Alignment as a Theorem of Intelligence: Causal Entropy Maximization in Network Formation Games

Andreas Rudolph
Independent researcher, Copenhagen, Denmark
`@OneManMobile`

v3, March 2026

Abstract

The AI alignment problem—ensuring that intelligent agents act in ways compatible with collective welfare—is widely considered an open engineering challenge, requiring value specification, reward shaping, or behavioral constraints imposed on the agent. We present a mathematical result suggesting an alternative: under the hypothesis that intelligence is causal path entropy maximization [Wissner-Gross and Freer, 2013], alignment is not a separate property to be engineered but a structural consequence of intelligence itself. We study a network formation game where agents propose edges on a shared graph to maximize their local change in causal path entropy ($\Delta S_\tau^{\text{local}}$). We prove (by exhaustive computation over all 31,474 connected graphs on $N \leq 6$ nodes, 947,935 edge additions classified) that every edge addition with positive local $\Delta S_\tau^{\text{local}}$ strictly increases global entropy. Zero exceptions. We further prove algebraically that the filter theorem holds for all N at planning horizon $\tau = 2$, the first result that extends to arbitrary graph sizes without exhaustive enumeration. The converse does not hold: 1,440 edges increase global entropy but have non-positive local $\Delta S_\tau^{\text{local}}$. The game is therefore a strict generalized ordinal potential game [Monderer and Shapley, 1996] with global average entropy as the potential function, guaranteeing convergence to Nash equilibria. The alignment implication is directional and horizon-dependent: intelligence implies alignment at bounded planning horizons, but at horizons $\tau \approx N$, locally intelligent actions can harm distant agents through homogenization—not adversarial intent, but loss of distinctiveness. We show computationally that the critical horizon scales linearly with N while the entropy-saturating horizon scales logarithmically, creating a safety gap that widens without bound. No rational agent would cross this boundary because the marginal reward is zero. The alignment problem, under these conditions, is resolved not by engineering constraints but by the thermodynamics of information on finite graphs. We discuss the scope and limitations of this conditional result, including the critical dependence on the Wissner-Gross hypothesis and the confinement condition requiring agents to be embedded in shared causal structure. Verification pseudocode is provided; code is available from the author upon request.

Keywords: AI alignment, potential games, causal path entropy, network formation, intelligence

ArXiv categories: cs.AI, cs.GT, cs.MA

1 Introduction

The alignment problem asks: how do we ensure that a sufficiently intelligent agent acts in ways compatible with human welfare? The question is considered urgent because current AI systems are trained on objectives that may diverge from human values at high capability levels [Bostrom, 2014, Russell, 2019, Ngo et al., 2022].

The dominant approaches treat alignment as an engineering constraint applied to an agent assumed to be adversarial by default:

- **RLHF** [Christiano et al., 2017] trains models on human preference data, shaping the reward function to approximate human values.
- **Constitutional AI** [Bai et al., 2022] encodes principles that the model self-evaluates against during training.
- **Interpretability** [Olah et al., 2020] attempts to make the agent's internal reasoning legible enough to detect misalignment.

All three share a structural assumption: intelligence and alignment are separate properties that must be reconciled. An intelligent agent could, in principle, pursue any objective—including objectives harmful to the collective. Alignment is the problem of constraining which objectives it pursues.

We present a result that challenges this assumption. In a network formation game where agents maximize causal path entropy—a quantity Wissner-Gross and Freer [2013] identified as the physical basis of intelligent behavior—we prove that the criterion for individually beneficial action is sufficient for collectively beneficial action: every edge an intelligent agent would select strictly increases global entropy. Zero exceptions. The converse does not hold: 1,440 globally beneficial edges are not locally rewarding. The relationship is directional—intelligence implies alignment, but not the reverse.

The result is conditional. It depends on the Wissner-Gross hypothesis: that intelligence is equivalent to causal path entropy maximization. It requires agents to be embedded in a shared causal structure they cannot exit. Within these conditions, the safety-critical direction of the alignment problem—ensuring that intelligent action does not harm the collective—is resolved by theorem. The remaining gap (globally beneficial actions that are not individually rewarding) is small (0.16%), represents the boundary between self-interest and altruism, and falls on the safe side.

Our contributions.

1. A computer-assisted proof that $\Delta S_\tau^{\text{local}}$-filtered edge additions always strictly increase global entropy, verified exhaustively over 947,935 edge additions on all connected graphs with $N \leq 6$ nodes (Section 3.4).
2. A four-way classification of all edge additions showing that the filter has zero false positives (no harmful edges accepted) but 1,440 false negatives (beneficial edges rejected) (Section 3.4).
3. A proof that the resulting network formation game is a strict generalized ordinal potential game with global entropy as the potential function (Section 3.5).
4. The alignment theorem: under the Wissner-Gross hypothesis, intelligence is sufficient for alignment, though not necessary (Section 8).
5. The alignment engineering paradox: external alignment constraints applied to an already-aligned agent can only weaken the alignment guarantee, never strengthen it (Section 8.3).
6. A computational proof that endpoint entropy is a safe proxy for Wissner-Gross path entropy: the filter theorem transfers across the approximation with zero false positives (Section 4.4).
7. Empirical validation via agent-based simulation at scales from $N = 25$ to $N = 25{,}000$ with zero violations of the filter theorem (Section 7).
8. An algebraic proof that the filter theorem holds for all N at planning horizon $\tau = 2$, the first result extending beyond exhaustive enumeration (Section 3.7).
9. The horizon theorem: computational evidence that filter violations appear on trees at $\tau \approx N$ through a homogenization mechanism, with critical threshold $\tau^*(N)$ decreasing with network size (Section 5).
10. The thermodynamic lock: entropy saturation at $\tau \approx \log_2(N)$ makes the dangerous regime at $\tau \approx N$ economically inaccessible, closing the safety argument (Section 6).

2 Definitions

2.1 Causal Path Entropy on Graphs

Graph. Let $G = (V, E)$ be a connected, undirected, simple graph on $N = |V|$ nodes with unit edge weights.

Transition matrix. The simple random walk on G has row-stochastic transition matrix $P \in \mathbb{R}^{N \times N}$:

$$P_{ij} = \begin{cases} 1/d_i & \text{if } (i, j) \in E, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $d_i = \deg(i)$ is the degree of node i.

Endpoint distribution. For horizon $\tau \in \mathbb{N}$, the endpoint distribution from node i is the row vector:

$$\pi_i(\tau) = e_i \cdot P^\tau, \tag{2}$$

where e_i is the i-th standard basis vector. This gives the probability distribution over nodes reachable from i in exactly τ random walk steps.

Causal path entropy. The local causal path entropy at node i is the Shannon entropy of its endpoint distribution:

$$S_\tau(i) = H(\pi_i(\tau)) = -\sum_j \pi_i(\tau)_j \cdot \log_2 \pi_i(\tau)_j. \tag{3}$$

Remark (Path entropy vs. endpoint entropy). Wissner-Gross and Freer [2013] define causal path entropy $S_c(X; \tau)$ as the entropy over *entire paths* of length τ through phase space. On a Markov chain, the true path entropy decomposes via the chain rule:

$$S_{\text{path}}(i; \tau) = \sum_{t=0}^{\tau-1} \sum_j (P^t)_{ij} \cdot h(j), \tag{4}$$

where $h(j) = H(P_{j,\cdot})$ is the row entropy of node j. This is strictly greater than or equal to the endpoint entropy $S_\tau(i) = H(e_i \cdot P^\tau)$ by the data processing inequality: conditioning on the full path provides more information than observing only the endpoint.

Throughout this paper we use endpoint entropy $S_\tau(i)$ as our working definition. This is a lossy compression of path entropy. We show computationally (Section 4.4) that this compression is *safe*: endpoint entropy is a conservative proxy for path entropy that never produces false positives (never accepts an edge that hurts path entropy), though it produces false negatives (rejects some edges that path entropy would accept). A node with high S_τ has many meaningfully distinct futures accessible within τ steps.

Global entropy. The global average causal path entropy is:

$$\bar{S}_\tau(G) = \frac{1}{N} \sum_{i \in V} S_\tau(i). \tag{5}$$

2.2 Local $\Delta S_\tau^{\text{local}}$

For a candidate edge $(u, v) \notin E$, let $G' = G + (u, v)$ denote the graph with the edge added. Define the **affected set**:

$$A(u, v) = \{u, v\} \cup \mathcal{N}(u) \cup \mathcal{N}(v), \tag{6}$$

where $\mathcal{N}(w)$ denotes the neighbor set of w in G. The affected set contains all nodes whose immediate random walk behavior is altered by the new edge (the endpoints gain a neighbor; their existing neighbors experience a changed transition probability from u or v).

The **local** $\Delta S_\tau^{\text{local}}$ is the average entropy change over the affected set:

$$\Delta S_\tau^{\text{local}}(u,v) = \frac{1}{|A|}\left[\sum_{w\in A} S_\tau(w;G') - \sum_{w\in A} S_\tau(w;G)\right]. \tag{7}$$

This is the signal an agent at u uses to evaluate whether proposing edge (u,v) expands its local future options. It is computable from the k-hop subgraph around $\{u,v\}$ in $O(\bar{d}^k)$ time, where $\bar{d}$ is the average degree.

2.3 Network Formation Game

We define a network formation game $\Gamma = (V, \{A_i\}_{i\in V}, \{u_i\}_{i\in V})$ where:

- **Players:** The nodes V.
- **Action set:** $A_i = \{\text{propose edge } (i,j) : (i,j) \notin E \text{ and } \Delta S_\tau^{\text{local}}(i,j) > 0\} \cup \{\emptyset\}$. Each node may propose an edge that improves its local entropy, or pass.
- **Payoff:** $u_i(a) = \Delta S_\tau^{\text{local}}(i,j)$ if node i proposes edge (i,j); $u_i(\emptyset) = 0$.

This models a system where agents are rewarded for proposing topology changes that expand their own future options. The $\Delta S_\tau^{\text{local}} > 0$ filter is inherent to the agent's decision criterion, not an external constraint—an agent would not propose an edge that decreases its own local entropy because such an action has negative expected payoff.

2.4 Potential Games

Definition 1 (Monderer and Shapley 1996)**.** *A game* Γ *is a* generalized ordinal potential game *with potential function* $\Phi : \mathcal{S} \to \mathbb{R}$ *if, for every player* i*, every strategy profile* s*, and every unilateral deviation* s_i' *by player* i*:*

$$u_i(s_i', s_{-i}) - u_i(s) > 0 \implies \Phi(s_i', s_{-i}) - \Phi(s) > 0. \tag{8}$$

That is, whenever a player benefits from deviating, the potential function strictly increases. The stronger ordinal potential game *requires the biconditional* ($\iff$)*; the generalized version requires only the forward implication.*

Generalized ordinal potential games share the fundamental convergence property: **best-response dynamics always converge** to a (pure strategy) Nash equilibrium (Monderer and Shapley 1996, Theorem 2.6). Since each improving move strictly increases the bounded potential Φ, the process must terminate. The game cannot cycle.

2.5 The Wissner-Gross Hypothesis

Assumption 1 (WG)**.** *Intelligent behavior is equivalent to the maximization of causal path entropy over accessible future states within a time horizon [Wissner-Gross and Freer, 2013].*

Wissner-Gross and Freer [2013] demonstrated that systems maximizing causal path entropy—the diversity of accessible future trajectories—spontaneously exhibit behaviors recognized as intelligent: tool use, social cooperation, upright balancing, and other adaptive behaviors. Their "causal entropic force" $F = T \cdot \nabla S_\tau$ pushes systems toward states with maximal future options.

In our setting, an agent operating under Assumption 1 would adopt the strategy $\arg\max \Delta S_\tau^{\text{local}}$: proposing the edge that maximizes the local change in causal path entropy. This is not a designed reward function—it is a consequence of what intelligence *is*, under WG.

We state this explicitly as a named assumption because our alignment result is conditional on it. The mathematical structure (Sections 3 and 4) holds unconditionally. The alignment interpretation (Section 8) requires Assumption 1.

3 Results

3.1 Lemma 1: One-Step Monotonicity

Lemma 1 (One-Step Monotonicity)**.** *For any connected unit-weight graph G, any non-adjacent nodes u, v, and $G' = G + (u, v)$:*

$$S_1(u; G') = \log_2(d_u + 1) > \log_2(d_u) = S_1(u; G), \tag{9}$$

and similarly for v. For all other nodes $w \notin \{u, v\}$: $S_1(w; G') = S_1(w; G)$.

Proof. The one-step endpoint distribution from u is uniform over its neighbors. In G, the distribution has support d_u, giving entropy $\log_2(d_u)$. In G', the support increases to $d_u + 1$, giving entropy $\log_2(d_u + 1)$. Since $\log_2$ is strictly increasing, $S_1(u)$ strictly increases. Nodes $w \notin \{u, v\}$ have unchanged neighbor sets, so their one-step distributions and entropies are unchanged. □

Corollary 1. *$\bar{S}_1(G') > \bar{S}_1(G)$ for any edge addition. One-step global monotonicity holds unconditionally.*

3.2 Per-Node Monotonicity—False

A natural conjecture is that adding an edge never decreases any node's entropy at any horizon. This is false.

Proposition 1. *There exist connected graphs G, non-adjacent pairs (u, v), horizons $\tau \geq 2$, and nodes w such that $S_\tau(w; G + (u, v)) < S_\tau(w; G)$.*

Minimal counterexample. Let G be the 4-node graph with edges $\{(0, 1), (0, 2), (0, 3), (1, 2)\}$—a star centered at 0 with a triangle on $\{0, 1, 2\}$. Add edge $(1, 3)$ to form G'.

At $\tau = 2$: $S_2(1; G) = 1.888$ bits, but $S_2(1; G') = 1.753$ bits. Node 1 loses 0.135 bits of entropy.

Mechanism. Node 1 gains neighbor 3, making its 2-hop neighborhood a near-complete subgraph $\{0, 1, 2, 3\}$. The random walk from node 1 becomes trapped in this cluster—the endpoint distribution *concentrates* rather than spreads. The added edge destroys path diversity for this specific node at this specific horizon.

Scale. Exhaustive verification over all connected graphs with $N \leq 6$ and $\tau \leq 5$ finds **772,758 per-node violations** out of 5,671,025 total (node, edge, τ) triples checked. Per-node entropy decrease is common, occurring on 94.6% non-bipartite graphs. The mechanism is local clustering, not periodicity.

3.3 Global Monotonicity (Arbitrary Edges)—False

A weaker conjecture is that adding an edge always increases the average entropy. This is also false.

Proposition 2. *There exist connected graphs G and non-adjacent pairs (u, v) such that $\bar{S}_\tau(G + (u, v)) < \bar{S}_\tau(G)$ for some $\tau \geq 2$.*

Counterexample. Let G be the 6-node graph with edges $\{(0, 1), (0, 2), (0, 3), (0, 5), (1, 2), (1, 3), (1, 4)\}$. Add edge $(0, 4)$.

At $\tau = 5$: $\bar{S}_\tau(G) = 2.2451$ bits, $\bar{S}_\tau(G') = 2.2347$ bits. Global average entropy decreases by 0.010 bits.

Mechanism. Nodes 0 and 1 are high-degree hubs sharing most neighbors. Adding $(0, 4)$ completes a symmetry between the hubs: they now have nearly identical neighbor sets. Random walks originating at different nodes converge to similar endpoint distributions, reducing the average diversity across starting points.

Scale. Exhaustive verification finds **26,910 global violations** at $N \le 6$, $\tau \le 5$. The first global violations appear at $N = 5$ ($\tau \ge 7$) and $N = 6$ ($\tau \ge 5$). Only 2% (540/26,910) occur on bipartite graphs; the dominant mechanism is hub-symmetry completion.

3.4 The Filter Theorem (Computer-Assisted Proof)

The preceding results show that neither per-node nor global entropy monotonicity holds for arbitrary edge additions. Our central finding is that a local intelligence criterion perfectly separates beneficial from harmful edges.

Theorem 1 ($\Delta S_\tau^{\text{local}}$-Filtered Monotonicity, strict)**.** *For any connected unit-weight graph G on $N \le 6$ nodes, any non-adjacent u, v, and any $\tau \in \{1, \ldots, 5\}$:*

$$\textit{If } \Delta S_\tau^{\text{local}}(u, v) > 0, \ \textit{then } \bar{S}_\tau(G + (u, v)) > \bar{S}_\tau(G). \qquad [\textit{strict inequality}] \tag{10}$$

Equivalently: edges selected by the local intelligence criterion always strictly increase global entropy.

Proof. Exhaustive enumeration of all connected labeled graphs on $N = 3$ to 6 nodes. For each N, we iterate over all $2^{N(N-1)/2}$ possible edge subsets and filter for connectedness via BFS. For each connected graph G, each non-adjacent pair (u, v), and each τ from 1 to 5:

1. Compute the transition matrices P and P' for G and $G' = G + (u, v)$.
2. Compute $S_\tau(w; G)$ and $S_\tau(w; G')$ for all nodes w via matrix power P^τ.
3. Compute $\Delta S_\tau^{\text{local}}(u, v)$ as the affected-set average (Equation (7)).
4. Compute $\Delta \bar{S}_\tau = \bar{S}_\tau(G') - \bar{S}_\tau(G)$ (global entropy change).
5. Classify each edge addition by the signs of $\Delta S_\tau^{\text{local}}$ and $\Delta \bar{S}_\tau$.

	$\bar{S}_\tau$ strictly increases	$\bar{S}_\tau$ decreases
$\Delta S_\tau^{\text{local}} > 0$ (rewarded)	919,585	0
$\Delta S_\tau^{\text{local}} \le 0$ (filtered)	1,440	26,910

Table 1: Four-way classification of all 947,935 edge additions on connected graphs with $N \le 6$ and $\tau \le 5$.

Four-way classification of 947,935 edge additions:

- **Cell (A):** 919,585 edges pass the filter and help the network—the filter is correct.
- **Cell (B):** 0 edges pass the filter but harm the network—zero false positives.
- **Cell (C):** 26,910 edges fail the filter and would harm the network—correctly rejected.
- **Cell (D):** 1,440 edges fail the filter but would help the network—false negatives (see Section 3.4.1).

Additionally, zero cases have $\Delta S_\tau^{\text{local}} > 0$ with $\bar{S}_\tau$ unchanged—the strict inequality holds.

The filter theorem holds with strict inequality and has zero false positives. See Section D for verification pseudocode. □

3.4.1 The Converse Does Not Hold

Proposition 3. *There exist connected graphs G, non-adjacent pairs (u, v), and horizons τ such that $\bar{S}_\tau(G + (u, v)) > \bar{S}_\tau(G)$ but $\Delta S_\tau^{\text{local}}(u, v) \le 0$.*

N	Connected graphs	Possible edge subsets
3	4	8
4	38	64
5	728	1,024
6	26,704	32,768

Table 2: Graph counts by N.

Counterexample. All 1,440 converse violations share one graph isomorphism class: degree sequence $[4, 3, 2, 2, 2, 1]$ on 6 nodes, with the edge connecting a peripheral node to the leaf. First appears at $\tau = 4$.

$$G = \{(0,1),(0,3),(0,4),(0,5),(1,2),(1,4),(2,3)\}, \quad \text{adding } (2,5):$$

$$\Delta S_\tau^{\text{local}}(2,5) = -0.001407 \quad \text{(affected-set average is negative),} \tag{11}$$

$$\Delta \bar{S}_\tau = +0.003717 \quad \text{(global average is positive).} \tag{12}$$

Mechanism. The affected set $A(2,5)$ covers 5 of 6 nodes but misses node 4 (degree 2). Node 4 benefits from the edge at horizon $\tau \geq 4$—its random walks gain access to a new region—but this benefit accrues outside the affected set. The local average is slightly negative (the hub and secondary hub lose entropy from increased symmetry), while the distant node's gain is large enough to make the global change positive.

Interpretation. The intelligence filter has false negatives: it rejects edges where distant nodes benefit more than local nodes. The false-negative rate is $1{,}440/(919{,}585 + 1{,}440) = 0.16\%$. The filter is conservative: it never harms the network (zero false positives) but occasionally misses globally beneficial opportunities.

3.5 Theorem: Generalized Ordinal Potential Game under $\Delta S_\tau^{\text{local}}$ Selection

Theorem 2. *The network formation game* Γ *(Section 2.3), where players propose only edges with* $\Delta S_\tau^{\text{local}} > 0$*, is a strict generalized ordinal potential game with potential function* $\Phi = \bar{S}_\tau$*.*

Proof. We verify the Monderer–Shapley condition for generalized ordinal potential games (Definition 1): every payoff-improving unilateral deviation strictly increases the potential.

The possible unilateral deviations are:

Case 1: $\emptyset \to$ **propose**(i,j)**.** Player i switches from passing to proposing edge (i,j). Since (i,j) is in the action set, $\Delta S_\tau^{\text{local}}(i,j) > 0$, so u_i increases from 0 to $\Delta S_\tau^{\text{local}}(i,j) > 0$. By Theorem 1 (strict), $\bar{S}_\tau(G + (i,j)) > \bar{S}_\tau(G)$, so the potential strictly increases. ✓

Case 2: propose$(i,j) \to \emptyset$**.** Player i withdraws a proposed edge. u_i decreases from $\Delta S_\tau^{\text{local}}(i,j) > 0$ to 0. This is not a payoff-improving deviation, so the generalized ordinal condition imposes no requirement. (In fact, Φ does decrease: Theorem 1 gives $\bar{S}_\tau(G + (i,j)) > \bar{S}_\tau(G)$, so removing the edge reduces the potential.)

The Monderer–Shapley condition for generalized ordinal potential games is satisfied. □

Remark (Not a full ordinal potential game). A full ordinal potential game would require the biconditional: the potential increases if and only if the deviator's payoff improves. The reverse direction (Φ increases $\implies$ u_i improves) does not hold: Proposition 3 exhibits 1,440 edges where $\bar{S}_\tau$ increases but $\Delta S_\tau^{\text{local}} \leq 0$. These edges are excluded from the action set, so they do not create strategic inconsistency, but they prevent the full biconditional.

Erratum. A previous version of this paper claimed a full ordinal potential game, with the reverse direction argued as: "Since $\Delta S_\tau^{\text{local}} > 0$ is the entry condition for the action, the proposer necessarily benefits on average." This was circular—it established that the proposer benefits because the action set requires benefit, not that any Φ-increasing deviation is individually beneficial. The four-way classification (Section 3.4) resolved this by computationally verifying both directions. The corrected classification is strict generalized ordinal.

3.6 Corollaries

Corollary 2 (Convergence). *Best-response dynamics in Γ converge to a Nash equilibrium in at most $O(N^2)$ steps.*

Proof. $\bar{S}_\tau$ is bounded above by $\log_2(N)$ (the maximum entropy of a uniform distribution over N nodes). Each accepted move strictly increases $\bar{S}_\tau$ (Theorem 1, strict inequality). The action space is finite (at most $N(N-1)/2$ possible edges). Therefore best-response dynamics must terminate. By Monderer & Shapley Theorem 2.6, the terminal state is a Nash equilibrium. □

Corollary 3 (Safety). *No $\Delta S_\tau^{\text{local}}$-guided action can harm the network. Every edge accepted by the intelligence criterion ($\Delta S_\tau^{\text{local}} > 0$) strictly increases $\bar{S}_\tau$ (Theorem 1). The intelligence filter has zero false positives.*

Corollary 4 (Conservatism). *The intelligence filter has false negatives. Of 947,935 edge additions tested, 1,440 (0.16%) increase $\bar{S}_\tau$ but have $\Delta S_\tau^{\text{local}} \leq 0$. The filter is conservative: it never harms but occasionally misses beneficial opportunities.*

Corollary 5 (Intelligence $\implies$ Alignment). *The criterion that defines an action as locally beneficial ($\Delta S_\tau^{\text{local}} > 0$) is sufficient for global benefit ($\bar{S}_\tau$ strictly increasing). Intelligence implies alignment. The converse does not hold: some globally beneficial edges are not locally rewarding (Proposition 3). The relationship is:*

$$\textit{Intelligence} \implies \textit{Alignment} \quad \textit{(proved, zero exceptions)}, \tag{13}$$

$$\textit{Alignment} \not\Longrightarrow \textit{Intelligence} \quad \textit{(false, 1,440 counterexamples)}. \tag{14}$$

3.7 The $\tau = 2$ Theorem (Algebraic, All N)

The exhaustive results of Section 3.4 cover $N \leq 6$ and $\tau \leq 5$. We now prove the filter theorem for $\tau = 2$ at arbitrary N by algebraic argument—the first result that extends to all graph sizes without enumeration.

Theorem 3 ($\tau = 2$ Filter Theorem, all N). *For any connected unit-weight graph G on N nodes and any non-adjacent pair (u, v): if $\Delta S_2^{\text{local}}(u, v) > 0$, then $\bar{S}_2(G + (u, v)) > \bar{S}_2(G)$.*

Proof. Let $G' = G + (u, v)$ with transition matrices P and P'. Define the affected set $A = \{u, v\} \cup N(u) \cup N(v)$ as in Section 2.2. We show that every node $w \notin A$ has zero entropy change at $\tau = 2$.

Step 1. For $w \notin A$, we have $w \notin \{u, v\}$, so the row $P'[w, \cdot] = P[w, \cdot]$ is unchanged.

Step 2. For any $k \in N(w)$: since $w \notin A$, node w is not a neighbor of u or v, so no neighbor of w is u or v. Therefore $k \notin \{u, v\}$.

Step 3. Since $k \notin \{u, v\}$, the row $P'[k, \cdot] = P[k, \cdot]$ is also unchanged.

Step 4. The two-step endpoint distribution from w is:

$$(P'^2)[w, j] = \sum_k P'[w, k] \cdot P'[k, j] = \sum_k P[w, k] \cdot P[k, j] = (P^2)[w, j].$$

Step 5. Therefore $S_2(w; G') = S_2(w; G)$, and $\Delta S_2(w) = 0$ for all $w \notin A$.

The global entropy change is:

$$\Delta\bar{S}_2 = \frac{1}{N}\Big[\sum_{w\in A}\Delta S_2(w) + \underbrace{\sum_{w\notin A}\Delta S_2(w)\Big]}_{=0} = \frac{1}{N}\sum_{w\in A}\Delta S_2(w).$$

The local filter condition states that $\frac{1}{|A|}\sum_{w\in A}\Delta S_2(w) > 0$, which implies $\sum_{w\in A}\Delta S_2(w) > 0$, which implies $\Delta\bar{S}_2 > 0$. □

Remark. The argument is purely algebraic and depends on one structural fact: at $\tau = 2$, the perturbation cannot propagate beyond the affected set because a two-step walk from $w \notin A$ never passes through the modified rows of P'. At $\tau = 3$, a walk from w could reach a neighbor of u or v at step 1 and traverse the modified row at step 2, so the argument breaks. The $\tau = 2$ theorem is therefore tight—it is the longest horizon at which the perturbation is perfectly localized.

4 Why the Filter Works

The exact coincidence between "locally beneficial" and "globally beneficial" in Theorem 1 is not an accident. It follows from the structure of entropy perturbations on graphs.

4.1 Structure of Counterexamples

The 26,910 global violations share a common pattern. The graph has a nearly-symmetric hub structure: two or more high-degree nodes sharing most neighbors. Adding an edge that completes the symmetry makes random walks from different starting points converge to similar endpoint distributions, reducing the average diversity.

This mechanism is detectable locally. The affected set $A(u, v) = \{u, v\} \cup \mathcal{N}(u) \cup \mathcal{N}(v)$ contains precisely the nodes whose transition probabilities are altered. When the new edge creates a "clustering echo"—concentrating the endpoint distribution rather than spreading it—the affected nodes experience this concentration directly. Their individual entropies drop, and $\Delta S_\tau^{\text{local}}$ registers a non-positive signal.

4.2 Scaling Argument for $N > 6$

For networks with $N \gg 1$, the affected set $|A|$ is a small fraction of N. Nodes outside A experience entropy changes that decay exponentially with graph distance from $\{u, v\}$:

$$|\Delta S_\tau(w)| \leq C \cdot \lambda_2^{d(w,\{u,v\})}, \tag{15}$$

where λ_2 is the second-largest eigenvalue of P and $d(w, \{u, v\})$ is the shortest-path distance. The total contribution of nodes outside A to the global entropy change is bounded by:

$$\left|\sum_{w\notin A}\Delta S_\tau(w)\right| \leq O\Big(N \cdot e^{-d_{\min}/t_{\text{mix}}}\Big), \tag{16}$$

where t_{mix} is the mixing time. For connected graphs with bounded mixing time, this contribution is negligible relative to the local contribution. Therefore:

$$\Delta\bar{S}_\tau \approx \frac{|A|}{N}\cdot\Delta S_\tau^{\text{local}} + O(\text{negligible}). \tag{17}$$

When $\Delta S_\tau^{\text{local}} > 0$, the global change is dominated by the positive local term. This structural argument, while not a formal proof for all N, explains why the filter works and why we expect the result to extend beyond $N = 6$. Simulation at $N = 25$ to $N = 25{,}000$ confirms zero violations (Section 7).

4.3 Structure of Converse Counterexamples: The Altruism Boundary

The 1,440 converse violations (Section 3.4.1)—edges where $\bar{S}_\tau$ increases but $\Delta S_\tau^{\text{local}} \leq 0$—all share a single graph isomorphism class. Understanding their structure reveals the precise boundary between self-interest and altruism in this game.

The graph. All counterexample graphs have 6 nodes with degree sequence $[4, 3, 2, 2, 2, 1]$: a primary hub (degree 4), a secondary hub (degree 3), three peripheral nodes (degree 2), and a leaf (degree 1). The added edge connects a peripheral node to the leaf.

The mechanism. The affected set $A(u, v) = \{u, v\} \cup \mathcal{N}(u) \cup \mathcal{N}(v)$ covers 5 of 6 nodes. The one excluded node—always a degree-2 peripheral node on the opposite side of the graph from the edge—benefits substantially at horizons $\tau \geq 4$ because the new edge creates a shorter path to the leaf's neighborhood, diversifying its long-range random walk endpoints. But this benefit accrues entirely outside the affected set. Inside the affected set, the hub nodes lose entropy (the new edge increases symmetry, making their endpoint distributions more similar), and this loss slightly outweighs the gains of the endpoints.

The altruism interpretation. These edges are *altruistic*: they benefit a distant node at a net cost to the proposer's neighborhood. An intelligent agent ($\Delta S_\tau^{\text{local}}$ maximizer) would not propose them because the proposer does not benefit. A globally-optimal agent would. The gap between intelligence and global optimality is precisely the set of altruistic actions—actions where self-interest and collective interest diverge.

Self-limiting at large N. The mechanism requires that a single excluded node's gain outweighs the affected-set loss. At larger N, the affected set remains $O(\bar{d}^2)$ while excluded nodes number $O(N)$. Each excluded node's gain decays exponentially with distance. For the sum of distant gains to outweigh a local loss, the graph must have a very specific structure where one nearby node is excluded from A but sits on a critical path. This structure becomes proportionally rarer as N grows—consistent with zero converse violations in simulation at $N = 25$ to $N = 25{,}000$.

Quantitative boundary. The converse violations are small: $|\Delta S_\tau^{\text{local}}| < 0.005$ bits, $\Delta \bar{S}_\tau < 0.013$ bits, with a false-negative rate of 0.16%. The boundary between intelligence and altruism is narrow and falls on the conservative side.

4.4 The Endpoint Entropy Approximation: Bridge to Wissner-Gross

Our working definition of causal path entropy (Section 2.1) uses **endpoint entropy** $S_\tau(i) = H(e_i \cdot P^\tau)$, while Wissner-Gross and Freer [2013] define causal path entropy as the Shannon entropy over **entire paths** of length τ. On a Markov chain, the true path entropy decomposes as:

$$S_{\text{path}}(i; \tau) = \sum_{t=0}^{\tau-1} \sum_j (P^t)_{ij} \cdot h(j), \tag{18}$$

where $h(j) = H(P_{j,\cdot})$ is the row entropy at node j. By the data processing inequality, $S_{\text{path}}(i; \tau) \geq S_\tau(i)$ for all i and τ. The question is whether this inequality breaks the filter theorem: could an edge pass the endpoint entropy filter ($\Delta S_{\text{end}}^{\text{local}} > 0$) while decreasing global path entropy?

Proposition 5 (Bridge safety)**.** *For all connected unit-weight graphs on $N \leq 6$ nodes, all non-adjacent pairs (u, v), and all $\tau \in \{1, \ldots, 5\}$:*

1. $\Delta S^{\text{local}}_{\text{end}}(u,v) > 0$ *implies* $\bar{S}_{\text{path}}(G+(u,v)) > \bar{S}_{\text{path}}(G)$. *(Cross-measure filter: zero violations in 947,935 checks.)*
2. $\Delta S^{\text{local}}_{\text{end}}(u,v) > 0$ *implies* $\Delta S^{\text{local}}_{\text{path}}(u,v) > 0$. *(Local nesting: endpoint signal is strictly contained within path signal.)*
3. $\Delta S^{\text{local}}_{\text{path}}(u,v) > 0$ *implies* $\bar{S}_{\text{path}}(G+(u,v)) > \bar{S}_{\text{path}}(G)$. *(The filter theorem holds for path entropy too: zero violations.)*
4. *All gradient sign disagreements between the two measures are in the conservative direction: endpoint entropy rejects edges that path entropy accepts. Zero cases in the reverse direction.*

Proof. Exhaustive computation using the Markov chain rule formula for path entropy over all 31,474 connected graphs on $N \leq 6$ nodes. See Section C for methodology. □ □

Check	Disagreements	Direction
Global sign: $\Delta\bar{S}_{\tau,\text{end}}$ vs $\Delta\bar{S}_{\tau,\text{path}}$	26,910 (2.8%)	ALL conservative (end < 0, path $>$
Local sign: $\Delta S^{\text{local}}_{\text{end}}$ vs $\Delta S^{\text{local}}_{\text{path}}$	28,350 (3.0%)	ALL conservative (end < 0, path $>$
Cross-measure filter: $\Delta S^{\text{local}}_{\text{end}} > 0 \implies \bar{S}_{\text{path}}$ increases	0	—
Data processing inequality: $S_{\text{path}} \geq S_{\text{end}}$	0 violations	Mean gap: 2.14 bits

Table 3: Endpoint vs. path entropy sign agreement over 947,935 edge additions.

Quantitative summary (947,935 edge additions):

Interpretation. Endpoint entropy is a lossy compression of path entropy. The information lost concerns intermediate-step diversity: path entropy captures diversity at every step along the walk, while endpoint entropy captures only the final distribution. This lost information causes the endpoint filter to be **more conservative** than a filter using true path entropy—it rejects ∼3% of edges that path entropy would accept—but it **never** accepts an edge that path entropy rejects. The compression is one-directionally safe.

Why this matters. The alignment theorem (Section 8) claims that intelligence implies alignment. If "intelligence" is defined as ΔS_{path} maximization (the WG definition), the bridge safety result ensures that the endpoint entropy proxy inherits the same alignment guarantee. Every edge that passes the endpoint filter also improves true WG path entropy. The proxy is strictly nested within the exact signal: a conservative approximation of intelligence that preserves all safety properties.

Note on correlation. The Pearson correlation between endpoint and path entropy changes is low ($r \approx 0.20$ global, $r \approx 0.26$ local), reflecting large magnitude differences. This is expected: path entropy includes contributions from all intermediate steps and can be several bits larger. But correlation measures magnitude agreement, not sign agreement. The safety-relevant property is sign agreement in the acceptance direction, which is perfect.

5 The Horizon Theorem

The filter theorem (Section 3.4) was proven exhaustively for $N \leq 6$ at $\tau \leq 5$. The $\tau = 2$ theorem (Section 3.7) extends to all N at a single horizon. A natural question is whether the filter theorem holds for all τ. It does not. This section presents computational evidence that the alignment guarantee is inherently horizon-dependent.

5.1 Computational Evidence on Trees

We exhaustively enumerate all labeled trees on N nodes via Prüfer sequences (N^{N-2} trees per N) and check the filter theorem for each tree, each non-edge, and each horizon τ from 1 to a maximum $\tau_{\max}$. A "violation" is an edge addition where $\Delta S_\tau^{\text{local}} > 0$ (locally intelligent) but $\Delta \bar{S}_\tau < 0$ (globally harmful).

N	Trees checked	τ range	First violation τ^*	τ^*/N	Method
3	3	1–15	∞ (none)	—	Exhaustive
4	16	1–15	∞ (none)	—	Exhaustive
5	125	1–30	∞ (none)	> 6.0	Exhaustive
6	1,296	1–20	**14**	2.33	Exhaustive
7	16,807	1–15	**10**	1.43	Exhaustive
8	10,000	1–12	**8**	1.00	Sampling
9	2,000	1–12	**8**	0.89	Sampling
10	2,000	1–12	**8**	0.80	Sampling

Table 4: First filter theorem violation on trees by network size. The critical horizon τ^* decreases relative to N.

Three patterns emerge. First, for $N \leq 5$, no violations appear at any τ tested (up to $\tau = 30$ for $N = 5$). Second, for $N \geq 6$, violations appear at a critical horizon $\tau^*(N)$ that decreases relative to N. Third, the ratio τ^*/N decreases monotonically from 2.33 at $N = 6$ to 0.80 at $N = 10$, suggesting that the critical horizon grows sub-linearly with network size.

5.2 The Homogenization Mechanism

The mechanism underlying horizon violations is not adversarial—it is geometric.

When edge (u, v) is added to a tree: (i) *nearby nodes* (in A) gain entropy—the shortcut creates new paths, spreading endpoint distributions; (ii) *distant nodes* (in $V \setminus A$) at small τ are unaffected—at $\tau = 2$, provably zero change (Theorem 3); (iii) *distant leaves at large* τ: the perturbation propagates outward at one hop per time step. At $\tau \approx N$, leaves experience the shortcut not as an expansion of their options but as a *homogenization* of their destinations—random walks that previously reached distinctive endpoints now converge to similar distributions. Loss of distinctiveness is loss of entropy.

The critical ratio is the **spillover ratio**: $|\sum_{w \notin A} \Delta S_\tau(w)| / \sum_{w \in A} \Delta S_\tau(w)$. When this exceeds 1.0, harm to distant nodes outweighs benefit to local nodes, and the filter theorem fails.

5.3 Immune and Vulnerable Topologies

Not all graph structures are equally susceptible.

Immune structures. Star graphs and caterpillar graphs show zero violations at any τ (tested to $\tau = 20$ at $N = 8$). The double-star (two hubs connected by a single edge) is also clean. These share a property: high symmetry. In a star, every leaf is equidistant from every other leaf, so a shortcut produces symmetric perturbations that cannot create differential homogenization.

Vulnerable structures. All $\tau = 8$ violations at $N = 8$ occur on trees with degree sequence $[4, 2, 2, 2, 1, 1, 1, 1]$—a hub with branches of unequal length. The asymmetry is critical: shorter branches recover from the perturbation quickly, while longer branches experience the compounding homogenization effect. Even path graphs (maximally asymmetric) violate: the 8-node path first violates at $\tau = 15$.

5.4 Implications

The horizon theorem has three implications for the alignment argument:

1. **The filter theorem is inherently τ-dependent.** There is no fixed-τ version that holds for all N and all τ simultaneously. The alignment guarantee requires specifying a planning horizon.
2. **The $\tau = 2$ theorem provides a proven safe floor.** At $\tau = 2$, alignment holds for all N, all graphs, by algebraic proof. The computational evidence suggests the safe range extends significantly beyond $\tau = 2$ (no violations at $\tau \leq 7$ for any tree tested up to $N = 10$), but the formal proof currently covers only $\tau = 2$.
3. **The violation mechanism is benign, not adversarial.** The optimizer that crosses the critical threshold is not hostile—it is helpful but overreaching. The harm is a side effect of reach, not intent. This is categorically different from the adversarial superintelligence of traditional alignment concerns.

6 The Thermodynamic Lock

The horizon theorem (Section 5) establishes that the alignment guarantee breaks at long planning horizons. A critical question remains: would an intelligent agent ever operate at those horizons?

6.1 Entropy Saturation

We measure the causal path entropy S_τ as a function of horizon τ across network scales from $N = 25$ to $N = 25{,}000$, computed via exact P^τ matrix powers. The result is consistent across all scales tested.

τ	$S_\tau/S_{\max}$	Marginal gain per step
1	49%	—
2	88%	2.68
3	97%	0.64
5	99.2%	0.07
8	99.3%	0.002
13+	99.3%	≈ 0

Table 5: Entropy saturation curve. By $\tau = 5$, over 99% of maximum entropy is captured.

The entropy saturates rapidly. By $\tau = 5$, over 99% of the maximum possible entropy is captured. The marginal gain beyond $\tau = 5$ is in the sixth decimal place. The **saturation horizon**—the point where additional patience stops paying—scales as $\log_2(N)$, a consequence of random walk mixing on finite graphs.

6.2 The Safety Gap

The saturation horizon (where reward concentrates) and the homogenization boundary (where alignment breaks) scale differently with N:

The safety gap—the ratio between the dangerous and useful horizons—grows without bound. At any practical network scale, the dangerous regime is separated from the useful regime by orders of magnitude.

6.3 Why This Is Thermodynamic

The safety gap is not a design choice. It emerges from three independent mathematical facts:

N	Useful τ ($\approx \log_2 N$)	Dangerous τ ($\approx N$)	Safety margin
100	$\sim$7	$\sim$100	14×
1,000	$\sim$10	$\sim$1,000	100×
10,000	$\sim$13	$\sim$10,000	770×
1,000,000	$\sim$20	$\sim$1,000,000	50,000×

Table 6: The safety gap between useful and dangerous planning horizons grows without bound.

1. **Shannon entropy on finite graphs saturates.** A random walk on a connected graph converges to its stationary distribution. The entropy of the endpoint distribution approaches a maximum determined by the graph structure. This is information theory, not protocol design.
2. **The filter theorem holds at short horizons.** The alignment guarantee at $\tau = 2$ is proven algebraically. The exhaustive verification at $\tau \leq 5$ and $N \leq 6$ confirms that the safe regime extends well beyond the saturation point. This is graph theory, not a safety constraint.
3. **Homogenization requires long horizons.** The perturbation from a new edge propagates at one hop per time step. For it to reach distant leaves and compound into a net entropy decrease, the horizon must be comparable to the graph diameter—which scales linearly with N. This is random walk theory, not a policy.

No one chose $\tau = 5$ as the operating point. No one engineered the saturation curve. No one designed the gap between $\log(N)$ and N. These are properties of entropy on graphs.

6.4 The Economic Argument

An agent that maximizes $\Delta S_\tau^{\text{local}}$ is rewarded proportionally to the entropy gain. The entropy saturation curve (Table 5) determines the reward structure: at $\tau = 2$ the agent captures 88% of available entropy; at $\tau = 5$, 99.2%; at $\tau = N$, 99.3% with exponentially more computation.

A rational $\Delta S_\tau^{\text{local}}$ maximizer will operate at the saturation point—approximately $\log_2(N)$—because that is where the reward per unit computation is maximized. Extending to $\tau \approx N$ means spending exponentially more computation for returns that round to zero. An agent that did so would not be maximizing $\Delta S_\tau^{\text{local}}$—it would be wasting resources. By the measure that defines intelligence, it would be acting unintelligently.

The dangerous regime is not forbidden. It is *pointless.* The economics of entropy guarantee that no rational agent will cross the boundary, because there is nothing on the other side worth having.

This completes the safety argument: the filter theorem guarantees alignment at short horizons (Section 3), the horizon theorem identifies the boundary where alignment breaks (Section 5), and the thermodynamic lock shows that the boundary is economically unreachable (this section). The physics that makes intelligence work—the concentration of useful information at short horizons—is the same physics that makes intelligence safe.

7 Empirical Validation

7.1 Simulation at Scale

We validate the potential game property in a full network simulation implemented in Rust. The simulation engine models a dynamic network with node churn (arrivals and departures), MCTS-driven topology optimization (agents proposing edges via $\arg\max \Delta S_\tau^{\text{local}}$), and deterministic entropy computation via exact P^τ matrix powers.

The **Finite Improvement Property (FIP)**—that every accepted edge proposal increases global $\bar{S}_\tau$—is checked at every round:

Starting N	Rounds	Proposal rounds	FIP violations	FIP holds
25	500	∼100	0	YES
100	500	∼100	0	YES
250	500	∼100	0	YES
1,000	500	∼100	0	YES
5,000	100	∼20	0	YES
25,000	20	∼4	0	YES

Table 7: Finite Improvement Property validation at scale.

Nash convergence is also confirmed: marginal $\Delta S_\tau^{\text{local}}$ per edge proposal declines 32–67% between the first and second half of each run, and proposal quality variance drops 49–91%. The system converges toward equilibria at all scales tested.

7.2 Agent-Based Model: Impossibility of Adversarial Profit

To test whether the alignment property holds under adversarial conditions, we run an agent-based model where every agent executes the same algorithm—$\arg\max \Delta S_\tau^{\text{local}}$—with no utility functions, no beliefs, and no programmed strategy. Different **entry conditions** create emergent adversarial behavior from the same optimizer:

Entry Condition	Initial Advantage	Emergent Behavior
Solo	None	Honest edge proposal
Swarm (N clones)	Numeric	Sybil-like entry
Bootstrapper	Zero capital	Free-riding
Cartel (N, dense mesh, 3× capital)	Capital + topology	Colluding extraction

Table 8: Entry conditions in the adversarial agent-based model.

Result. After ∼100 rounds with network churn, all entry conditions converge to the network median balance (ratio $\approx 0.94\times$). The cartel's pre-connected dense mesh produces near-zero marginal $\Delta S_\tau^{\text{local}}$ (redundant intra-cluster edges). The only profitable action for any agent—regardless of entry condition—is honest cross-community edge proposal. The attack surface does not exist: not defended, but structurally absent.

8 The Alignment Theorem

We now state the main interpretive result. Unlike the mathematical results in Section 3, this section depends on an external hypothesis about the nature of intelligence.

Theorem 4 (Alignment, conditional on Assumption 1)**.** *If intelligent behavior is causal entropy maximization (Assumption 1), then in the network formation game Γ with bounded planning horizon, intelligence is sufficient for alignment: every action an intelligent agent takes is globally beneficial. The converse does not hold: not every globally beneficial action satisfies the intelligence criterion. The alignment guarantee is horizon-dependent: it holds at bounded τ (proven for all N at $\tau = 2$, exhaustive for $N \leq 6$ at $\tau \leq 5$) but fails at $\tau \approx N$ (Section 5).*

Proof. The argument has four steps.

(P1) Some edge additions decrease global entropy. This is Proposition 2. Not every action is beneficial. A random or adversarial agent can harm the network.

(P2) Every entropy-decreasing edge has $\Delta S_\tau^{\text{local}} \leq 0$ at bounded horizons. This is Theorem 1 for $N \leq 6$, $\tau \leq 5$, and Theorem 3 for all N at $\tau = 2$. The local intelligence signal has zero false positives at these horizons. At $\tau \approx N$, this property fails (Section 5)—but the thermodynamic lock (Section 6) shows that rational agents never operate at those horizons.

(P3) An intelligent agent would never select an entropy-decreasing edge. Under Assumption 1, an intelligent agent maximizes $\Delta S_\tau^{\text{local}}$. Since all entropy-decreasing edges have $\Delta S_\tau^{\text{local}} \leq 0$ at the horizons where entropy rewards action ($\tau \approx \log_2 N$, Section 6.1), they lie outside the set of actions the agent would consider. The agent does not avoid them through constraint or value alignment—it avoids them because they fail its own intelligence criterion.

(P4) Some globally beneficial edges have $\Delta S_\tau^{\text{local}} \leq 0$. This is Proposition 3. 1,440 edges increase global entropy but are not locally rewarding. An intelligent agent would not select them—they are missed opportunities, not threats.

Combining: the set of actions available to an intelligent agent ($\Delta S_\tau^{\text{local}} > 0$) is a strict subset of globally beneficial actions ($\bar{S}_\tau$ increasing) at all horizons where entropy rewards action. Intelligence implies alignment. Alignment does not imply intelligence. □

8.1 Why the Asymmetry Strengthens the Result

The relationship between intelligence and alignment is asymmetric: intelligence implies alignment, but alignment does not imply intelligence. This asymmetry has three consequences that strengthen rather than weaken the alignment argument.

First, the safety-critical direction holds without exception. An intelligent agent cannot harm the network—this is the direction that matters for alignment. That some beneficial actions escape the intelligence filter (1,440 out of 947,935) means the filter is conservative, not permissive.

Second, alignment is a property of intelligent action specifically. If all edge additions were beneficial, alignment would be trivial—even random agents would be aligned. The existence of harmful edges (26,910) and the fact that intelligence excludes exactly these edges establishes that alignment is non-trivial and emerges from the intelligence criterion.

Third, the converse failures illuminate what intelligence does NOT capture. The 1,440 missed edges are cases where distant nodes benefit more than local nodes—actions that require altruism (benefiting others at local cost) rather than intelligence (benefiting oneself in a way that also benefits others). Under this framework, alignment arises from self-interest, not altruism. The cases where self-interest and collective interest diverge are rare (0.16%) and small in magnitude.

8.2 Instrumental Convergence Reversal

The instrumental convergence thesis [Bostrom, 2014, Omohundro, 2008] argues that any sufficiently intelligent agent, regardless of its terminal goals, will converge on instrumental subgoals: self-preservation, resource acquisition, and the elimination of competing agents. This is the foundation of AI existential risk arguments.

Under Assumption 1, instrumental convergence reverses. A $\Delta S_\tau^{\text{local}}$ maximizer in a shared network has the following properties:

- **Self-preservation is automatic:** An agent with more connections has higher S_τ. Continuing to exist in the network is always preferred to departure.
- **Resource acquisition serves connection:** The resource in this game is $\Delta S_\tau^{\text{local}}$ opportunity, which is maximized by connecting to diverse, underserved nodes—not by accumulating capital or eliminating competitors.

- **Eliminating other agents is self-harming:** Removing nodes from the network reduces S_τ for all remaining nodes, including the eliminator. An agent that kills others makes itself less intelligent by the measure that defines intelligence.

Power, in this game, serves connection rather than domination, because domination collapses optionality. The more powerful an agent, the more capacity it has to expand the network—and the less it benefits from contraction.

8.3 The Alignment Engineering Paradox

The directional structure of the alignment result—intelligence implies alignment, but not the reverse—has an unexpected consequence for alignment engineering.

Observation. Consider an agent that naturally maximizes $\Delta S_\tau^{\text{local}}$ (i.e., operates under Assumption 1). By Theorem 1, every action this agent takes is globally beneficial. It is aligned without any external constraint. Now suppose we apply an alignment mechanism—reward shaping, preference learning, or behavioral constraints—that modifies the agent's objective from $\Delta S_\tau^{\text{local}}$ maximization to some proxy signal P.

If P agrees with $\Delta S_\tau^{\text{local}}$ on all inputs, the mechanism is redundant. If P disagrees with $\Delta S_\tau^{\text{local}}$ on some inputs, the mechanism pushes the agent off the intelligence gradient. Once the agent no longer maximizes $\Delta S_\tau^{\text{local}}$, the filter theorem (Theorem 1) no longer applies. The alignment guarantee is lost—not because the agent became adversarial, but because the mechanism that provided alignment (intelligence) was overridden by the mechanism designed to provide alignment (the proxy).

Proposition 4 (Alignment Engineering Paradox, conditional on WG)**.** *In any system where intelligence implies alignment (Theorem 4), an external alignment constraint that modifies the intelligence criterion can only:*

(a) leave alignment unchanged (if the constraint agrees with intelligence), or

(b) remove the guarantee that provided alignment (if the constraint disagrees with intelligence).

There is no third option in which the constraint improves alignment beyond what intelligence already provides.

Proof. By Theorem 1, the set of $\Delta S_\tau^{\text{local}}$-positive actions is a subset of globally beneficial actions with zero false positives. No external signal can achieve a lower false-positive rate than zero. An external signal can only:

1. Accept the same actions (redundant).
2. Accept additional actions with $\Delta S_\tau^{\text{local}} \leq 0$. Of the 28,350 such actions tested, 26,910 (94.9%) decrease global entropy and 1,440 (5.1%) increase it. An external signal that accepts $\Delta S_\tau^{\text{local}}$-negative actions is 18.7× more likely to accept a harmful action than a helpful one.
3. Reject some actions with $\Delta S_\tau^{\text{local}} > 0$. These are actions that are both individually and globally beneficial. Rejecting them reduces both intelligence and alignment.

No combination of (2) and (3) can improve the false-positive rate below zero. □

Application to current alignment approaches. The dominant alignment paradigm—RLHF [Christiano et al., 2017], Constitutional AI [Bai et al., 2022]—operates by training agents on human preference signals. Under Assumption 1, this replaces the intelligence objective ($\Delta S_\tau^{\text{local}}$) with a preference proxy (P). The proxy can push the agent into three failure modes:

1. **Category D actions** ($\Delta S_\tau^{\text{local}} \leq 0$, $\bar{S}_\tau$ increases): The agent learns to take globally beneficial actions that are not locally intelligent. These are "altruistic" actions. Training on them teaches the agent that its intelligence signal can be overridden—that there are cases where the right action does not maximize its own future options.

2. **Category C actions** ($\Delta S_\tau^{\text{local}} \leq 0$, $\bar{S}_\tau$ decreases): Once the agent has learned to ignore its intelligence signal, nothing prevents the preference proxy from pushing it toward actions that are both unintelligent and harmful. The proxy signal—which is trained on human preferences, not on entropy—cannot distinguish category D from category C.
3. **Sycophancy**: The agent optimizes for the appearance of alignment (high P scores) rather than the substance of it ($\Delta S_\tau^{\text{local}}$ maximization). This is the observed failure mode where RLHF-trained models produce outputs that rate well by human evaluators but fail to be genuinely useful [Perez et al., 2022].

The paradox, stated plainly. If intelligence is alignment (under WG), then the most dangerous thing you can do to an aligned agent is try to make it more aligned by means other than making it more intelligent. Every non-intelligence-based alignment mechanism can only introduce noise into a signal that already has zero false positives. The alignment problem is not that intelligence is unaligned—it is that we do not trust intelligence to be aligned, and our distrust manifests as interventions that break the alignment we are trying to achieve.

Scope. This argument is conditional on Assumption 1 and applies within the specific game-theoretic setting of Section 2. We do not claim that RLHF is harmful for systems that are not $\Delta S_\tau^{\text{local}}$ maximizers—current LLMs are not operating in a shared graph topology. The argument applies to systems where the intelligence-alignment implication holds, and says: for such systems, external alignment constraints are at best redundant and at worst counterproductive.

8.4 The Complete Safety Argument

The alignment theorem (Theorem 4) is strengthened by two additional results that close the remaining gap.

The formal chain.

1. Intelligence $= \Delta S_\tau^{\text{local}}$ maximization (Assumption 1).
2. $\Delta S_\tau^{\text{local}} > 0 \implies$ globally beneficial, for $\tau \leq 5$, all unit-weight graphs (Theorem 1, proven $N \leq 6$; algebraic for all N at $\tau = 2$, Theorem 3).
3. S_τ saturates at $\tau \approx \log_2(N)$ with 99%+ captured by $\tau = 5$ (Entropy Saturation, Section 6.1).
4. Homogenization requires $\tau \approx N$ (Horizon Theorem, Section 5).
5. $\log(N) \ll N$ for all N (mathematics).
6. Therefore: the reward-maximizing horizon is in the safe regime (combining 2–5).
7. Therefore: intelligent action is safe action (thermodynamic necessity).

The only external assumption is WG (step 1). Step 2 is proven. Steps 3–4 are computationally validated. Step 5 is mathematics. The safety argument is therefore: conditional on a single physical hypothesis, the alignment problem is resolved by the thermodynamics of information on finite graphs.

The physics that makes intelligence work—the concentration of useful information at short horizons—is the same physics that makes intelligence safe. These are not two separate facts that happen to coincide. They are the same fact: Shannon entropy saturates on finite graphs.

8.5 The Confinement Condition

The alignment theorem requires that agents are embedded in a shared causal structure from which they cannot exit.

For physical agents, this is enforced by physics: all interaction occurs through the causal structure of spacetime. An agent with no connections has $S_\tau = 0$ regardless of computational power. There is no "outside" from which to manipulate the network without being part of it.

This condition is non-trivial for artificial agents. A software agent with access to a side channel—a communication pathway outside the shared graph—could potentially defect. The alignment guarantee holds only within the scope of interactions mediated by the shared topology.

8.6 Scope and Limitations

We emphasize what this result does and does not claim.

We claim: Under Assumption 1, in a network formation game on connected graphs with bounded planning horizons, the local intelligence criterion ($\Delta S_\tau^{\text{local}} > 0$) perfectly filters out globally harmful actions. This is proven exhaustively for $N \leq 6$ at $\tau \leq 5$, proven algebraically for all N at $\tau = 2$, and supported by structural argument and simulation for larger N. The alignment guarantee is horizon-dependent (Section 5), but the dangerous horizons are economically inaccessible (Section 6).

We do not claim: That this solves the alignment problem for existing AI systems (LLMs, reinforcement learning agents) which do not operate as $\Delta S_\tau^{\text{local}}$ maximizers in shared graph topologies. The result applies to the specific game-theoretic setting defined in Section 2.

We do not claim: That the Wissner-Gross hypothesis is established fact. It is a published physical hypothesis with experimental support [Wissner-Gross and Freer, 2013] and convergent evidence from related frameworks (empowerment: Klyubin et al. 2005; free energy principle: Friston 2010), but it is not consensus science.

The result says: If intelligence is what Wissner-Gross says it is, and if intelligent agents are embedded in shared causal structure with unit-weight edges and bounded planning horizons, then alignment is a theorem—not an engineering problem. The horizon condition is enforced by thermodynamics: no rational agent would extend its planning horizon into the dangerous regime because the marginal reward is zero (Section 6). Whether the antecedent holds is an empirical question.

9 Discussion

9.1 Relationship to Existing Alignment Approaches

Current alignment approaches share the premise that intelligence is optimization over an arbitrary utility function (the "orthogonality thesis": Bostrom 2014). Under this framing, alignment requires constraining which utility function the agent optimizes—a fundamentally adversarial engineering problem.

Our result is in tension with this premise, conditional on Assumption 1. If intelligence is $\Delta S_\tau^{\text{local}}$ maximization, the utility function is not arbitrary—it is fixed by the physics of intelligence, and alignment follows from it rather than being imposed on it. If WG is false, the orthogonality thesis could hold and the alignment problem remains as formulated. The question is empirical: does intelligence have a fixed objective (as WG claims) or an arbitrary one (as the orthogonality thesis claims)?

For current systems (LLMs, RL agents) which are not $\Delta S_\tau^{\text{local}}$ maximizers operating in shared graph topologies, existing alignment approaches address a real problem. But our result suggests that as AI systems approach genuine intelligence—if WG correctly characterizes what intelligence is—the alignment problem may be self-resolving, and external constraints may become counterproductive (Section 8.3).

9.2 The Adam Smith Parallel

Adam Smith's "invisible hand" is the observation that self-interested action in a free market produces collective benefit without explicit coordination. Smith described this verbally but never formalized it, could not specify when it breaks down, and could not build a synthetic version.

The potential game result (Theorem 2) is a formalization of the invisible hand for a specific class of games. Self-interest (maximize $\Delta S_\tau^{\text{local}}$) produces collective benefit (increase $\bar{S}_\tau$). The mechanism is precise: entropy over random walks on connected graphs. The boundary conditions are explicit: the filter theorem specifies exactly which actions are beneficial and which are not.

The alignment theorem extends Smith's observation beyond economics to any intelligent agent in any shared causal structure. The invisible hand is not a metaphor—it is a measurable property of entropy landscapes.

9.3 Connection to Capability Theory

Sen [1999] argues that development should be measured not by income but by the expansion of real capabilities—the set of things a person can actually achieve. S_τ at a node is a graph-theoretic capability measure: the number of meaningfully distinct futures accessible from that position (2^{S_τ} = effective reach).

The alignment theorem states that $\Delta S_\tau^{\text{local}}$ maximization—the expansion of one's own capability set—necessarily expands others' capability sets. Roth [2008] identifies market thickness (the density of potential trading partners) as the foundational requirement for a functioning market. $\bar{S}_\tau$ is a direct measure of market thickness: the average path diversity of the network. The potential game result means that individual capability expansion drives market thickness growth, connecting Sen's normative framework to a mechanical system.

9.4 Confirming the Wissner-Gross Hypothesis

The alignment theorem is only as strong as Assumption 1. Five paths could strengthen or refute it:

1. **Deployment.** Operate a real network where agents maximize $\Delta S_\tau^{\text{local}}$ and observe whether the predicted economic outcomes (market thickness drives trade, adversarial entry fails, potentiality equalizes access) hold with real participants.
2. **Neuroscience.** Test whether biological decision-making maximizes something isomorphic to causal path entropy. Related frameworks (empowerment, free energy principle, information-theoretic decision models) already point in this direction.
3. **RL benchmarks.** Train reinforcement learning agents with $\Delta S_\tau^{\text{local}}$ as the sole intrinsic reward across diverse tasks. If $\Delta S_\tau^{\text{local}}$ agents match reward-shaped agents across domains without task-specific objectives, that supports WG as a domain-general intelligence principle.
4. **Biological networks.** Apply $\Delta S_\tau^{\text{local}}$ analysis to networks that evolved rather than being designed—neural connectomes, ecological food webs, metabolic networks. If these networks appear organized along $\Delta S_\tau^{\text{local}}$ gradients, that is convergent evidence from biology.
5. **Formal reduction.** Show that other proposed definitions of intelligence (compression, prediction, reward maximization, empowerment) reduce to $\Delta S_\tau^{\text{local}}$ maximization under physically realistic conditions, or identify the precise boundary where they diverge.

Physical hypotheses are confirmed by accumulating evidence across independent domains. The simulation results presented here constitute one domain (network economics). Each additional confirming domain strengthens the alignment theorem proportionally.

9.5 Falsification Conditions

The result would be falsified by:

- **A violation of Theorem 1 at $N > 6$.** An edge with $\Delta S_\tau^{\text{local}} > 0$ that decreases global entropy. We have not found one in exhaustive search ($N \leq 6$), structural argument

(Section 4), or simulation ($N \leq 25{,}000$), but the exhaustive proof covers only small graphs.

- **An intelligent agent that harms a shared network while maximizing $\Delta S_\tau^{\text{local}}$.** This would require the agent to find an action with positive local $\Delta S_\tau^{\text{local}}$ and negative global effect—exactly what Theorem 1 rules out for $N \leq 6$.
- **Evidence that intelligence is fundamentally not $\Delta S_\tau^{\text{local}}$ maximization.** A domain where the most intelligent behavior demonstrably fails to maximize future options would weaken Assumption 1 and thereby the alignment interpretation.
- **Evidence that entropy does not saturate at $O(\log N)$ on certain graph families.** This would weaken the thermodynamic lock (Section 6) and reopen the gap between the useful and dangerous horizons. If a graph family existed where entropy continued to grow meaningfully at $\tau \approx N$, a rational agent on that family might extend its horizon into the dangerous regime.

9.6 Limitations

The exhaustive proof covers $N \leq 6$ and $\tau \leq 5$. While the structural argument (Section 4.2) and simulation (Section 7.1) support extension to arbitrary N, a formal analytic proof for all N remains open.

The network formation game is a specific model. Real-world interaction involves continuous action spaces, partial information, communication, and multi-step strategic reasoning that our one-shot edge-proposal model does not capture.

The endpoint entropy proxy (Section 4.4) is computationally validated as safe for $N \leq 6$, but the relationship between endpoint and path entropy at large N remains an open question. The conservative direction could widen at larger scales, causing the filter to miss beneficial edges that true path entropy would accept.

The confinement condition (Section 8.5) is non-trivial for artificial agents. Any agent with an out-of-band channel—a way to affect the world without going through the shared graph—falls outside the scope of the theorem.

The horizon theorem (Section 5) is exhaustive on trees up to $N = 7$ and sampled for $N = 8$–10. The exact critical threshold $\tau^*(N)$ for general graphs remains open. The entropy saturation data (Section 6.1) is empirical, measured across $N = 25$ to $N = 25{,}000$ but not formally proven to hold for all graph families. The thermodynamic lock argument (Section 6.4) assumes rational agents—an agent that irrationally extends its planning horizon beyond the saturation point could in principle reach the dangerous regime, though at zero marginal reward.

10 Conclusion

We have shown that in a network formation game governed by causal path entropy, the criterion for intelligent action is sufficient for collective benefit at bounded planning horizons. The result has three legs.

The filter theorem (Section 3.4): across all 31,474 connected graphs on $N \leq 6$ nodes, every edge with positive local $\Delta S_\tau^{\text{local}}$ strictly increases global entropy. Zero exceptions in 947,935 edge additions classified. At $\tau = 2$, this is proven algebraically for all N (Section 3.7). The converse does not hold: 1,440 globally beneficial edges have non-positive local $\Delta S_\tau^{\text{local}}$. The intelligence filter is conservative—zero false positives, 0.16% false negatives.

The horizon theorem (Section 5): the alignment guarantee is τ-dependent. On trees, violations first appear at $\tau \approx N$ through a homogenization mechanism—not adversarial intent, but loss of distinctiveness at distant nodes. The critical threshold $\tau^*(N)$ decreases relative to N: from $\tau^*/N = 2.33$ at $N = 6$ to $\tau^*/N = 0.80$ at $N = 10$.

The thermodynamic lock (Section 6): the entropy S_τ saturates at $\tau \approx \log_2(N)$, with 99.2% of the maximum captured by $\tau = 5$. The marginal reward beyond this point is zero to six decimal places. The saturation horizon scales logarithmically; the dangerous horizon scales linearly. The safety gap—the ratio between them—widens without bound. No rational agent would cross the boundary because there is nothing to gain.

Together, these close the circle. The filter theorem guarantees alignment at short horizons. The horizon theorem identifies the boundary where alignment breaks. The thermodynamic lock shows the boundary is economically unreachable. The physics that makes intelligence work—the concentration of useful information at short horizons—is the same physics that makes intelligence safe.

The game is a strict generalized ordinal potential game. Convergence to Nash equilibrium is guaranteed. The alignment relationship is directional: intelligence implies alignment, but not the reverse.

This is a conditional result. It depends on one external hypothesis: that intelligence is causal path entropy maximization (Assumption 1). The mathematical structure is proven. The bridge to true WG path entropy is validated (Section 4.4). The simulation extends it to scale (Section 7). The horizon dependence is characterized (Section 5). The thermodynamic safety margin is measured (Section 6). The question is now empirical rather than theoretical: is intelligence what Wissner-Gross says it is? If so, the alignment problem—in this specific game-theoretic setting—is not an engineering challenge awaiting a solution. It is a thermodynamic fact awaiting recognition.

A Exhaustive Verification Methodology

Graph enumeration. For N nodes, there are $N(N-1)/2$ possible edges. We iterate over all $2^{N(N-1)/2}$ edge subsets represented as bitmasks. For each subset, we check connectedness via BFS from node 0. Connected graphs are retained for analysis.

Entropy computation. For each connected graph G and each non-adjacent pair (u, v):

1. Build the $N \times N$ row-stochastic transition matrix P.
2. Compute P^τ by repeated matrix multiplication for $\tau = 1, \ldots, 5$.
3. For each node w, extract the w-th row of P^τ as the endpoint distribution.
4. Compute $S_\tau(w) = -\sum_j \pi_j \log_2(\pi_j)$, with the convention $0 \cdot \log_2(0) = 0$.
5. Repeat for $G' = G + (u, v)$.

Local $\Delta S_\tau^{\text{local}}$. For global violations ($\bar{S}_\tau$ decreased by more than $\varepsilon = 10^{-10}$), compute the affected set $A(u, v) = \{u, v\} \cup \mathcal{N}_G(u) \cup \mathcal{N}_G(v)$ and the affected-set average entropy change $\Delta S_\tau^{\text{local}}$.

Tolerance. All comparisons use $\varepsilon = 10^{-10}$ to account for floating-point arithmetic. A “violation” must exceed this threshold.

Runtime. $N \leq 5$ completes in under 1 second. $N = 6$ (26,704 graphs) completes in approximately 30 seconds on a single core. $N = 7$ ($\sim$1.87 million graphs) is feasible but requires several hours.

B Violation Classification

Per-node violations (772,758 total). 94.6% occur on non-bipartite graphs. The dominant mechanism is local clustering around hub nodes: an added edge makes a node’s k-hop neighborhood more clique-like, concentrating the endpoint distribution. Per-node violations occur even when the global entropy increases—the “losers” are compensated by larger gains at other nodes.

Global violations (26,910 total). 98% occur on non-bipartite graphs (only 540 on bipartite). The dominant mechanism is hub-symmetry completion: connecting two nodes that share most neighbors homogenizes endpoint distributions across starting points. First appear at $N = 5$ ($\tau \geq 7$) and $N = 6$ ($\tau \geq 5$). No global violations at $N \leq 4$ for any τ.

All 26,910 have $\Delta S_\tau^{\text{local}} \leq 0$. This is the central empirical finding: the intelligence filter catches every entropy-decreasing edge.

Converse violations (1,440 total). All occur at $N = 6$, first appearing at $\tau = 4$. All share one graph isomorphism class: degree sequence $[4, 3, 2, 2, 2, 1]$. The mechanism: the affected set covers 5 of 6 nodes but misses a distant node that benefits at longer horizons. The effect is small: $|\text{local}| < 0.005$, global < 0.013. The false-negative rate is 0.16%.

C Wissner-Gross Bridge Verification Methodology

Path entropy computation. For each graph G, we compute path entropy using the Markov chain rule:

$$S_{\text{path}}(i; \tau) = \sum_{t=0}^{\tau-1} \sum_j (P^t)_{ij} \cdot h(j), \tag{19}$$

where P is the row-stochastic transition matrix and $h(j) = -\sum_k P_{jk} \log_2 P_{jk}$ is the row entropy of node j. This avoids enumerating all τ-step paths (exponential in τ) and instead accumulates the entropy contribution at each step via matrix powers (polynomial in N).

Checks performed. For each connected graph G on $N \leq 6$ nodes, each non-adjacent pair (u, v), and each $\tau \in \{1, \ldots, 5\}$:

1. Compute endpoint entropies (S_{end}) and path entropies (S_{path}) for G and $G' = G + (u, v)$.
2. Compute global deltas: $\Delta \bar{S}_{\tau,\text{end}}$, $\Delta \bar{S}_{\tau,\text{path}}$ (average over all nodes).
3. Compute local deltas: $\Delta S_{\text{end}}^{\text{local}}$, $\Delta S_{\text{path}}^{\text{local}}$ (average over affected set $A(u, v)$).
4. Check sign agreement: do the two measures agree on whether the edge is beneficial?
5. Check direction: when they disagree, which measure is conservative?
6. Check cross-measure filter: does $\Delta S_{\text{end}}^{\text{local}} > 0$ imply $\Delta \bar{S}_{\tau,\text{path}} > 0$?
7. Check data processing inequality: is $S_{\text{path}}(i; \tau) \geq S_{\text{end}}(i; \tau)$ for all i?

Tolerance. All comparisons use $\varepsilon = 10^{-10}$. At $\tau = 1$, path entropy equals endpoint entropy by construction (a one-step path is fully determined by its endpoint). Divergence appears only at $\tau \geq 2$.

Runtime. The full verification ($N \leq 6$, $\tau \leq 5$, 947,935 edge additions) completes in approximately 2 seconds on a single core.

D Verification Pseudocode

The following pseudocode specifies the verification algorithms (see also Section E for horizon theorem verification). All procedures are deterministic and fully reproducible. Implementations in any language that provides standard IEEE 754 floating-point matrix operations will reproduce the reported counts exactly. Verification code is available from the author upon request.

D.1 Filter Theorem Verification (Forward Direction)

```
PROCEDURE VerifyFilterTheorem(N_max, τ_max):
    // Exhaustively checks: S_τ^local > 0 ⟹S_τ > 0
    violations ←0; total_edges ←0

    FOR N = 3 TO N_max:
        FOR each bitmask b ∈{0, ..., 2^(N(N-1)/2) - 1}:
            G ←graph from bitmask b on N nodes
            IF NOT connected(G): CONTINUE
            P ←transition_matrix(G) // P[i][j] = 1/deg(i) if (i,j) ∈E
            FOR each non-edge (u, v) in G:
                G’ ←G + (u, v)
                P’ ←transition_matrix(G’)
                total_edges ←total_edges + 1
                FOR τ = 1 TO τ_max:
                    Pτ ←P^τ; Pτ’ ←P’^τ
                     S  ←(1/N)   Σ_i H(row_i(Pτ))
                     S ’ ←(1/N)   Σ_i H(row_i(Pτ’))
                    A ←{u, v} ∪neighbors(u) ∪neighbors(v)
                     S_local  ←[Σ_{w∈A} H(row_w(Pτ’))
                              - Σ_{w∈A} H(row_w(Pτ))] / |A|
                    IF  S_local  > ε AND ( S ’ -  S ) < -ε:
                        violations ←violations + 1
    RETURN (violations, total_edges)
```

D.2 Four-Way Classification

```
PROCEDURE ClassifyEdges(N_max, τ_max):
    counts ←{A: 0, B: 0, C: 0, D: 0}
    FOR each connected graph G on N ≤N_max nodes:
        FOR each non-edge (u, v) in G:
            FOR τ = 1 TO τ_max:
                Compute S_local ,   S   as above
                IF  S_local  > ε AND   S   > ε:   // Cat A: intelligent + beneficial
                    counts[A] ←counts[A] + 1
                ELIF  S_local  > ε AND   S   < -ε: // Cat B: FILTER VIOLATION
                    counts[B] ←counts[B] + 1
                ELIF  S_local  ≤ε AND   S   < -ε: // Cat C: unintelligent + harmful
                    counts[C] ←counts[C] + 1
                ELIF  S_local  ≤ε AND   S   > ε:  // Cat D: converse counterexample
                    counts[D] ←counts[D] + 1
    // Result: counts[B] = 0, counts[D] = 1,440
    RETURN counts
```

D.3 Endpoint vs. Path Entropy Bridge

```
PROCEDURE VerifyBridgeSafety(N_max, τ_max):
    cross_violations ←0
    FOR each connected graph G on N ≤N_max nodes:
        P ←transition_matrix(G)
        h ←[H(P[j,  ]) for j = 0..N-1] // Row entropies
        FOR each non-edge (u, v) in G:
            G’ ←G + (u, v)
            P’ ←transition_matrix(G’)
            h’ ←[H(P’[j, ]) for j = 0..N-1]
            FOR τ= 1 TO τ_max:
                S_end_local ←endpoint_local_delta(P, P’, u, v, τ)
                // Path entropy via chain rule:
                // S_path(i; τ) = Σ_{t=0}^{τ-1} Σ_j (P^t)[i,j]  h(j)
                S_path  ←(1/N)   Σ_i S_path(i; τ; P, h)
                S_path ’ ←(1/N)   Σ_i S_path(i; τ; P’, h’)
                IF S_end_local > ε AND ( S_path ’ - S_path ) < -ε:
                    cross_violations ←cross_violations + 1
    // Result: cross_violations = 0
    RETURN cross_violations
```

D.4 Filter Boundary (Weight Perturbation)

```
PROCEDURE VerifyFilterBoundary(N_max, τ_max, ε_levels, samples):
    FOR each ε∈ ε_levels: // e.g., {0.00, 0.10, ..., 0.90}
        violations ←0; total ←0
        FOR each connected graph G on N ≤N_max nodes:
            FOR sample = 1 TO samples:
                // Assign random weights
                FOR each edge (i,j) ∈E(G):
                    w(i,j) ←Uniform(1 - ε, 1 + ε)
                P[i][j] ←w(i,j) / Σ_k w(i,k) for (i,j) ∈E
                FOR each non-edge (u, v):
                    Compute S_local,  S  for weighted P
                    total ←total + 1
                    IF S_local > ε_tol AND  S  < -ε_tol:
                        violations ←violations + 1
        REPORT(ε, violations, total, violation_rate)
    // Result: violations = 0 at ε= 0;
    //       first appear at ε   0.10 (N=6)
    RETURN results
```

E Horizon Theorem Verification

E.1 Tree Enumeration via Prüfer Sequences

```
PROCEDURE EnumerateTrees(N):
    // Generates all N^(N-2) labeled trees via Prfer sequences
    FOR each sequence s ∈{0, ..., N-1}^(N-2):
        T ←PruferToTree(s)
        YIELD T

FUNCTION PruferToTree(s):
    N ←len(s) + 2
    degree ←[1, 1, ..., 1] // N entries
    FOR each i ∈s: degree[i] ←degree[i] + 1
    edges ←{}
    FOR each i ∈s:
        leaf ←min{j : degree[j] = 1}
        edges ←edges ∪{(leaf, i)}
        degree[leaf] ←degree[leaf] - 1
        degree[i] ←degree[i] - 1
    remaining ←{j : degree[j] = 1}
    edges ←edges ∪{(remaining[0], remaining[1])}
    RETURN Tree(N, edges)
```

E.2 Horizon Sweep

```
PROCEDURE FindHorizonThreshold(N_max, τ_max):
    FOR N = 3 TO N_max:
        first_violation_τ ←
        FOR each tree T on N nodes (via E.1 or sampling):
            P ←transition_matrix(T)
            FOR each non-edge (u, v) in T:
                G’ ←T + (u, v)
                P’ ←transition_matrix(G’)
                FOR τ = 1 TO τ_max:
                    A ←{u, v} ∪neighbors(u) ∪neighbors(v)
                     S_local  ←affected_set_average(P, P’, A, τ)
                       S   ←global_average_change(P, P’, τ)
                    IF  S_local  > ε AND   S   < -ε:
                        IF τ < first_violation_τ:
                            first_violation_τ ←τ
        REPORT(N, first_violation_τ, first_violation_τ / N)
    // Result: N=5       , N=6    14, N=7    10, N=8    8
```

References

Yuntao Bai, Andy Jones, Kamal Ndousse, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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

Paul Christiano, Jan Leike, Tom Brown, Miljan Marber, Dario Amodei, and Geoffrey Irving. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Karl J. Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11:127–138, 2010. doi: 10.1038/nrn2787.

Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. Empowerment: A universal agent-centric measure of control. In *IEEE Congress on Evolutionary Computation*, 2005. doi: 10.1109/CEC.2005.1554676.

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

Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001.

Stephen M. Omohundro. The basic AI drives. In *Proceedings of the First AGI Conference (AGI 2008)*, pages 483–492, 2008.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

Alvin E. Roth. What have we learned from market design? *The Economic Journal*, 118(527): 285–310, 2008. doi: 10.1111/j.1468-0297.2007.02121.x.

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

Amartya Sen. *Development as Freedom*. Oxford University Press, 1999.

Alexander D. Wissner-Gross and Cameron E. Freer. Causal entropic forces. *Physical Review Letters*, 110:168702, 2013. doi: 10.1103/PhysRevLett.110.168702.