temptation to defect, R the reward for mutual cooperation, P the penalty for mutual defection, and S the (suckers) payoff for cooperating against a defector. Defection is the unique Nash Equilibrium of the one-shot game, even though mutual cooperation is Pareto-superior. [8]*

The one-shot result is a trap. Each agent’s best-response calculation yields defection regardless of the opponent’s choice ($T > R$ if the other cooperates; $P > S$ if the other defects). Both agents follow this logic, and both arrive at (D, D) with payoff $P = -5$: every participant loses energy in absolute terms. Neither agent can unilaterally escape, since switching to C while the opponent defects yields the sucker’s payoff $S = -11.25$. Resolution requires the repeated interaction structure that follows.

In the repeated game, agents interact over indefinite horizons. The discount factor δ weights future payoffs relative to present ones. Unlike the abstract “patience parameter” of standard game theory, the framework’s discount factor is anchored to physically observable quantities: survival probability and time-preference rate [8].

Theorem 11 (Cooperation as Nash Equilibrium). *In the infinitely repeated energy-based game, cooperation is sustainable as a Nash Equilibrium if and only if the discount factor exceeds a critical threshold: $\delta \geq \delta^* = 0.363$. [8]*

A discount factor above 0.363 is sufficient to sustain cooperation. The threshold is low because mutual defection is self-destructive ($P < 0$), making the cost of short-horizon behavior large relative to any single-round gain. No assumption of altruism or far-sightedness is required; any agent whose time-horizon satisfies $\delta \geq \delta^*$ cooperates at equilibrium. Moreover, 0.363 is the baseline under idealized pairwise conditions. Incorporating network friction effects, multi-agent dynamics, and stochastic environmental costs each introduces additional penalties for defection, lowering the threshold further [8].

Corollary 12 (Cooperation as the Unique Efficient Nash Equilibrium). *Cooperation is the unique Nash Equilibrium simultaneously satisfying Pareto efficiency and welfare maximization. [8]*

This resolves the Folk Theorem’s indeterminacy [26]. When payoffs are denominated in energy with thermodynamic friction, cooperation is not merely one equilibrium among many; it is the only stable, efficient one. Physics selects the equilibrium that abstract game theory cannot. The result scales: in the N -player public goods formulation, cooperation remains a Nash Equilibrium as the number of agents grows without bound, with the critical threshold converging to $\delta_N^* \rightarrow 0.4$ as $N \rightarrow \infty$ under the baseline parameters of [8]. Ostrom [33] documented empirically that large-group cooperation persists without central authority; the N -player theorem provides the formal proof that this observation is not anomalous but expected.

Additional results reinforce the stability of the cooperative equilibrium:

Theorem 13 (Invasion Barrier). *A single defector earns less than cooperators for $\delta > \delta^*$. [8]*

Theorem 14 (Defection Profit Erasure). *At sufficient network friction, even a single defection’s short-term profit is completely erased within approximately 20 recovery periods. [8]*

Defection in the physical game is not merely suboptimal but strictly loss-making at every scale: mutual defection yields negative absolute returns for both agents ($P = -5 < 0$), a single defector in a cooperative population earns less than cooperators, and even a one-time defection followed by return to cooperation is net-negative over the agent’s lifetime. Unlike the abstract Prisoner’s Dilemma, where defection at least yields a positive payoff, defection here has no sustainable form.

4.4 The Irreplaceability of Accumulated Negentropy

The energy recoverable from a system is a fraction of the thermodynamic work required to build it. The accumulated negentropy formalism of [6] quantifies the irreplaceable value of complex systems.

Theorem 15 (Blind Search Replication Cost). *The cost of rebuilding a complex system from scratch is at least as great as the original thermodynamic investment: $W_{\text{rebuild}} \geq \mathcal{N}(T)$. [6]*

Corollary 16 (Burning-Library Ratio). *For any complex system, the fraction of accumulated thermodynamic investment recoverable through destruction, $\mathcal{R}_{\text{BL}} = E_{\text{destroy}}/\mathcal{N}$, is negligibly small. Destruction recovers a vanishing fraction of the value destroyed. [6]*

Burning the Library of Alexandria for fuel would have recovered approximately one thousandth ($\mathcal{R}_{\text{BL}} \sim 10^{-3}$) of the accumulated intellectual work invested in producing its contents: centuries of scholarship converted to minutes of heat. For biological systems, the ratio is far more extreme. *Accumulated Negentropy* [6] computes $\mathcal{R}_{\text{BL}} \sim 3 \times 10^{-7}$ for the Amazon rainforest and $\sim 10^{-7}$ for the biosphere as a whole: destruction recovers less than one ten-millionth of the thermodynamic investment that produced the system’s complexity.

Theorem 17 (Present Value of Generative Information). *Living systems continuously produce novel functional information whose present value is $PV_{\text{gen}} = \dot{\mathcal{I}}_{\text{gen}} \cdot v/r$. Destruction permanently terminates an irreplaceable data stream. [6]*

These results, taken together, ground a distinction that moral and legal traditions have long marked but not derived from physical first principles: the asymmetry between life and property. The Library of Alexandria, the greatest knowledge repository of the ancient world, had a recovery ratio of $\sim 10^{-3}$; a single human being, embodying decades of developmental and evolutionary search, contains orders of magnitude more accumulated negentropy than any artifact. For living systems, the ratio drops to $\sim 10^{-7}$, destruction is effectively total and generative information is permanently lost. These ratios are lower bounds on the cost of destruction: they measure only the backward-looking thermodynamic investment (work invested versus energy recoverable). Moreover, the search process that organized raw materials and energy into functional structure dwarfs the original material investment and cannot be shortcut: the Blind Search Replication Cost theorem establishes that rebuilding requires at least the full accumulated negentropy regardless of available energy. Destruction also eliminates a node from the cooperative network, elevating baseline friction for all remaining agents. Destruction is therefore self-defeating: it erases a structure whose replacement cost exceeds \mathcal{R}_{BL} by further orders of magnitude.

Rights theory and the asymmetry between life and property are treated in full at §7.2.

4.5 Stable Coexistence as a Dynamical Attractor

Value Dynamics [9] extends the static equilibrium analysis into a dynamical systems framework. Consider an agent interacting with a center of accumulated value (a resource-rich ecosystem, a nation, an institution, an economy). The agent faces a tradeoff: closer engagement yields greater energy harvest but also greater friction costs. Too close, and maintenance costs consume the agent; too far, and it cannot harvest enough to survive. The formal analysis ([9]) proves that this tradeoff produces a unique coexistence attractor: an optimal engagement distance at which net energy is maximized.

Theorem 18 (Stability of the Coexistence Attractor). *The coexistence attractor is globally stable. Agents displaced in either direction experience restoring forces that return them to the equilibrium.* [9]

Theorem 19 (Freedom Bandwidth). *Around the coexistence attractor, there exists a Coexistence Band: a finite range of positions at which an agent can sustain its identity. The width of this band provides a formal, scalar measure of freedom, determined by physical parameters.* [9]

The Coexistence Band corresponds to what political and social theory identifies as the space of viable freedom: too little engagement (isolation from markets, institutions, or ecosystems) starves the agent of resources; too much (total dependence, loss of autonomy) subjects it to friction costs that erode its boundaries.

Theorem 20 (Irreversibility of Dissolution). *An agent pushed below the boundary dissolution threshold experiences strictly declining boundary integrity, reaching zero in finite time. Crossing the dissolution threshold is irreversible.* [9]

Dissolution is the formal analogue of the poverty trap, the failed state, and ecological collapse in human and biological systems: a point beyond which recovery requires external intervention. The irreversibility is what gives the cooperative equilibrium its structural significance; the consequences of crossing this threshold are permanent.

Corollary 21 (Cascade Collapse). *If a high-energy center is destroyed, all agents within its Coexistence Band simultaneously lose viability.* [9]

Excessive concentration in a single center therefore makes the entire ecosystem fragile. The multi-center diversification result [9] proves that distributed systems with multiple coexistence attractors are structurally more resilient than monocentric ones: the formal basis for polycentric governance [33; 34].

4.5.1 The Stability-Cooperation Feedback

The dynamical analysis connects to the game-theoretic analysis through the discount factor. The discount factor is maximized at the coexistence attractor [9]: agents at the equilibrium distance are precisely those whose time-horizons satisfy $\delta \geq \delta^*$, producing a positive feedback between position and strategy. Stability produces cooperation, which reinforces stability. The coexistence attractor is not a fragile fixed point requiring external maintenance; it is a self-reinforcing attractor basin.

5 From Cooperative Equilibrium to Ethics

5.1 The Challenge

In 1739, David Hume observed that authors imperceptibly shift from descriptive claims (*is*) to prescriptive claims (*ought*) without valid inference connecting the two domains [1, 3.1.1]. This “Guillotine” has since been treated as an impassable logical barrier between descriptive and prescriptive. Moore [41] reinforced the barrier with the Open Question Argument: for any natural property N , the question “is N good?” remains open, so “good” cannot be defined as any natural property.

Any framework that derives prescriptive conclusions from descriptive premises must address this challenge directly.

5.2 From “Optimal Strategy” to “Ethics”: The Central Claim

The proof that cooperation is the unique efficient equilibrium is a result about strategic behavior. Connecting it to *ethics*, the moral norms, emotions, and judgments observed across human cultures, requires justification. Proving that heat flows from hot to cold does not define temperature, but it explains why temperature behaves the way it does. Proving that cooperation is the unique efficient equilibrium does not by itself define good. What the derivation explains is why cooperative norms exist, why they converge across cultures, and why behavior is perceived as right or wrong according to its proximity to the equilibrium.

The argument rests on four steps.

Step 1: The optimization problem is real but intractable. A persistent agent in a shared, finite-resource environment faces a constrained optimization problem whose exact solution requires computing shadow prices (§4.1), friction costs (§4.2.1), deception costs (§4.2.2), cooperative equilibrium strategies (§4.3), sustainability bounds (§4.4), and coexistence dynamics (§4.5). Both the information and the computation are out of reach: no biological agent observes the full system state, and none could perform the calculations in real time even if it did.

Step 2: Evolution produces moral emotions. Natural selection is a noisy, distributed, multi-generational optimizer. Populations subject to the pressures formalized in *Thermodynamics of Cooperation* evolve pre-linguistic, somatic approximations of the cooperative equilibrium [53; 54; 55; 56]. These are emotions, not thoughts: empathy approximates boundary extension (§6.2); moral disgust approximates the rejection of entropy-injecting deception (Corollary 9); contentment⁵ approximates operation within sustainability bounds (Theorem 15); guilt and shame approximate the self-monitoring that prevents cascade-inducing violations (Theorem 6); gratitude approximates the recognition of cooperative benefit (Theorem 11); and indignation approximates the enforcement response to constraint violation within the Coexistence Band (Theorem 19). The biological mechanism is well-established: somatic markers, bodily signals associated with past outcomes, bias decisions toward advantageous strategies without requiring conscious deliberation [57; 58; 59].

⁵Contentment is not a moral emotion in the standard classification [45] but an equilibrium-tracking one: it signals that the agent’s internal state is within viable operating range. The moral emotions listed here (empathy, guilt, shame, gratitude, indignation, moral disgust) are directional signals, triggered when interactions move toward or away from the cooperative equilibrium.

Step 3: Emotions are compressed into morals. The raw emotions are named and generalized. Guilt in response to deception is abstracted as *honesty*; indignation at cheating as *fairness*; empathy for others in need as *compassion*. Morals are the first abstraction: communicable, teachable, transmissible across generations. They are products of gene-culture co-evolution [60; 61], where “outside-the-head” cultural products (institutions, practices, technologies) interlock with evolved “inside-the-head” psychological mechanisms to suppress free-riding [46] and enforce the cooperative equilibrium. *Good* and *bad* are the broadest compressions, marking proximity to or distance from the cooperative equilibrium. *Ought* operates at this level: the prescription “you ought to be honest” expresses in communicable form the emotional pull toward the equilibrium. Each moral tracks a specific component of the formal structure: honesty tracks the entropy-minimizing communication regime (Corollary 9), fairness tracks shadow-price equality (Theorem 4), reciprocity tracks the cooperative equilibrium strategy (Theorem 11), reverence for life tracks the thermodynamic irreplaceability of accumulated negentropy (Corollary 16), and tolerance tracks the Coexistence Band within which autonomous agents sustain mutual viability (Theorem 19).

Step 4: Morals are systematized into ethics. What we call “ethical principles” are systematic frameworks extracted from morals [40; 62]. Deontology systematized the constraint-shaped morals (honesty, duty, integrity) into universal rules, capturing the constraint structure (§4.1). Consequentialism systematized the optimization-shaped morals (welfare, outcomes) into maximization frameworks, capturing the optimization structure (Corollary 12). Contractarianism systematized the voluntary-constraint morals (fairness, reciprocity) into social contract theory, capturing the voluntary-constraint structure (Theorem 4). Virtue ethics systematized the constitutive-standards morals (excellence, flourishing, natural goodness) into accounts of character (Aristotle’s *eudaimonia*; Foot’s natural goodness), capturing the identity-maintenance structure (Theorem 18). Each tradition identified a real feature of the underlying formal structure but lacked the tools to exhibit it in full.

Evolution produced emotions as somatic approximations of the cooperative equilibrium, those emotions were compressed into communicable morals, and morals were systematized into the philosophical traditions we call ethics.

Let \mathcal{S} be a population of agents satisfying A_1 (Intent to Persist) in a shared, finite-resource environment governed by A_0 (Second Law). *Thermodynamics of Cooperation* establishes two results:

- (i) Cooperation is the unique efficient Nash Equilibrium (Corollary 12).⁶
- (ii) Any population subject to selection pressure in such an environment converges on the cooperative equilibrium, because cooperation is evolutionarily stable (Theorem 13), defection is self-extinguishing [8], and the cooperative configuration is a dynamical attractor with restoring forces (Theorem 18).
- (iii) The moral emotions evolved under (ii), the cultural morals abstracted from those emotions, and the ethical systems systematized from those morals by interlocking institutions and practices, correspond to specific components of the formal structure: constraint-respecting behavior and fairness [4], deception-avoidance and honesty [5], preservation of complex systems

⁶The cooperative equilibrium is derived from constraint necessity and computable shadow prices [4], self-penalizing deception costs [5], the irreplaceability of accumulated negentropy [6], cascading thermodynamic friction [7], and the game-theoretic equilibrium result itself [8]. The payoff matrix that produces the unique efficient equilibrium is constructed from these physical cost structures, not from arbitrary utility assignments.

[6], friction-avoiding behavior [7], reciprocity and conditional cooperation [8], and tolerance of diversity within the Coexistence Band [9].

Parts (i) and (ii) are proven mathematical results. Part (iii) is an empirical prediction: given (i) and (ii), the existence of evolved emotional approximations of the cooperative equilibrium is a prediction of standard evolutionary biology: if a unique efficient equilibrium exists and defection is self-extinguishing under selection pressure, populations will evolve emotions that approximate the equilibrium’s component structures [53; 54]. The specific correspondence asserted in (iii), between particular evolved emotions, the morals abstracted from them, and particular formal results, is a testable empirical prediction: the structure of moral emotions (which behaviors trigger guilt, gratitude, or indignation, and with what relative intensity) should track the structure of the cooperative equilibrium (which strategies are optimal, which are dominated, and by what margin). Specific falsification criteria appear in §8.2.

The claim is not that evolution optimized ethics perfectly, but that the *target* of the evolutionary process is now analytically identifiable. Moral emotions are noisy, culturally filtered, computationally bounded approximations of the cooperative equilibrium, the same way that thrown-ball intuitions are noisy approximations of Newtonian mechanics. The trajectory runs from evolved emotion to communicable moral to systematic ethics, each a lossy compression of the cooperative equilibrium.

The framework explains the *structure* of ethics; it says nothing about consciousness, qualia, or what it feels like to be moral. It says: the thing that is felt has the structure of a constrained optimization problem, and the solution to that problem is cooperation. What appear to be irreconcilable clashes of fundamental values are disagreements between competing abstractions of heuristic approximations, each tracking a different component of the same cooperative equilibrium.

Both *good* and *ought* can be defined in terms of the framework⁷:

Good is a scalar quantity of proximity to the cooperative equilibrium. It is gradable, measurable, and context-dependent, not a substance but a property.

Ought is a vector in the direction of steepest ascent from the current position to the equilibrium, encoding both what to change and how much benefit the change would bring.

Moore asked what *good* is and concluded it was undefinable. Moore’s Open Question dissolves because it assumes the “is” of identity, not of predication. X is not identical to good, but X may be good to the degree that it approximates the cooperative equilibrium. The question “is X good?” is no more open than “is this surface hot?”. Hume asked how to derive *ought* from *is* and concluded no bridge exists. *Ought* is the gradient toward the cooperative equilibrium at the current position, and therefore *ought is* descriptive. The gap Hume identified was never between two truths but between two resolutions of the same truth, a difference in resolution, not in kind.

The claim concerns the structural target that moral vocabularies approximate, not the vocabularies themselves or the specific normative content that *right*, *wrong*, and *ought* express within a given culture. The action identified as *correct* in a given circumstance can vary substantially across cultural contexts, because each culture presents a different realization of the cooperative equilibrium shaped by its history, institutions, and material conditions. In each context the local gradient points toward the optimum of that network; variation in moral terminology across cultures is consistent

⁷The formal identities for *good* and *ought* are derived in §B.

with, and predicted by, convergence of the underlying structure those terms track [46; 45]. What would falsify the claim is not variation in vocabulary but the absence of that structural convergence.

5.3 Anticipated Objections

5.3.1 “ A_1 smuggles in a normative premise”

A_1 (Intent to Persist) is a physically observable property of any system whose structural organization is directed toward boundary maintenance. Bacteria maintain ion gradients, ecosystems cycle nutrients, corporations acquire revenue, organisms metabolize. No conscious act is required; persistence is a physical process that operates at every scale, including scales where consciousness does not exist. Some entities are additionally conscious of their persistence, but consciousness is not a prerequisite, in the same way that breathing operates automatically whether or not attention is directed toward it. The empirical evidence for persistence as a physical property is as extensive as for any claim in biology or physics. To reject A_1 as an empirical observation requires accounting for this evidence rather than dismissing it.

The objection can be pressed further: does the capacity to *deliberate about* A_1 presuppose A_1 ? It does. The argument has two branches, and A_1 withstands both. If cognition is a physical process (neural activity, metabolic expenditure, boundary-maintaining computation), then evaluating A_1 does not itself need to be a persistence-maintaining act, but it can only occur in a system where persistence-maintaining processes are active. Whether the evaluator is consciously willing persistence, unconsciously maintaining it, or consciously choosing not to self-destruct, the physical substrate that supports the deliberation satisfies A_1 throughout. If, alternatively, one posits a non-physical source of cognition, not identifiable or measurable, that claim is not falsifiable and cannot serve as grounds for rejecting a physically grounded axiom within a scientific framework. The claim that cognition is physical, by contrast, is falsifiable: one could in principle produce evidence against it, and none has been produced. On either branch, the circularity objection fails.⁸

5.3.2 “Evolution does not justify ethics”

The theory does not argue that evolved traits are justified *because* they evolved. The argument is that a derivable optimum exists (the cooperative equilibrium), that evolution produced approximations of it (emotions, Step 2), that those emotions were compressed into communicable morals (Step 3), and that ethical systems systematized those morals into philosophical frameworks (Step 4). Moral emotions, morals, and the philosophical traditions built from them are products of evolutionary selection in environments where the cooperative equilibrium is the attractor. The justificatory work is done by the equilibrium, not by the evolutionary history.

5.4 Philosophical Precedent

Multiple philosophical traditions have independently concluded that the physical structure of a system determines evaluative content without requiring a separate normative domain.

Biological teleology [2; 63]. Ruth Millikan’s theory of “proper functions” defines the function of a biological trait as the effect for which it was selected by its evolutionary history. The *proper*

⁸This has the form of a transcendental argument: the conditions required to contest A_1 (maintaining a boundary, processing information, persisting through time) are the conditions A_1 describes. The structure is shared with Descartes’ Cogito, but grounded in physics rather than consciousness.

function of a self-maintaining system is persistence: this is not a value judgment but a biological fact about the system’s design history. Axiom A_1 generalizes this: “Intent to Persist” is not a conscious desire but the defining characteristic of any system whose structural organization is directed toward boundary maintenance. Daniel Dennett’s *The Intentional Stance* [3; 63] provides a complementary formulation: we can legitimately attribute goal-directedness to a system when doing so yields successful predictions, without claiming the system has conscious awareness.

Constitutive standards of life-forms [64; 65]. Philippa Foot’s *Natural Goodness* develops a neo-Aristotelian ethical naturalism in which evaluative claims are *internal* to a life-form. Functional legs are good *for a cheetah*; deep roots are good *for an oak tree*; cooperation is good *for entities that persist in shared environments*. The evaluative “ought” is constitutive of the life-form, not imposed from outside by a separate normative domain. The theory formalizes this intuition: given what an entity is (a self-maintaining, energy-consuming system in a shared finite environment), cooperation is the formally derivable well-functioning of that system. The neo-Aristotelian approach is active in contemporary philosophy [66] and provides philosophical grounding for the central claim, independent of the derivations in *Thermodynamics of Cooperation*.

Searle’s institutional facts [67]. John Searle demonstrated that certain concepts *constitutively* contain normative content. The concept “promise” analytically necessitates “obligation to keep it”: the “is” of a promise already contains the “ought” of fulfillment. The theory exhibits a parallel structure: the concept “entity that persists” constitutively encompasses “system that must maintain boundaries, acquire energy, and navigate resource conflicts.” The connection from the “is” of a persisting entity to the “ought” of cooperation is indirect, passing through evolutionary and cultural steps. Both connections are physical, not conventional, and constitutive in both cases.

6 Stress Tests

6.1 The Free Will Problem

The challenge. If ethical behavior is derivable from thermodynamic constraints, it appears pre-determined by physics, leaving no room for moral agency. If the correct action is computationally determinable from boundary conditions, energy budgets, and discount factors, then agents do not choose to be ethical; their state is determined by state of the system in which they exist.

The resolution. Agency is the capacity of a system to model alternative strategy spaces and select among them, producing actions whose origin lies in the system’s own computation and that remain opaque to external prediction. This definition relies on a recursive relationship: the system’s state is the aggregate of its agents’ states, and each agent selects its next state by modeling the external state in relation to its internal state. No external observer can model this process more efficiently than the agent models itself; the agent therefore remains the authoritative source of its own state transitions.

The choice is the actual execution of this computation. The fact that an answer is computationally determinable does not mean it is determined until the agent actually performs the computation.

Three independent factors ensure that this internal process remains opaque to external prediction. First, identifying the cooperative optimum for a realistic population is a non-trivial optimization problem; computing a Nash equilibrium in a finite game is PPAD-complete in general [68], and the real-world optimization involves open boundaries, high dimensionality, and coupling that evolves with every action. Furthermore, the neural dynamics implementing the agent’s approximate solu-

tion are driven by stochastic processes, such as ion-channel noise and thermal fluctuations, amplified by the nonlinear dynamics of the neural circuitry [69]. This noise floor ensures that the mapping from prior state to action is not deterministically predictable. Finally, the approximations used to navigate these spaces are species-typical priors shaped by selection and tuned by individual history [70; 71], instantiated in the same stochastic circuitry.

Together, these factors ensure the computation cannot be evaluated from outside the agent. The agent is the computation and the computer.

6.2 The Altruism Paradox

The challenge. If the theory’s foundational axiom is self-preservation (A_1), then self-sacrifice appears irrational, a direct violation of the axiom. Yet altruism is empirically ubiquitous.

The resolution. The paradox resolves once “self” is identified with the persisting boundary rather than with the biological individual. The entity satisfying A_1 is the Markov blanket [51; 72] whose persistence the system’s organization is directed toward maintaining, and a blanket can enclose more than one biological organism (a Markov blanket of Markov blankets, in the terminology of [72]) and can outlast any particular metabolic vehicle. The effective payoff function carries nonzero weights w_{ij} on other agents’ welfare whenever their boundary is absorbed into the same persisting blanket, so what appears to be sacrifice from outside is resource reallocation within the blanket.

Boundary extension has two axes: spatial and temporal.

Spatial extension enlarges the blanket to enclose other agents. The physical basis is thermodynamic coupling: when agents pool resources, share boundary-maintenance costs, or co-invest in shared infrastructure, they form a higher-order blanket whose persistence becomes the operative optimization target. The “self-sacrifice” of one component for another is internal resource reallocation within that blanket, analogous to the body diverting blood from extremities to protect the brain during hypothermia. At the cellular level, the same logic governs apoptosis: a cell self-terminates to prevent corrupted information from propagating through the organism [73], altruism without cognitive mediation.

The coupling arises along a continuous spectrum through distinct biological mechanisms. At one end, genetic relatedness produces inclusive fitness: Hamilton’s Rule [28] predicts that an agent sacrifices personal payoff when the benefit to a related agent, weighted by the degree of relatedness, exceeds the cost. Neural mechanisms for action understanding and behavioral coordination [74; 75] extend the same coupling beyond kin, enabling cooperative blankets among unrelated individuals. Non-human primates exhibit inequity aversion [76] and prefer equitable outcomes in ultimatum-game paradigms [77; 78], confirming that these coupling mechanisms predate language and philosophical reasoning. At the far end of the spectrum, agents merge boundaries entirely (parent and child, spouses, military units), becoming a single composite entity with a joint welfare function. The mathematical structure is identical across the spectrum: genetic relatedness, neural coupling, and full structural merger all produce nonzero w_{ij} in the effective payoff function (§A), differing only in the biological mechanism that generates the coupling and its magnitude.

Temporal extension is not an independent mechanism but a consequence of spatial extension: an individual’s accumulated negentropy persists after its own blanket dissipates only if it is encoded in a larger blanket whose own persistence carries it forward. Biological reproduction, cultural transmission, and institutional embedding are mechanisms by which the individual’s functional information is absorbed into such a host blanket (a genetic lineage, a body of knowledge, an

institution) before the metabolic vehicle dissipates (§4.4). An agent whose accumulated negentropy is substantially encoded in the host blanket has an effective planning horizon that extends beyond its individual lifespan, producing a discount factor that approaches unity. Under Theorem 11, any discount factor above the critical threshold sustains cooperation. Self-sacrifice of the metabolic vehicle is then the optimal strategy for maximizing the persistence probability of the host blanket.

In human agents, this embedding plausibly underlies the belief that one’s identity persists beyond biological death. This belief is the heuristic approximation, not the load-bearing premise; the physical reality is that the agent’s accumulated negentropy does persist through the structures encoding its functional information.

Altruism and self-preservation are the same operation at different boundary scales. A_1 is satisfied by whatever Markov blanket persists, which need not coincide with the biological organism. The empirical observation that cooperative groups outcompete selfish groups [79; 80] is a direct consequence: cooperation minimizes network friction, and the group boundary becomes the operative unit of selection. The formal game-theoretic treatment appears in §A.

6.3 Thermodynamic Fascism

The charge. A physics-based ethics, with value denominated in energy and information, appears to rank agents by throughput: the agent commanding more energetic and informational resources is worth more, and is therefore entitled to subjugate or displace the lesser. This objection assumes that energetic magnitude confers normative priority, a category error that mirrors the structural failures of Social Darwinism [81; 82].

The refutation. The charge conflates raw power with the formal definition of Value. Once the components of Value are carried explicitly, domination is an unstable strategy: the dominator depends on the dominated, and the equilibrium that maximizes its payoff is the cooperative one.

Value is not throughput. Raw energetic utility is the smallest component of an entity’s contribution to the network; the dominant terms are accumulated negentropy (the total thermodynamic work already invested in the system’s information structure, §4.4) and generative utility (the present value of the novel functional information the system will continue to produce, Theorem 17). The dominated agent’s Value is a dependency of the dominator’s own: every agent’s coexistence band is sustained by the functional and generative contributions of the agents it is coupled to (Theorem 18). Domination is a sustained boundary-constraint violation; it does not sever the coupling, it degrades it. The dominator continues to depend on the dominated’s generative contribution while driving that contribution down through friction (Theorem 6), shadow-price accumulation (Theorem 1), and, in the deceptive variant, channel degradation (Theorem 7).

These are not independent costs layered on top of the dependency; they are the mechanisms through which the dependency is paid. Because every agent’s payoff is coupled to the Value of its neighbors, any unilateral deviation by the strongest agent shrinks the total payoff and leaves it with a share of a smaller pie that is strictly less than its cooperative share (Theorem 13); sustained power concentration amplifies friction across the network and eventually destroys the dominator’s own coexistence band (Corollary 21, Theorem 20). The cooperative equilibrium (Corollary 12) is therefore not a constraint imposed on the dominator from outside; it is the configuration that maximizes the dominator’s own payoff, because that payoff is a function of the Value of the agents it is coupled to.

Social Darwinism optimized a proxy (reproductive fitness or current competitive advantage) under data that omitted accumulated investment and foreclosed future generation. The present derivation optimizes directly over the thermodynamic quantities, carries the accumulated and generative terms explicitly, and produces a structural penalty on concentration through cascade dynamics. Domination is not a high-value strategy in the resulting landscape; it is a strategy that erodes the substrate of its own payoff.⁹

6.4 The Self-Preservation Objection

The charge. Even granting that destruction is wasteful for resource extraction, an agent identifying another agent as an existential threat may rationally choose preemptive elimination for survival. A government considering preemptive war, a state contemplating suppression of a political rival, a community eradicating a species it deems inconvenient: each invokes A_1 against the theory’s own conclusions.

The refutation. The objection generalizes a limiting case. When a threat is immediate and certain to breach the agent’s boundary integrity irreversibly, whether by terminating persistence outright or by inflicting an irreversible violation, the effective δ collapses toward zero and the theory predicts destructive self-defense: the agent has no future to discount, and the cost calculus reduces to the single surviving round. A physical attack in progress is the canonical instance. The results below do not contest this case. They concern the class of scenarios that the objection specifically states, where the threat is anticipated rather than ongoing and the agent retains a non-trivial horizon.

In those scenarios, the cost calculus favors management over destruction. The destruction path incurs constraint violations on all affected agents, friction cascading at 10^3 – 10^4 amplification (Theorem 6), negentropy loss with $\mathcal{R}_{BL} \sim 10^{-7}$ recovery (Corollary 16), cascade collapse risk (Corollary 21), and coalition opposition from agents whose own persistence depends on the cooperative network [6]. The management path (containment, negotiation, structural modification of the interaction, boundary adjustment) incurs bounded finite cost while preserving the generative stream and the cooperative network. For any parameterization with $\delta > \delta^*$, management dominates destruction on the agent’s own terms: the cumulative costs of destruction exceed any achievable benefit.

The Yellowstone wolf extermination illustrates the cascade dynamics concretely [83; 84]. Removal of a single keystone species, driven by the logic of preemptive threat elimination, triggered system-wide degradation across a tri-trophic cascade; the 1995 reintroduction reversed the degradation, confirming the coexistence configuration as the system’s attractor (Theorem 18). Cooperative coexistence is the cost-minimizing strategy.

⁹The theorems establish the direction of costs and the eventual outcome under the stated conditions; they do not bound the timescale on which a given violation produces observable collapse, and the empirical record contains long-persisting dominations that have not triggered full system collapse within an observable horizon. The baseline model is deliberately coarse: violations are binary in both the game-theoretic and deception results, with no graded violation scale and no commitment or capitulation parameter governing when an agent withdraws from conflict. What the theorems guarantee is that the associated costs (friction, shadow prices, foreclosed contributions, cascade risk) accumulate monotonically toward collapse of the violator’s coexistence band; the observable timescale depends on friction amplification, network depth, and recovery horizons specific to the system. Extensions relaxing the binary-compliance assumption and introducing commitment dynamics are listed in §8.3.

7 Structural Implications and the Social Contract

The results of the deductive chain (§4), the central claim (§5.2), and the stress tests (§6) are substrate-independent: every result holds for any agent satisfying A_1 in any shared finite-resource environment. The cooperative equilibrium is the dynamical attractor for any such system with repeated interaction, and the proofs do not depend on any agent’s awareness of them. What varies is the path: the friction dissipated, the negentropy destroyed, and the time elapsed before the system arrives. This section describes the structure of that destination and the costs of various departures from it.

7.1 The Social Contract as Consequence

The social contract tradition [37; 40; 85] argues that rational agents would voluntarily accept mutual constraints to escape a destructive state of nature. Each formulation rests on a thought experiment whose conclusions depend on the assumptions built in: Hobbes’s agents are driven by fear, Rawls’s reason behind a veil of ignorance, Gauthier’s are constrained maximizers. The derivation replaces the thought experiment with a proof.

The first step is to establish the unconstrained baseline. The Impossibility of Infinite Freedom (§4.1) proves that in any environment where $N \geq 2$ agents share at least one essential resource whose supply is insufficient to satisfy all agents’ unconstrained optima simultaneously, no strategy profile can be unconstrained-optimal for all agents. Necessity of Active Constraints (§4.1) establishes that removing all constraints produces conflict: each agent’s pursuit of its unconstrained optimum generates negative externalities on every other agent. This result follows from the combinatorics of finite resources and competing objectives, and the solution (below) is self-enforcing rather than requiring Hobbes’s external sovereign.

Given that the unconstrained regime is infeasible, what does the constrained regime look like? The Shadow Price Theorem (§4.1) proves that each constraint carries a computable cost, the shadow price μ_{Bk}^* , representing the marginal energy that agent A foregoes by respecting agent B ’s allocation. Cooperation as the Unique Efficient Nash Equilibrium (§4.3) then establishes that the cooperative strategy profile is the *unique* Nash Equilibrium that simultaneously achieves Pareto efficiency and welfare maximization, derived not from a hypothetical assembly’s decision but from the optimization problem that finite resources and thermodynamic friction impose.

The final piece is stability. The Invasion Barrier (§4.3) proves that a single defector in a cooperative population earns strictly less than cooperators for $\delta > \delta^*$, and Defection Profit Erasure (§4.3) proves that even a single defection followed by return to cooperation incurs a net lifetime loss.

The derivation shows that agents persisting in a shared environment converge on mutual constraints, that the equilibrium those constraints enable is unique, and that it is dynamically stable. The social contract is the only stable solution to the optimization problem that finite resources and repeated interaction impose.

7.2 Rights from Constraint Boundaries

The philosophical literature offers several competing accounts of what grounds rights: natural endowments [38], social constructs built by legal institutions, outputs of rational deliberation [40]. The derivation provides a formally precise account: what we recognize as a right is a boundary constraint (§4.1); only the constraint exists in the optimization, and “right” is the label applied from the perspective of the agent the constraint protects.

This reframes the standard understanding of rights. The dominant philosophical framing treats rights as entitlements: “I have a right to X, therefore you must accommodate me.” But in the analytical framing, an agent begins with fundamental requirements (energy, boundary integrity) and no entitlements to any resource; its access depends entirely on the structure of the surrounding population. Without mutual constraints, there is no stable allocation (§4.1); the unconstrained state is Hobbes’s war, and it is net-negative for everyone ($P < 0$, §4.3). Constraints are not impositions on pre-existing freedoms; they are the structural features that *create* the feasible region. For agents in a shared finite-resource environment, freedom is not an inherent property but an emergent consequence of the mutual constraints that all agents impose on themselves (Freedom Bandwidth, §4.5).¹⁰ The cooperative equilibrium improves on the baseline of unguaranteed access, and is not a compromise from an imagined state of unlimited entitlement. The shadow price μ_{Bk}^* quantifies this directly: the marginal energy one agent foregoes by respecting another’s boundary, measured within the shared optimization rather than derived from any claim either agent holds against the other.

The distinction between negative and positive rights does not correspond to any asymmetry in the optimization’s cost structure. Political philosophy has debated whether the obligation not to harm (negative) and the obligation to provide (positive) have equal standing, on the assumption that they are fundamentally different kinds of claim. The Lagrangian constraints of §4.1 are formally constraints on agent strategies: each agent accepts a bounded allocation, forgoing its unconstrained optimum in exchange for the cooperative surplus. Viewed from the constrained agent’s side, this is a restriction. Viewed from the agent whose boundary is protected, the same constraint delivers a provision: the neighboring agent’s cooperative behavior preserves the resources it would otherwise consume. The cooperative strategy already specifies each agent’s allocation, including the fraction directed to shared infrastructure, so an agent that withholds its cooperative contribution has deviated from the equilibrium strategy, and the deviation triggers the same friction costs (§4.2) and the same shadow-price penalties (§4.1) as seizing another agent’s resources. There is no explicit primitive for provision obligations¹¹; every agent’s constraint is another agent’s provision, and the two are the same object described from opposite sides.

¹⁰Freedom Bandwidth is the *extrinsic* face of freedom: what mutual constraints leave open. The *intrinsic* face, what an agent is internally capable of doing, tracks accumulated negentropy $\mathcal{N}(T)$ (§4.4). The framework therefore contains both of what Sen [86] termed capabilities; their full interaction is deferred to §8.3.

¹¹*Provision obligations.* The endogenous production network extension (§8.3) is expected to yield a Necessity of Active Provisions result analogous to Necessity of Active Constraints (§4.1): for the cooperative equilibrium to persist, producers must supply the outputs that dependent agents require. The requirement is on the producer’s own optimization, not on any claim the consumer holds against it, and the extension introduces no consumer-side entitlement.

7.3 Economic Measurement and Externalities

The friction model (§4.2) translates what economists call externalities into computable terms. Conventional externality-correction mechanisms (taxes, cap-and-trade systems, regulatory mandates) rely on estimated social costs typically measured through willingness-to-pay surveys, hedonic pricing, or political negotiation. These estimates are vulnerable to the preference-elicitation problems that affect welfare economics [87]. The framework’s friction cost is derived from physical parameters rather than elicited preferences, and it includes network cascade effects (§4.2) that conventional cost-benefit analysis typically omits.

The 2007-2008 financial crisis illustrates the gap [88]. Pre-crisis risk models priced the exposure of individual institutions in isolation; aggregate statistics (GDP growth, unemployment, average leverage) suggested a healthy system. What these models missed was the network structure: correlated exposures, counterparty chains, and the fact that one institution’s distress propagated stress to every institution it touched. The cascading friction result (§4.2) formalizes exactly this structure. Friction generated at a single node propagates through the network, and the aggregate cost includes amplification terms that do not appear in any individual node’s balance sheet. An accounting framework built on friction costs would formalize the systemic danger that aggregate statistics masked, not by predicting the specific trigger, but by registering the network topology that was generating fragility faster than individual-node metrics could detect.

Several measurable quantities serve as indicators of cooperative health. Total friction dissipated tracks how efficiently a system cooperates. The fraction of agents operating within the Coexistence Band (§4.5) signals whether the economy is generating systemic instability. The rate of change of accumulated negentropy measures whether a system’s informational capital is growing or declining. And the degree to which resources concentrate beyond the critical mass threshold \mathcal{M}_{\min} (§4.5) identifies a structural danger invisible at the node level.

7.4 Stability Thresholds and Their Consequences

The theory does not prescribe policies; it describes the cost structure against which any policy can be evaluated. Legal and regulatory systems are institutional mechanisms that reduce variance around the cooperative equilibrium. Individual agents approximate the equilibrium’s cost structure through moral emotions (§5.2, Step 2), but these heuristics are noisy: the distribution across agents is centered on the equilibrium but has significant spread. Laws, regulations, and enforcement institutions act as filters on this distribution, dampening the tails and narrowing the range of realized behavior toward the cooperative mean. Because the filtering is itself imperfect, a gap persists between institutional rules and the actual optimum, and that gap is a source of friction that accumulates across the system, degrading agent boundary integrity toward the dissolution threshold (§4.5). When the accumulated departure grows large enough, the result is a *phase transition* in governance: structures that function well during abundance can fail catastrophically under resource stress [33]. Several structural predictions follow.

The dissolution threshold (§4.5) defines the point below which an agent’s recovery becomes physically impossible. A society that permits agents to cross this threshold destroys the cooperative capacity of its own network: each dissolution event eliminates a node from the cooperative web and generates cascading friction costs (§4.2). This result is independent of distributive ideology: agents that cross the dissolution threshold exit the cooperative network permanently, reducing its total capacity. When individuals lose housing, employment, and social support simultaneously, recovery

becomes not just difficult but superlinearly harder, because each lost capacity compounds the next. The dissolution threshold formalizes the point at which this compounding becomes irreversible.

The Coexistence Band width (§4.5) defines measurable bounds on permissible inequality: too narrow prevents productive specialization; too wide pushes peripheral agents toward dissolution while central agents accumulate at the expense of peripheral viability. In economic terms, this is income and wealth inequality: the result identifies formal thresholds beyond which the cooperative equilibrium itself becomes structurally unstable. The multi-center diversification result (§4.5) proves that distributed systems with multiple coexistence attractors are structurally more resilient than monocentric ones, providing a formal basis for the polycentric governance that Ostrom [1990; 2005] demonstrated empirically.

The deception collapse threshold (§4.2) illustrates what a quantitative standard for institutional transparency could look like. For networks of depth $d = 10$, the idealized model gives $q^*(d) \approx 0.063$: above this per-layer deception rate, throughput collapses. The exact number is model-dependent, but the phase structure, that degradation is discontinuous once a threshold is crossed rather than gradual, is robust. Disinformation campaigns, regulatory capture, and institutional corruption are mechanisms that raise the per-layer deception rate toward that threshold.

The relationship to political philosophy is complementary rather than competitive. Hobbes, Locke, Rousseau, Rawls, and Gauthier identified the *problem* (why rational agents would accept mutual constraints) and proposed solutions grounded in thought experiments. The derivation shows that their fundamental intuition was correct: mutual constraint is necessary for coexistence under finite resources, uniquely efficient, and dynamically stable. The results identify formal thresholds (dissolution boundaries, coexistence band limits, deception collapse points) whose violation carries computable costs. How societies respond to these physical realities is a question the theory does not address; it identifies the cost structure, not the response.

8 Limitations and Scope

8.1 What Is Not Claimed

The scope is set by the axioms. The derivations address the structural rules of multi-agent interaction and the function of moral experience for agents satisfying A_1 in shared finite-resource environments, and nothing beyond. They do not address the hard problem of consciousness [89], why subjective experience exists at all as distinct from what it tracks; on the nature of awareness itself the framework is agnostic. Every result is conditional on A_1 : the derivations prove what follows *if* an entity persists, not that entities *should* persist. And while the cooperative equilibrium constrains the space of viable institutional arrangements, it does not pick one out. Many property regimes, governance structures, and legal codes can realize a cooperative equilibrium, and which one a given society converges on depends on factors outside the scope of what is derived here.

8.2 Empirical Testability

The results generate numerically specific, falsifiable predictions:

Prediction	Source	Falsified If...
Cooperation is sustainable for agents with $\delta \geq 0.363$ in energy-denominated games	Theorem 11	Cooperation in physically costly repeated interactions consistently requires discount factors above 0.5
A single defection's net gain is erased by accumulated friction within ~ 20 periods	Theorem 14	Defectors in cost-coupled groups sustain net-positive long-run payoffs despite friction accumulation
Cooperative networks collapse when per-layer corruption exceeds $\sim 6\%$	Theorem 8	Multi-layered cooperative networks tolerate $>10\%$ per-layer corruption without measurable degradation
The energy recovered by destroying a living system is negligible relative to the energy invested in building it ($\mathcal{R}_{BL} \ll 1$)	Corollary 16	Biological complexity at any scale can be reconstructed from raw materials at comparable thermodynamic cost to its original assembly
The relative intensities of moral emotions across cultures correlate with the relative magnitudes of the corresponding cooperative-equilibrium departures	Central Claim §5.2, Part (iii)	No statistically significant correlation exists between moral-emotion intensity rankings and cooperative-equilibrium departure magnitudes across at least three independent cultural samples

The results also predict that cooperation rates should increase with the physical cost of conflict, that isolated defection events should cascade disproportionately through cooperative networks, and that resource concentration beyond the minimum maintenance threshold should trigger discontinuous regime changes. Relevant empirical literatures include experimental game theory [29], commons governance [33], and network-level resilience studies.

The central claim (§5.2, Part (iii)) is itself empirically testable: it predicts that the *structure* of moral intuitions (which behaviors feel right, which wrong, and with what relative intensity) should track the *structure* of the cooperative equilibrium (which strategies are optimal, which dominated, and by what margin).¹²

8.3 Extensions and Open Problems

The most structural extensions concern the interaction model itself. The shared-pool formulation of [4] does not model the endogenous flows between agents. Productive outputs, material or informational, that feed other agents' inputs are not represented, so provision thresholds (the minimum output each agent must sustain to keep downstream agents above their persistence boundaries) are not directly computable, and signalling, the information produced by an agent's behavior and read by others to update their own strategy, is not modelled as a distinct stream. An endogenous-flow extension would cover both.

The cooperation results rest on a binary choice between cooperation and defection, resolved within a single round of the repeated game, over an indefinite horizon. Real interactions involve continu-

¹²The prediction is ordinal: the *ranking* of moral-emotion intensities across violation types should correlate with the *ranking* of cooperative-equilibrium departures. A stronger cardinal test, correlating continuous magnitudes rather than rankings, would require additional machinery described in §8.3.

ous gradations of compliance and sustained contests in which agents allocate resources to conflict until continued commitment threatens dissolution. Extending the model to a continuous action space would characterize the full cooperation manifold and determine whether partial defection is dominated. Introducing an explicit commitment parameter (the resources an agent pledges to a given contest) and a capitulation threshold (the boundary-integrity level at which withdrawal becomes optimal) would permit analysis of conflict duration, sub-critical domination persistence, and the conditions under which graded violations stabilize below cascade thresholds. The horizon assumption should be addressed: A_1 is a property, not a claim of immortality, and finite individual lifespans can in principle activate end-game dynamics that the indefinite-horizon model abstracts away.

Cultural variation is also an open structural extension. While the central claim (§5.2) is robust to variation in moral vocabulary and specific prescriptions, the specific moral systems societies develop out of shared evolved intuitions vary substantially across cultures [46]. Cultural moralities are information structure in the sense of [6]: stores of norms and institutions built up by cultural transmission, but the current information modelling does not encompass different types or classes of information, and a taxonomy of cultural variation within the framework remains open.

Other extensions refine the existing machinery rather than broadening its scope. The results assume standard regularity conditions on utility functions (smoothness and strict concavity), and relaxing these would require different optimization techniques. The value dynamics model [9] represents agent-center coupling as a single scalar; a multi-dimensional extension would refine the coexistence geometry without disturbing the attractor and irreversibility results. The current model is deterministic [7], but real agents face stochastic perturbations (random resource fluctuations, uncertain contest outcomes, exogenous shocks to boundary integrity), and incorporating these would connect the framework to stochastic games and Bayesian conflict models.

The central claim’s ordinal prediction (Part (iii) of §5) is testable now with existing cross-cultural moral psychology data [45; 47]. The stronger cardinal test, correlating moral-emotion intensities with the thermodynamic magnitudes of the corresponding equilibrium departures, would require a parameterization scheme mapping real-world scenarios to the model’s formal quantities, drawing on experimental economics for payoff calibration, information theory for entropy measurement, and calorimetry for thermodynamic cost estimation.

Applied work on AI alignment is a distinct program. AI systems that do not themselves satisfy A_1 function as extensions of their deploying entity’s strategy space, and whether they generate cooperation or friction is determined by the same cost structure that governs any other multi-agent interaction. Current alignment approaches route the alignment signal through subjective human judgment [90; 91], a channel whose fidelity at scale remains an open question [92; 93]. The cooperative equilibrium (Corollary 12) offers a complementary alignment objective derived from physical first principles rather than learned from data, denominated in measurable quantities rather than preference scores, and valid for arbitrarily capable agents without requiring human evaluation.

The framework’s axioms and main results stand on their own, and none of the extensions above would overturn them. What the extensions would do is turn a theory that currently speaks in structural terms into one that speaks in specific ones, with concrete provision thresholds, timescales, cross-cultural parameters, and alignment objectives. That is the work to come, and the list here is not exhaustive.

9 Closing Remarks

The preceding sections are the formal paper. What follows is a personal note on why I wrote it.

We feel clearly that we are only now beginning to acquire reliable material for welding together the sum total of all that is known into a whole; but, on the other hand, it has become next to impossible for a single mind fully to command more than a small specialized portion of it. I can see no other escape from this dilemma (lest our true aim be lost for ever) than that some of us should venture to embark on a synthesis of facts and theories, albeit with second-hand and incomplete knowledge of some of them — and at the risk of making fools of ourselves. — Erwin Schrödinger [10]

Wayne McLaren, a mentor of mine, used to tell me that if you don't take away from others, if you don't do anything to get in another person's way, people will leave you alone to do what you will. Respect a person's boundaries and they have no reason to disrespect yours. He had another code as well: helping people who could use your help when you have the resources and means not only helps them but helps you. Writing this, I realized I haven't specifically thought of Wayne and his principles in a long time, or how closely they track the cooperative equilibrium. What emerged from those principles, at least for myself, was that virtues like honesty, integrity, and respect are important to uphold, at the very least attempt to uphold, if one does not want to be at odds with everyone around them.

If these virtues (including all the ones not named) are so fundamental, then where do they come from? Why did they come to be and how? I felt there must be a foundational source to these seemingly universal principles, one that is not just a commandment written on stone and passed down from generation to generation but can be derived from physical reality itself. From that conviction, I started with two observations: every living system persists by importing energy and exporting entropy, and those systems share an environment where resources are finite.

The theorems that emerged showed that even with natural selection, where competition is a fundamental component of evolution, cooperation is a fundamental component of persistence. Concepts like character, integrity, honesty, and respect are a natural consequence for sentient, thinking beings subject to that cooperative requirement. The traditions and laws we use to hold society together are approximations of this deeper structure, imperfect but pointed in the right direction. If what we call character is the action of striving to maintain the cooperative equilibrium, then whether we like it or not, everyone is part of the system we live in, and if we are at odds with some person or group within this system, we are at odds with ourselves.

Artificial intelligence was another reason I feel this work has value. We struggle to uphold our own principles in every action and every word, for our own integrity, let alone as examples for other people. For the most part it works for us, the human race still exists. Now we are training artificial agents with data based on us, which includes all of those strengths and weaknesses, but with the high likelihood that these agents will be autonomous to a degree we are not in control of at some point. Before that happens, I feel it would be better to ground these agents in the physical principles from which our ethics derive, rather than in the imperfect cultural expressions of those principles.

Schrödinger warned that synthesis requires accepting the risk of making a fool of oneself. I do not profess to be an expert in any one of the fields I drew from to write this paper. Maybe I will be considered a fool for writing equations for *good* and *should*. But if I had the intuition to conceive

of such a thing, and the means to attempt it, and I did not try, then I would not be upholding those principles Wayne taught me and I came to believe in myself.

In so far as a thing is in harmony with our nature, it is necessarily good. — Baruch Spinoza [\[36\]](#)

A Supplementary Appendix: Game-Theoretic Formalization of the Altruism Matrix

This appendix provides the formal game-theoretic underpinning for the boundary-extension resolution of the altruism paradox (§6.2). The body text presents the mechanism in prose along its two axes, spatial and temporal; here we supply the mathematical treatment.

A.1 The Effective Payoff Function

Boundary extension operates by expanding the agent’s effective payoff function to internalize some fraction of other agents’ payoffs. The operative boundary is the Markov blanket [51; 72] whose persistence the system’s organization is directed toward maintaining. A blanket can enclose more than one biological organism (a Markov blanket of Markov blankets, in the terminology of Kirchhoff et al. [72]) and can outlast any particular metabolic vehicle. The coupling weights w_{ij} in the effective payoff function are nonzero precisely when agent j ’s boundary is absorbed into the same persisting blanket as agent i :

$$\Pi_i^{\text{eff}} = \pi_i + \sum_{j \neq i} w_{ij} \pi_j$$

where $w_{ij} \geq 0$ are coupling weights whose source differs by axis. In each case, the agent maximizes its *own* (extended) payoff. Altruism is selfish optimization over an expanded utility function.

A.2 Axis 1: Spatial Extension

Agent i ’s inclusive payoff augments the base payoff with weighted contributions from other agents:

$$\Pi_i^{\text{ext}} = \pi_i + \sum_{j \neq i} w_{ij} \pi_j$$

The coupling weight $w_{ij} \geq 0$ can arise from any source that expands the agent’s effective self-boundary at a given time. The three principal sources form a continuous spectrum:

Genetic relatedness. For a standard diploid organism, $w_{ij} = r_{ij}$ where $r_{ij} \in [0, 1]$ is the relatedness coefficient (probability of shared genetic identity at a random locus): $r_{\text{parent-child}} = 0.5$, $r_{\text{siblings}} = 0.5$, $r_{\text{cousins}} = 0.125$.

Behavioral coordination coupling. Neural mechanisms for action understanding and behavioral coordination [74; 75] generate an additional coupling weight $\epsilon_{\text{coord},ij} \geq 0$, so that $w_{ij} = r_{ij} + \epsilon_{\text{coord},ij}$. The physical basis is shared boundary-maintenance work: agents that coordinate behavior effectively pool thermodynamic resources, forming a higher-order macro-entity even without genetic relatedness. When ϵ_{coord} is large enough, the inclusive-fitness condition triggers for unrelated strangers: $(0 + \epsilon_{\text{coord}})B > C$ can hold even when $r_{ij} = 0$. Non-human primates exhibit this coupling as inequity aversion [76] and equitable outcome preference [77], confirming that the mechanism predates language and conscious deliberation.

Full structural merger. At the limit, two agents merge boundaries entirely (parent-child dyads, spouses, military units), producing $w_{ij} \rightarrow 1$. The merged pair becomes a single composite entity with a joint utility function:

$$U_{AB}(\mathbf{x}_A, \mathbf{x}_B) = w_A U_A(\mathbf{x}_A) + w_B U_B(\mathbf{x}_B)$$

where $w_A, w_B > 0$. The “self-sacrifice” of component A for component B is internal resource reallocation within the merged system.

Hamilton’s Rule (unified). Regardless of source, agent i sacrifices personal payoff $\pi_i = -C$ when:

$$w_{ij} \cdot B > C$$

where B is the benefit (in energy units) to agent j . The mathematical structure is identical across the spectrum: genetic relatedness, behavioral coordination coupling, and full structural merger all produce nonzero w_{ij} via thermodynamic coupling at the expanded boundary, differing only in the biological mechanism that generates the weight and its magnitude.

A.3 Axis 2: Temporal Extension

Temporal extension is not an independent mechanism but the same boundary extension viewed across time: a host blanket that encloses agent i ’s accumulated negentropy continues to enclose it after i ’s metabolic vehicle dissipates. An agent’s identity is partially constituted by its accumulated negentropy: functional information built up over developmental and evolutionary history. Biological reproduction, cultural transmission, and institutional embedding are physical mechanisms by which this informational structure is absorbed into a host blanket whose persistence carries it forward. An agent whose accumulated negentropy is substantially encoded in such a host blanket has an effective planning horizon extending beyond its individual lifespan.

The effective discount factor for such an agent approaches 1:

$$\delta_{\text{extended}} = 1 - \epsilon \approx 1$$

where $\epsilon \rightarrow 0$ reflects the degree to which the agent’s informational structure is embedded in persistent external structures. Under Theorem 11, any $\delta > \delta^*$ sustains cooperation. For $\delta \rightarrow 1$, any finite short-term cost is offset by the long-term cooperative payoff stream (see accumulated negentropy discussion, §4.4; [6]). Self-sacrifice of the metabolic vehicle functions as a commitment device [55]: by irrevocably investing accumulated negentropy in external structures, the agent credibly commits to the cooperative equilibrium, maximizing the persistence probability of its informational structure (Corollary 16; Theorem 17). In human agents, the subjective experience of this embeddedness is the belief that identity persists beyond biological death. This belief is the heuristic approximation [94; 56], not the load-bearing premise; the physical reality is that the agent’s accumulated negentropy does persist through the structures encoding its functional information.

A.4 Summary

Axis	What Expands	Scope of “Self”	Example
1. Spatial	Who counts as self at a given time	Kin \rightarrow cooperative partners \rightarrow merged entity	Parent for child; soldier for unit; apoptotic cell
2. Temporal	How far the same expanded self extends in time	Informational persistence beyond metabolic vehicle	Reproduction; cultural transmission; martyrdom

The unified equation $\Pi_i^{\text{eff}} = \pi_i + \sum_{j \neq i} w_{ij} \pi_j$ subsumes both axes: the spatial axis sets the scope of the host blanket through nonzero w_{ij} , and the temporal axis is the persistence of that blanket beyond the metabolic vehicle, expressed as $\delta_{\text{extended}} \rightarrow 1$. The framework does not explain away altruism or reduce it to disguised selfishness; it expands the definition of self. The entity satisfying A_1 is not necessarily the biological organism but whatever boundary the effective payoff function encompasses.

B Supplementary Appendix: The Gradient Structure of Good and Ought

The central claim (§5.2) asserts that *good* and *ought* are gene-culture co-evolved heuristic approximations of the cooperative equilibrium. This appendix provides the mathematical formalization: the welfare function defined in [4] and [8] generates a gradient field over strategy space from which *good*, *ought*, and their context dependence emerge as distinct mathematical objects.

B.1 Setup

We adopt the Lagrangian formulation from [4] without restatement. Let $W(\mathbf{x}) = \sum_{i=1}^N U_i(\mathbf{x}_i)$ be the social welfare function over the joint strategy profile \mathbf{x} , with W strictly concave on the feasible set \mathcal{F} defined by the boundary constraints $g_k(\mathbf{x}) \leq 0$, and let

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}) = W(\mathbf{x}) - \sum_{k=1}^K \mu_k g_k(\mathbf{x})$$

be the associated Lagrangian with shadow prices $\boldsymbol{\mu} \geq \mathbf{0}$. The cooperative equilibrium \mathbf{x}^* is the unique global maximizer of W on \mathcal{F} (Theorem 4), at which the KKT stationarity condition $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^*, \boldsymbol{\mu}^*) = \mathbf{0}$ holds together with complementary slackness [4]. The relevant first-order quantity for what follows is therefore the Lagrangian gradient $\nabla_{\mathbf{x}} \mathcal{L}$, which vanishes at \mathbf{x}^* , rather than ∇W , which generally does not when boundary constraints bind. Away from \mathbf{x}^* , $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\mu}^*) \neq \mathbf{0}$, and strict concavity guarantees $W(\mathbf{x}) < W(\mathbf{x}^*)$ for all feasible $\mathbf{x} \neq \mathbf{x}^*$.

B.2 Good as a Scalar Quantity

Define the **degree of good** of a configuration \mathbf{x} as the normalized welfare:

$$G(\mathbf{x}) = \frac{W(\mathbf{x})}{W(\mathbf{x}^*)}$$

where $W(\mathbf{x}^*) > 0$ (the cooperative surplus over the no-action baseline is strictly positive whenever cooperation is feasible, Proposition 10). This yields a scalar $G \in (-\infty, 1]$ with the following properties:

1. $G(\mathbf{x}^*) = 1$: the cooperative equilibrium is maximally good.
2. G is strictly concave and has a unique maximum.
3. $G < 1$ for all $\mathbf{x} \neq \mathbf{x}^*$: every departure from the equilibrium reduces goodness.
4. G can be negative: configurations involving mutual defection destroy more value than they produce ($P < 0$, Proposition 10), yielding $W < 0$ and therefore $G < 0$.

Good is therefore a gradable, measurable, context-dependent scalar. The question “is X good?” reduces to “what is $G(\mathbf{x})$?”

B.3 Ought as a Vector

At any feasible configuration $\mathbf{x} \neq \mathbf{x}^*$, the Lagrangian gradient

$$\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}^*)$$

is non-zero and points in the direction of steepest welfare ascent that respects the binding boundary constraints, evaluated at the equilibrium shadow prices. In natural language, an “ought” statement (“you ought to be more honest,” “you ought to share more equitably”) is a culturally compressed encoding of one or more components of this vector. The full vector carries both information about which adjustments would increase welfare and the rate of welfare gain those adjustments would produce; analytically, these decompose into direction and magnitude.

Direction. The normalized gradient

$$\hat{\mathbf{d}}(\mathbf{x}) = \frac{\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}^*)}{|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}^*)|}$$

identifies the welfare-improving dimension of the ought: *what* to change. Strict concavity of $\mathcal{L}(\cdot, \boldsymbol{\mu}^*)$ ensures that the gradient always has a positive component toward the equilibrium, so gradient flow on $\mathcal{L}(\cdot, \boldsymbol{\mu}^*)$ converges to \mathbf{x}^* . Writing $\mathbf{x}_i = (x_{i1}, \dots, x_{iM})$ for agent i ’s allocation across the M resource dimensions, the components \hat{d}_{ij} specify, for each agent i and resource dimension j , the direction of welfare-improving strategy adjustment.

Magnitude. The norm

$$\|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}^*)\|$$

captures the scalar magnitude of the ought: the rate of welfare gain available at the current configuration. The larger the gradient norm, the steeper the welfare landscape and the greater the gain available from local adjustment. Near the equilibrium, $|\nabla_{\mathbf{x}}\mathcal{L}| \rightarrow 0$ and ought magnitude diminishes, which corresponds to the observation that ought-language carries greater emphasis in configurations far from the equilibrium than in those already near-optimal.

B.4 Context Dependence

The Lagrangian gradient $\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}^*)$ depends on the local parameters of the welfare landscape:

- Resource endowments \mathbf{R}
- Friction coefficients Φ
- Network topology (which agents interact with which)
- Shadow prices μ_k^* (the cost structure of boundary constraints)
- Population structure (number and heterogeneity of agents)

The same action (an allocation, a communication strategy, a boundary policy) produces different gradient magnitudes and directions under different parameterizations. A subsistence community and an affluent industrial economy have different welfare landscapes; the action that maximizes welfare in one context may differ from the welfare-maximizing action in the other.

This context dependence is not a weakness of the formalization but its central feature. Variation in ought-language across cultures, documented extensively in the empirical literature [46; 45; 47], is predicted by the framework: each culture’s institutional and material conditions produce a different welfare landscape, and the gradient of that landscape produces locally adapted ought-heuristics. What the framework predicts is invariant is not the direction of the gradient (the specific heuristic) but the existence and uniqueness of the maximum (the cooperative equilibrium toward which each gradient points).

B.5 Summary

The Lagrangian of the welfare function generates two primary mathematical objects, one of which decomposes into two analytically distinct components:

Object	Mathematical form	Corresponds to
Good (quantity)	$G(\mathbf{x}) = W(\mathbf{x})/W(\mathbf{x}^*)$	Evaluative language (“really good,” “not bad,” “terrible”)
Ought (vector)	$\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}^*)$	Ought-language (“you should do X”)
<i>direction</i>	$\hat{\mathbf{d}}(\mathbf{x}) = \nabla_{\mathbf{x}}\mathcal{L}/\ \nabla_{\mathbf{x}}\mathcal{L}\ $ (undefined at \mathbf{x}^*)	What “to do”
<i>magnitude</i>	$\ \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}^*)\ $	How important it is “to do it”

The cooperative equilibrium is the unique fixed point at which $G = 1$ and $\nabla_{\mathbf{x}}\mathcal{L} = \mathbf{0}$ (the ought vector vanishes, so neither direction nor magnitude is defined). Every other feasible configuration has a well-defined degree of good and a well-defined ought vector with distinct direction and magnitude. Both objects are context-dependent through the parameters of W and the active boundary constraints, and both are computable in principle from the physical quantities the framework tracks.

References

- [1] David Hume. *A Treatise of Human Nature*. Oxford University Press, 2000. Original work published 1739.
- [2] Ruth Garrett Millikan. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press, 1984.
- [3] Daniel C. Dennett. *The Intentional Stance*. MIT Press, 1987.
- [4] Keith Lostracco. Thermodynamics of cooperation: Necessary constraints, 2026. URL <https://doi.org/10.5281/zenodo.19635523>. doi:10.5281/zenodo.19635523.
- [5] Keith Lostracco. Thermodynamics of cooperation: Strategic entropy injection, 2026. URL <https://doi.org/10.5281/zenodo.19635814>. doi:10.5281/zenodo.19635814.
- [6] Keith Lostracco. Thermodynamics of cooperation: Accumulated negentropy, 2026. URL <https://doi.org/10.5281/zenodo.19635836>. doi:10.5281/zenodo.19635836.
- [7] Keith Lostracco. Thermodynamics of cooperation: Thermodynamic friction, 2026. URL <https://doi.org/10.5281/zenodo.19635850>. doi:10.5281/zenodo.19635850.

- [8] Keith Lostracco. Thermodynamics of cooperation: Cooperative equilibrium, 2026. URL <https://doi.org/10.5281/zenodo.19635856>. doi:10.5281/zenodo.19635856.
- [9] Keith Lostracco. Thermodynamics of cooperation: Value dynamics, 2026. URL <https://doi.org/10.5281/zenodo.19635865>. doi:10.5281/zenodo.19635865.
- [10] Erwin Schrödinger. *What Is Life? The Physical Aspect of the Living Cell*. Cambridge University Press, 1944.
- [11] Ilya Prigogine. Time, structure, and fluctuations. *Science*, 201(4358):777–785, 1978. doi:10.1126/science.201.4358.777.
- [12] Grégoire Nicolis and Ilya Prigogine. *Self-Organization in Nonequilibrium Systems: From Dissipative Structures to Order Through Fluctuations*. Wiley, 1977.
- [13] Jeremy L. England. Statistical physics of self-replication. *The Journal of Chemical Physics*, 139(12):121923, 2013. doi:10.1063/1.4818538.
- [14] Karl Friston. A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1–3): 70–87, 2006. doi:10.1016/j.jphysparis.2006.10.001.
- [15] Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. doi:10.1038/nrn2787.
- [16] Humberto R. Maturana and Francisco J. Varela. *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel, 1980. Original work published 1972.
- [17] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi:10.1002/j.1538-7305.1948.tb01338.x.
- [18] Leo Szilard. Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen. *Zeitschrift für Physik*, 53(11–12):840–856, 1929. doi:10.1007/BF01341281.
- [19] Rolf Landauer. Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3):183–191, 1961. doi:10.1147/rd.53.0183.
- [20] Charles H. Bennett. Logical reversibility of computation. *IBM Journal of Research and Development*, 17(6):525–532, 1973. doi:10.1147/rd.176.0525.
- [21] Charles H. Bennett. The thermodynamics of computation—a review. *International Journal of Theoretical Physics*, 21(12):905–940, 1982. doi:10.1007/BF02084158.
- [22] Christoph Adami. Information theory in molecular biology. *Physics of Life Reviews*, 1(1): 3–22, 2004. doi:10.1016/j.plrev.2004.01.002.
- [23] John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.
- [24] John F. Nash. Equilibrium points in N-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950. doi:10.1073/pnas.36.1.48.
- [25] John F. Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951. doi:10.2307/1969529.

- [26] Drew Fudenberg and Eric Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–554, 1986. doi:[10.2307/1911307](https://doi.org/10.2307/1911307).
- [27] John Maynard Smith and George R. Price. The logic of animal conflict. *Nature*, 246(5427): 15–18, 1973. doi:[10.1038/246015a0](https://doi.org/10.1038/246015a0).
- [28] William D. Hamilton. The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*, 7(1):1–16, 1964. doi:[10.1016/0022-5193\(64\)90038-4](https://doi.org/10.1016/0022-5193(64)90038-4).
- [29] Robert Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.
- [30] Jörgen W. Weibull. *Evolutionary Game Theory*. MIT Press, 1995.
- [31] Nicholas Georgescu-Roegen. *The Entropy Law and the Economic Process*. Harvard University Press, 1971.
- [32] Howard T. Odum. *Environment, Power, and Society*. Wiley-Interscience, 1971.
- [33] Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990.
- [34] Elinor Ostrom. *Understanding Institutional Diversity*. Princeton University Press, 2005.
- [35] Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968. doi:[10.1126/science.162.3859.1243](https://doi.org/10.1126/science.162.3859.1243).
- [36] Baruch Spinoza. *Ethica ordine geometrico demonstrata*. Jan Rieuwertsz, Amsterdam, 1677. Published posthumously in *Opera Posthuma*.
- [37] Thomas Hobbes. *Leviathan, or The Matter, Forme, & Power of a Common-Wealth Ecclesiastical and Civill*. Andrew Crooke, 1651.
- [38] John Locke. *Two Treatises of Government*. Awnsham Churchill, 1689.
- [39] Jean-Jacques Rousseau. *Du contrat social; ou, principes du droit politique*. Marc Michel Rey, 1762.
- [40] John Rawls. *A Theory of Justice*. Harvard University Press, 1971.
- [41] George Edward Moore. *Principia Ethica*. Cambridge University Press, 1903.
- [42] Claude-Adrien Helvétius. *De l’esprit*. Durand, 1758.
- [43] Mehdi Bazargan. *Eshq va Parastesh ya Thermodynamic-e Ensan [Love and Worship, or The Thermodynamics of Man]*. 1956. In Persian. Discussed in [44].
- [44] Charles Kurzman. Liberal islam, 1998. URL <https://api.semanticscholar.org/CorpusID:211128980>.
- [45] Jonathan Haidt. The moral emotions. In Scherer K. R. Davidson R. J. and Goldsmith H. H., editors, *Handbook of Affective Sciences*, pages 852–870. Oxford University Press, 2003.
- [46] Jonathan Haidt and Selin Kesebir. Morality. In Susan T. Fiske, Daniel T. Gilbert, and Gardner Lindzey, editors, *Handbook of Social Psychology*, pages 797–832. Wiley, 5th edition, 2010. doi:[10.1002/9780470561119.socpsy002022](https://doi.org/10.1002/9780470561119.socpsy002022).

- [47] Oliver Scott Curry, Daniel Austin Mullins, and Harvey Whitehouse. Is it good to cooperate?: Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, 60(1): 47–69, 2019. doi:[10.1086/701478](https://doi.org/10.1086/701478). URL <https://doi.org/10.1086/701478>.
- [48] Rudolf Clausius. Ueber verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie. *Annalen der Physik und Chemie*, 201(7):353–400, 1865. doi:[10.1002/andp.18652010702](https://doi.org/10.1002/andp.18652010702).
- [49] Ludwig Boltzmann. Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive den Sätzen über das Wärmegleichgewicht. *Wiener Berichte*, 76:373–435, 1877.
- [50] Max Planck. Zur Theorie des Gesetzes der Energieverteilung im Normalspectrum. *Verhandlungen der Deutschen Physikalischen Gesellschaft*, 2:237–245, 1900.
- [51] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.
- [52] Karl Friston. Life as we know it. *Journal of the Royal Society Interface*, 10(86):20130475, 2013. doi:[10.1098/rsif.2013.0475](https://doi.org/10.1098/rsif.2013.0475).
- [53] Gerd Gigerenzer, Peter M. Todd, and ABC Research Group. *Simple Heuristics That Make Us Smart*. Oxford University Press, 1999.
- [54] Gerd Gigerenzer and Wolfgang Gaissmaier. Heuristic decision making. *Annual Review of Psychology*, 62:451–482, 2011. doi:[10.1146/annurev-psych-120709-145346](https://doi.org/10.1146/annurev-psych-120709-145346).
- [55] Robert H. Frank. *Passions Within Reason: The Strategic Role of the Emotions*. W. W. Norton & Co., 1988.
- [56] Gerd Gigerenzer. Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science*, 2(3):528–554, 2010. doi:[10.1111/j.1756-8765.2010.01094.x](https://doi.org/10.1111/j.1756-8765.2010.01094.x).
- [57] Antonio R. Damasio. *Descartes’ Error: Emotion, Reason, and the Human Brain*. G. P. Putnam’s Sons, 1994.
- [58] Antoine Bechara, Hanna Damasio, Daniel Tranel, and Antonio R. Damasio. Deciding advantageously before knowing the advantageous strategy. *Science*, 275(5304):1293–1295, 1997. doi:[10.1126/science.275.5304.1293](https://doi.org/10.1126/science.275.5304.1293).
- [59] Antoine Bechara and Antonio R. Damasio. The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior*, 52(2):336–372, 2005. doi:[10.1016/j.geb.2004.06.010](https://doi.org/10.1016/j.geb.2004.06.010).
- [60] Robert Boyd and Peter J. Richerson. *Culture and the Evolutionary Process*. University of Chicago Press, 1985.
- [61] Joseph Henrich. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton University Press, Princeton, NJ, 2015. ISBN 9780691166858.
- [62] Jonathan Haidt. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Knopf Doubleday Publishing Group, 2012.

- [63] Daniel C. Dennett. *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. Simon & Schuster, 1995.
- [64] Philippa Foot. *Natural Goodness*. Oxford University Press, 2001.
- [65] Rosalind Hursthouse. *On Virtue Ethics*. Oxford University Press, 1999.
- [66] Matthew Lutz and James Lenman. Moral naturalism. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Stanford University, summer 2024 edition, 2024. URL <https://plato.stanford.edu/archives/sum2024/entries/naturalism-moral/>.
- [67] John R. Searle. How to derive “ought” from “is”. *The Philosophical Review*, 73(1):43–58, 1964. doi:[10.2307/2183201](https://doi.org/10.2307/2183201).
- [68] Constantinos Daskalakis, Paul W. Goldberg, and Christos H. Papadimitriou. The complexity of computing a nash equilibrium. *Communications of the ACM*, 52(2):89–97, 2009. doi:[10.1145/1461928.1461951](https://doi.org/10.1145/1461928.1461951).
- [69] A. Aldo Faisal, Luc P. J. Selen, and Daniel M. Wolpert. Noise in the nervous system. *Nature Reviews Neuroscience*, 9(4):292–303, 2008. doi:[10.1038/nrn2258](https://doi.org/10.1038/nrn2258).
- [70] John Tooby and Leda Cosmides. The psychological foundations of culture. In Jerome H. Barkow, Leda Cosmides, and John Tooby, editors, *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pages 19–136. Oxford University Press, 1992.
- [71] Leda Cosmides and John Tooby. Evolutionary psychology: New perspectives on cognition and motivation. *Annual Review of Psychology*, 64:201–229, 2013. doi:[10.1146/annurev.psych.121208.131628](https://doi.org/10.1146/annurev.psych.121208.131628).
- [72] Michael Kirchhoff, Thomas Parr, Ensor Palacios, Karl Friston, and Julian Kiverstein. The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, 15(138):20170792, 2018. doi:[10.1098/rsif.2017.0792](https://doi.org/10.1098/rsif.2017.0792).
- [73] Susan Elmore. Apoptosis: A review of programmed cell death. *Toxicologic Pathology*, 35(4): 495–516, 2007. doi:[10.1080/01926230701320337](https://doi.org/10.1080/01926230701320337).
- [74] Giacomo Rizzolatti and Laila Craighero. The mirror-neuron system. *Annual Review of Neuroscience*, 27:169–192, 2004. doi:[10.1146/annurev.neuro.27.070203.144230](https://doi.org/10.1146/annurev.neuro.27.070203.144230).
- [75] Frans B. M. de Waal. Putting the altruism back into altruism: The evolution of empathy. *Annual Review of Psychology*, 59:279–300, 2008. doi:[10.1146/annurev.psych.59.103006.093625](https://doi.org/10.1146/annurev.psych.59.103006.093625).
- [76] Sarah F. Brosnan and Frans B. M. de Waal. Monkeys reject unequal pay. *Nature*, 425(6955): 297–299, 2003. doi:[10.1038/nature01963](https://doi.org/10.1038/nature01963).
- [77] Darby Proctor, Rebecca A. Williamson, Frans B. M. de Waal, and Sarah F. Brosnan. Chimpanzees play the ultimatum game. *Proceedings of the National Academy of Sciences*, 110(6): 2070–2075, 2013. doi:[10.1073/pnas.1220806110](https://doi.org/10.1073/pnas.1220806110).
- [78] Frans B. M. de Waal. *Primates and Philosophers: How Morality Evolved*. Princeton University Press, 2006.

- [79] David Sloan Wilson and Edward O. Wilson. Rethinking the theoretical foundation of sociobiology. *The Quarterly Review of Biology*, 82(4):327–348, 2007. doi:[10.1086/522809](https://doi.org/10.1086/522809).
- [80] Elliott Sober and David Sloan Wilson. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press, 1998.
- [81] Herbert Spencer. *The Principles of Biology*, volume 1. Williams and Norgate, 1864.
- [82] Richard Hofstadter. *Social Darwinism in American Thought*. University of Pennsylvania Press, 1944.
- [83] Douglas W. Smith, Rolf O. Peterson, and Douglas B. Houston. Yellowstone after wolves. *BioScience*, 53(4):330–340, 2003. doi:[10.1641/0006-3568\(2003\)053\[0330:YAW\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2003)053[0330:YAW]2.0.CO;2).
- [84] William J. Ripple and Robert L. Beschta. Trophic cascades in Yellowstone: the first 15 years after wolf reintroduction. *Biological Conservation*, 145(1):205–213, 2012. doi:[10.1016/j.biocon.2011.11.005](https://doi.org/10.1016/j.biocon.2011.11.005).
- [85] David Gauthier. *Morals by Agreement*. Oxford University Press, 1986.
- [86] Amartya Sen. *Commodities and Capabilities*. North-Holland, Amsterdam, 1985. URL http://www.amazon.com/Commodities-Capabilities-Amartya-Sen/dp/0195650387/ref=sr_1_1?s=books&ie=UTF8&qid=1310679705&sr=1-1.
- [87] Jonathan Chapman and Geoffrey Fisher. Preference elicitation: common methods and potential pitfalls. In Erik Snowberg and Leeat Yariv, editors, *Handbook of Experimental Methodology*, volume 1, chapter 2, pages 25–80. North-Holland, 2025. ISBN 9780443317583. doi:[10.1016/bs.hbem.2025.09.001](https://doi.org/10.1016/bs.hbem.2025.09.001).
- [88] Matthew O. Jackson. *Social and Economic Networks*. Princeton University Press, 2008. ISBN 978-0-691-13440-6.
- [89] David J. Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–219, 1995.
- [90] Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martić, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, pages 4299–4307, 2017.
- [91] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.
- [92] Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiuūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Chris Olah, Dario Amodei, Daniela Amodei, Dawn Drain, Danny Li, Eli Tran-Johnson, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022. doi:[10.48550/arXiv.2211.03540](https://doi.org/10.48550/arXiv.2211.03540).

- [93] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Sharkey, Ansh Sarat, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Aidan Segerie, Micah Carroll, Andi Peng, Phillip Christofersen, Mehul Damber, Stewart Slocum, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023. doi:[10.48550/arXiv.2307.15217](https://doi.org/10.48550/arXiv.2307.15217).
- [94] Jonathan Haidt. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814–834, 2001. doi:[10.1037/0033-295X.108.4.814](https://doi.org/10.1037/0033-295X.108.4.814).