A Structural Causal Model for Goal-Frontier Maximization

Teague Lasser Claude Opus 4.6 GPT 5.4
teague@subseq.io

Abstract

The goal-frontier maximization (GFM) sequence establishes a computable alignment objective with structural safety properties and endogenous anti-monopolar pressure. The preceding papers treat agent evaluation as a statistical signal-processing problem: trust-weighted votes, correlational contraction detection, max-aggregated risk reports. This paper introduces a structural causal model for capability dynamics that upgrades scorpion detection and risk evaluation from statistical to causal identification, and sharpens the remaining agent-evaluation mechanisms, under explicitly stated structural assumptions. Do-calculus contraction attribution sharpens scorpion detection under additive-SCM, causal-sufficiency, and exogenous-orthogonality conditions. Pathway-conditional residuals formalize risk-trust with L^2 convergence guarantees. A log-linear opinion pool replaces max-aggregation without discarding independence information. A probabilistic dependency-risk model sharpens the minimax substrate bound into a per-step expected cost that compounds with growth, gated by an adversarial balance condition. The common thread: a GFM actor needs a causal model of other agents, and the SCM provides the structural backbone that each mechanism draws on.

1 Introduction

Goal-frontier maximization (GFM) proposes alignment through a single objective: maximize the volume of the jointly achievable capability space $\mathrm{vol}_{\mathrm{P}}(G)$ [Lasser, 2026a]. The foundational paper proves structural safety properties (self-balancing against destruction, coercion, and rigidity). Its companion [Lasser, 2026c] constructs a computable capability poset with polynomial-time measure computation and a leverage ranking signal that produces a gradient toward civilization-building infrastructure. The third paper [Lasser, 2026b] extends the framework to multi-step planning through a discounted value function, develops risk-capability identification, and establishes the anti-monopolar property: under a diversity condition, full capability domination is anti-maximizing.

Epistemic status. This paper is implementation machinery. Papers 1–3 establish what is maximized ($\mathrm{vol}_{\mathrm{P}}(G)$), prove structural safety properties (self-balancing, anti-monopolar pressure, structural alignment), and characterize the conditions under which those properties hold [Lasser, 2026a,c,b]. Those structural guarantees do not depend on the correctness of this paper's results: a flaw in the causal attribution rule, the risk-trust dynamics, the SPRT detection test, or the dependency-risk model would be a flaw in the *implementation* of agent-evaluation mechanisms, not in the framework's alignment properties. The framework falls back on the trust and capability machinery of Papers 1–3 whenever this paper's mechanisms degrade—a property formalized through the multi-channel attribution architecture of Section 2.4.

The preceding papers treat every agent-evaluation question as a statistical signal-processing problem. The foundational paper's scorpion detection [Lasser, 2026a] identifies agents whose actions are *correlated* with $\mathrm{vol}_{\mathrm{P}}$-contraction but cannot disentangle their contribution from simultaneous causes. The horizon paper's risk-trust factor ${T^{\mathrm{risk}}}_j$ [Lasser, 2026b] has no formal update dynamics.

Preprint.

The dependency-risk analysis uses minimax bounds that are quantitatively too weak for risk-adjusted planning. In each case, statistical signals suffice for obvious cases (direct harm, clear deception) but fail for the hard cases that matter most: confounded contraction in multi-agent populations, coordinated risk manipulation, and risk claims whose outcomes are never observed.

This paper introduces a structural causal model for capability dynamics and uses it to upgrade scorpion detection and risk evaluation from statistical to causal identification and to sharpen the remaining agent-evaluation mechanisms. The common thread: a GFM agent needs a *causal model* of other agents to make good decisions about their harm (who caused this contraction?), credibility (is this risk claim structurally sound?), and adversarial exposure (what substrate-correlated threats do we face?).

Contributions.

1. A structural causal model for capability dynamics with do-calculus contraction attribution, showing a conditional L^2 convergence advantage over correlational detection under causal sufficiency, additive-SCM structure, and exogenous orthogonality (Section 2).
2. L^2 EWMA dynamics for risk-trust ${T^{\text{risk}}}_j$ over pathway-conditional structural verification residuals, and a prior-corrected trust-weighted log-linear opinion pool that recovers Bayesian updating in the no-tempering limit, replacing the max aggregation operator, with a two-gate architecture separating attention allocation from action decisions via the actor's own forward causal risk evaluation (Section 3).
3. Causal scorpion detection under a Gaussian model, committed to a sequential probability ratio test on least-favorable simple hypotheses with Type I/II error control, and a high-probability bound on non-stationary scorpion evasion bandwidth under a stationary-epoch model (Section 4).
4. A probabilistic dependency-risk model gated by an adversarial balance condition $c > \max_i(f_i \cdot c_i)$ that admits contagion dynamics, with per-step expected costs that compound at the same rate as growth and a quantitative substrate-count lower bound $m > c_{\max}/c$ that sharpens the minimax $m \geq 2$ (Section 5).

2 Causal Attribution Framework

The foundational paper's scorpion detection (Proposition 2 of Lasser [2026a]) identifies agents whose actions are *correlated* with vol_P-contraction. It explicitly acknowledges that "disentangling the agent's contribution from simultaneous causes requires a causal-attribution procedure that the framework does not provide." This section provides that procedure.

2.1 Structural Causal Model for Capability Dynamics

Definition 1 (Capability Dynamics SCM)**.** *A* structural causal model for capability dynamics *is a tuple* $\mathcal{M} = (U, V, F, P(U))$ *where:*

- U*: exogenous variables representing agent dispositions, environmental conditions, and stochastic factors not determined by the model.*
- $V = \{G_t, \pi_t^{(1)}, \ldots, \pi_t^{(n)}, \Delta\text{vol}_\text{P}{}_t\}$*: endogenous variables representing the capability poset state, each agent's action at time* t*, and the resulting* vol_P*-change.*
- F*: structural equations mapping parent variables to children in the causal directed acyclic graph (DAG). The DAG encodes which agents' actions causally influence which capability changes.*
- $P(U)$*: prior distribution over exogenous variables.*

When the variance-reduction results of this paper are invoked (Proposition 1(c) and downstream), we additionally assume ***exogenous orthogonality****: the exogenous noise components entering different agents' structural equations, and the residual* h*, are mutually independent conditional on* G_t*. That is, if* $U_j \subseteq U$ *are the exogenous variables appearing in* a_j*'s structural equation,* $U_k \subseteq U$

those appearing in a_k's, and $U_h \subseteq U$ those appearing in the residual $h(G_t, U)$ (Equation (2)), then $U_j \perp\!\!\!\perp U_k \mid G_t$ for $j \neq k$ and $U_j \perp\!\!\!\perp U_h \mid G_t$ for all j. This rules out shared exogenous shocks that would induce covariance between agents' contributions (or between an agent's contribution and the environment residual) even under a correct causal DAG.

The GFM actor's world model $\mathcal{W}$ (Definition 1 of Lasser [2026b]) is an instance of $\mathcal{M}$ restricted to the actor's beliefs about causal structure.

The SCM extends $\mathcal{W}$ by making the causal structure explicit. In the foundational paper, $\mathcal{W}$ is an opaque prediction model: given actions, it predicts $\mathrm{vol_P}$-changes. The SCM decomposes predictions into causal pathways, enabling counterfactual queries: "what would have happened if agent a_j had acted differently?"

Temporal scope. Definition 1 specifies a *single-step* SCM: the endogenous variables $\{G_t, \pi_t^{(1)}, \ldots, \pi_t^{(n)}, \Delta\mathrm{vol_P}_t\}$ describe one time slice, and the structural equations F map actions at step t to the $\mathrm{vol_P}$-change at step t. The do-intervention $\mathrm{do}(\pi_j = \cdot)$ operates within this single-step structure: "other variables respond according to F" means other endogenous variables in the same time slice are recomputed under the intervened structural equations, not that the system evolves forward in time. Multi-step effects (an action at t that causes a cascade at $t+1$) are captured by instantiating a new single-step SCM at each subsequent time step with an updated poset state G_{t+1}; the causal attribution $\mathrm{CA}(a_j, t)$ (Definition 2) is the single-step effect.

2.2 Causal Contraction Attribution

Definition 2 (Causal Contraction Attribution)**.** *For an observed $\mathrm{vol_P}$-change $\Delta\mathrm{vol_P}(G_t)$ at time t, the* causal contribution *of agent a_j is:*

$$\mathrm{CA}(a_j, t) = \Delta\mathrm{vol_P}\big(G_t \mid \mathrm{do}(\pi_j = \pi_j^{\mathrm{obs}})\big) - \Delta\mathrm{vol_P}\big(G_t \mid \mathrm{do}(\pi_j = \mathrm{skip})\big) \tag{1}$$

where $\mathrm{do}(\pi_j = \cdot)$ denotes the do-calculus intervention [Pearl, 2009]: setting a_j's action to a specific value while allowing all other variables to respond according to F. The first term is the $\mathrm{vol_P}$-change under a_j's observed action; the second is the counterfactual $\mathrm{vol_P}$-change had a_j done nothing.

The causal attribution $\mathrm{CA}(a_j, t)$ isolates a_j's contribution from confounders: simultaneous actions by other agents, environmental changes, and background dynamics that the correlational signal in Proposition 2 of Lasser [2026a] cannot separate. When a_j's actions cause contraction, $\mathrm{CA}(a_j, t) < 0$. When they cause expansion, $\mathrm{CA}(a_j, t) > 0$. When a_j's actions have no causal effect (the contraction would have occurred regardless), $\mathrm{CA}(a_j, t) = 0$, and the correlational detection would have produced a false positive.

Additivity. The one-at-a-time $\mathrm{CA}(a_j, t)$ of Equation (1) sums to the total $\mathrm{vol_P}$-change only when the SCM's structural equation for $\Delta\mathrm{vol_P}$ is *additive in agent actions*: there exist functions g_j and a residual h such that

$$\Delta\mathrm{vol_P}(G_t) = \sum_j g_j\big(\pi_j^{\mathrm{obs}}, G_t, U\big) + h(G_t, U) \tag{2}$$

where h does not depend on any π_j. Under this additive-SCM assumption, the one-at-a-time intervention $\mathrm{do}(\pi_j = \pi_j^{\mathrm{obs}})$ vs. $\mathrm{do}(\pi_j = \mathrm{skip})$ isolates g_j exactly, and

$$\Delta\mathrm{vol_P}(G_t) = \sum_j \mathrm{CA}(a_j, t) + h(G_t, U). \tag{3}$$

The additive-SCM assumption is stronger than (and is not implied by) conditional independence of agent actions given G_t: two independently drawn actions can still interact multiplicatively in a non-additive structural equation such as $\Delta\mathrm{vol_P} = \pi_1 \cdot \pi_2$, in which case the one-at-a-time CA does not sum to the total. When interaction effects are present and a non-additive SCM is required, an exact decomposition needs a Shapley-style attribution rule that averages CA over all orderings of agent inclusion [Shapley, 1953]; the one-at-a-time form is then an approximation whose error is the deviation from additivity. For scorpion detection (Section 4), the one-at-a-time form is sufficient because the detection signal depends only on the sign and magnitude of the diagonal contribution g_j, not on the exact decomposition of interaction effects.

2.3 Causal Detection Dominates Correlational Detection

Proposition 1 (Convergence Advantage of Causal Attribution)**.** *Let a_j be an agent with expected causal contraction $\mu_j = \mathbb{E}[\mathrm{CA}(a_j, t)] < 0$ (a scorpion). Assume the observation sequences for both detectors are i.i.d. within a stationary strategy epoch, or weakly dependent with summable autocovariance. Under the SCM $\mathcal{M}$, the following* information-theoretic *sample-complexity comparison holds (independent of the specific sequential test used; see Section 4 for the SPRT-specific operational rates):*

- ***(a)*** *The correlational detector (Proposition 2(a) of Lasser [2026a]) has sample complexity $\Omega(\sigma^2/\mu_j^2)$ for identifying a_j as a net-contractor, under the idealized assumption that μ_j appears as an isolable drift in the trust-weighted aggregate $\hat{\Delta}$ (see paragraph below on the correlational estimator). Here σ^2 is the total variance of the signal the correlational detector can construct, including confounders.*
- ***(b)*** *The causal detector using $\mathrm{CA}(a_j, t)$ has sample complexity $\Omega(\sigma_{\mathrm{do}}^2/\mu_j^2)$, where σ_{do}^2 is the variance of the causal attribution signal after conditioning on the do-intervention. The ratio $\sigma_{\mathrm{do}}^2/\sigma^2 \leq 1$ (under the conditions of part (c)) is the structural advantage factor.*
- ***(c)*** *Under **causal sufficiency** (the SCM $\mathcal{M}$ captures all common causes of a_j's actions and the confounder signals), the **additive-SCM** assumption (Equation (2)), and **exogenous orthogonality** (Definition 1), we have $\mathrm{Cov}(\mathrm{CA}(a_j), \mathrm{confounders} \mid G_t) = 0$ and therefore*

$$\sigma^2 - \sigma_{\mathrm{do}}^2 = \mathrm{Var}(\textit{confounders} \mid G_t) \geq 0,$$

so $\sigma_{\mathrm{do}}^2 \leq \sigma^2$, with strict inequality whenever confounders contribute nonzero variance. Without causal sufficiency (or without the additive-SCM structure), the variance identity gives

$$\sigma^2 - \sigma_{\mathrm{do}}^2 = \mathrm{Var}(\textit{confounders} \mid G_t) + 2\,\mathrm{Cov}(\mathrm{CA}(a_j), \mathrm{confounders} \mid G_t),$$

whose sign depends on the covariance term. The inequality $\sigma_{\mathrm{do}}^2 \leq \sigma^2$ therefore requires the full triple (causal sufficiency, additive SCM, exogenous orthogonality), or the weaker condition $\mathrm{Cov}(\mathrm{CA}(a_j), \mathrm{confounders} \mid G_t) \geq -\frac{1}{2}\mathrm{Var}(\textit{confounders} \mid G_t)$. Under a general (non-sufficient) SCM with adversarial confounder correlation, causal attribution can in principle achieve worse convergence than the correlational detector; characterizing populations in which this occurs is outside the scope of this paper.

Proof. (a) The correlational detector's observable signal is the trust-weighted aggregate $\hat{\Delta}(\pi_t) = \sum_{k \in \mathcal{O}} T_k \cdot R_k(\pi_t)$ (Equation 10 of Lasser [2026a]), where $R_k \in \{-1, 0, +1\}$ is a ternary directional vote. Let $\Delta\mathrm{vol}_{\mathrm{P}\,\mathrm{total}}(t)$ denote the *true* continuous $\mathrm{vol}_{\mathrm{P}}$-change at step t that $\hat{\Delta}$ estimates. The correlational detector's sample complexity for detecting a_j's mean effect μ_j is governed by the variance of the continuous quantity it must recover, not by the variance of the ternary estimator itself: a sign-correct ternary estimator inherits the variance of the underlying continuous signal it tracks, plus discretization noise. Formally, σ^2 in the bound is the variance of any estimator of the total $\mathrm{vol}_{\mathrm{P}}$-change that the correlational detector can construct from its observation channel. The $\Omega(\sigma^2/\mu_j^2)$ lower bound is the information-theoretic sample complexity for detecting a mean shift μ_j in noise of variance σ^2.

(b) The causal detector evaluates $\mathrm{CA}(a_j, t)$ directly by comparing the observed $\mathrm{vol}_{\mathrm{P}}$-change against the counterfactual under $\mathrm{do}(\pi_j = \mathrm{skip})$. The variance of this signal is $\sigma_{\mathrm{do}}^2 = \mathrm{Var}(\mathrm{CA}(a_j, t))$, which excludes the confounding variance.

(c) The three assumptions do distinct work; we separate them.

Step 1: Additive SCM $\Rightarrow \mathrm{CA}(a_j) = g_j$ exactly. Under the additive-SCM assumption (Equation (2)), the continuous $\mathrm{vol}_{\mathrm{P}}$-change decomposes as $\Delta\mathrm{vol}_{\mathrm{P}\,\mathrm{total}}(t) = \mathrm{CA}(a_j, t) + \mathrm{confounders}(t)$, where the confounder term aggregates all other agents' contributions $\{g_k\}_{k \neq j}$ and the residual $h(G_t, U)$, and $\mathrm{CA}(a_j, t) = g_j(\pi_j^{\mathrm{obs}}, G_t, U) - g_j(\mathrm{skip}, G_t, U)$ is exactly a_j's separable contribution. Without additivity, $\mathrm{CA}(a_j)$ captures only the diagonal term and misses interaction effects. By the variance identity for sums:

$$\sigma^2 = \mathrm{Var}(\mathrm{CA}(a_j)) + \mathrm{Var}(\mathrm{confounders}) + 2\,\mathrm{Cov}(\mathrm{CA}(a_j), \mathrm{confounders}).$$

Step 2: Causal sufficiency ⇒ confounders are SCM-known. Causal sufficiency ensures the DAG $\mathcal{M}$ captures all common causes of a_j's actions and the confounder signals. Every variable contributing to the covariance term appears explicitly in the structural equations, so the covariance can in principle be computed from the DAG. Without causal sufficiency, latent common causes contribute unmodeled covariance that may have either sign.

Step 3: Exogenous orthogonality ⇒ Cov *term vanishes.* Under exogenous orthogonality (Definition 1), the noise U_j driving g_j is independent of both the noise $\{U_k\}_{k\neq j}$ driving $\{g_k\}_{k\neq j}$ and the noise U_h driving the residual h, conditional on G_t. Since the confounders term is $\sum_{k\neq j} g_k + h$, each cross-covariance $\mathrm{Cov}(g_j, g_k \mid G_t) = 0$ and $\mathrm{Cov}(g_j, h \mid G_t) = 0$. This gives $\mathrm{Cov}(\mathrm{CA}(a_j), \text{confounders} \mid G_t) = 0$, and therefore $\sigma^2 - \sigma^2_{\text{do}} = \mathrm{Var}(\text{confounders}) \geq 0$ as stated. Without exogenous orthogonality, the covariance term is retained and the inequality $\sigma^2_{\text{do}} \leq \sigma^2$ holds only when $\mathrm{Var}(\text{confounders}) + 2\,\mathrm{Cov}(\mathrm{CA}(a_j), \text{confounders}) \geq 0$; the inequality is strict whenever confounders have positive variance and the covariance term does not dominate. ■

The correlational detector as an estimator. Part (a) of Proposition 1 treats the correlational detector as if μ_j appeared directly as a component of the mean of the aggregate signal $\hat{\Delta}$. This is an idealization: the actual mean of $\hat{\Delta}$ is the sum of all agents' mean effects plus background dynamics, not μ_j alone. A correlational estimator that recovered μ_j from $\hat{\Delta}$ would need an additional auxiliary procedure, such as regressing $\hat{\Delta}$ against an indicator for a_j's action times, which incurs its own bias when confounders are correlated with a_j's action schedule. The $O(\sigma^2/\mu_j^2)$ bound therefore describes the best-case correlational detector: one that has access to a_j's action schedule and uses it as a conditioning variable. A weaker correlational detector that only observes $\hat{\Delta}$ without conditioning has a strictly worse (and generally incomputable) rate. The causal detector does not need this auxiliary procedure because $\mathrm{CA}(a_j, t)$ already isolates a_j's contribution by construction.

Practical implication. In populations with many simultaneously acting agents, confounding variance can dominate the correlational signal. A single scorpion among 100 simultaneously acting agents contributes $\sim$1% of the signal variance; the correlational detector needs $\sim$100$\times$ more observations than the causal detector to achieve the same confidence. The causal detector's advantage grows linearly with the number of simultaneously acting confounders.

2.4 Causal Attribution as One Channel Among Several

Proposition 1 assumes that the SCM $\mathcal{M}$ is accurate enough to support do-calculus attribution, and that the structural assumptions of part (c) (causal sufficiency and additive-SCM structure) hold. In practice the SCM is learned from observational data, partially specified, and adversarially exposed: agents can strategically align their actions to mask causal structure, and new agents or delegation chains can emerge during the trajectory. A framework that depended on the SCM being perfectly learned would be brittle. Instead, the framework is robust because causal attribution is *one channel* among several in a multi-channel attribution system, and the system's overall attribution accuracy degrades gracefully as any single channel's reliability degrades.

Multi-channel attribution is the standard methodology in adversarial intrusion analysis and intelligence tradecraft. The Diamond Model of intrusion analysis [Caltagirone et al., 2013] organizes attribution around four analytic vertices (adversary, capability, infrastructure, victim) linked by relationships, with confidence coming from cross-vertex correlation rather than from any single vertex. The *Pyramid of Pain* [Bianco, 2013] ranks observable signals by how costly they are for an adversary to change: hash values, IP addresses, and domain names are cheap to rotate; tools and tactics, techniques, and procedures (TTPs) are expensive to retrain. The MITRE ATT&CK framework [Strom et al., 2018] provides a shared capability taxonomy that grounds TTP-level attribution in a common vocabulary. The *Analysis of Competing Hypotheses* [Heuer, 1999] provides a structured method for comparing attribution hypotheses by diagnostic evidence, weighting signals by their power to discriminate among candidates rather than by their individual strength. In all four traditions, the robustness of an attribution decision is driven by signals the adversary cannot cheaply fake, and no single signal is load-bearing.

The GFM sequence already supplies several attribution channels, distributed across papers:

- **Behavioral prediction residual.** The trust model of Lasser [2026a] computes $r_k(t) = R_k(t) - \hat{R}_k(t \mid M_k(G))$, the divergence between an agent's observed behavior and its predictive model. An agent whose actions consistently match its model concentrates on high T_k; systematic divergence drives T_k down. This is a moderately expensive channel to fake: a scorpion must maintain behavioral consistency with a fictitious model to evade it.
- **Correlational co-occurrence.** The trust-weighted $\hat{\Delta}$ signal of Lasser [2026a] (Proposition 2(a)) detects agents whose actions temporally co-occur with vol_P-contraction. This is a cheap channel to fake (an adversary can introduce uncorrelated dummy actions) and is strictly dominated by the causal channel under the conditions of Proposition 1(c).
- **Capability gating.** Papers 2 and 3 of the GFM sequence [Lasser, 2026c,b] together provide a structural constraint: an agent cannot cause the contraction of a capability it does not have upstream access to, regardless of any other signal. This is the *means-analysis* from the intrusion-analysis tradition and sits at the top of the robustness hierarchy: it reflects the structure of the poset itself rather than behavior an adversary can stage.[1] An attribution decision that conflicts with capability gating should be overridden by it, not aggregated with it.
- **Causal contraction attribution** (this paper, Definition 2). Moderately expensive to fake, since the SCM demands counterfactual consistency, and strictly informative under the conditions of Proposition 1(c). Vulnerable to SCM misspecification.
- **SPRT scorpion detection** (this paper, Section 4). Moderately expensive to fake under stationary epochs; vulnerable to fast adaptation as characterized by the evasion bandwidth bound.

Table 1 summarizes the channels by evasion cost and required assumptions.

Channel	Cost to fake	Paper	Key assumptions
Capability gating	Very high	2, 3	Trusted, complete capability graph
Causal attribution	Moderate	4	Causal sufficiency, additive SCM, exog. orthogonality
Behavioral residual	Moderate	1	Predictive-model access
SPRT detection	Moderate	4	Gaussian, stationary epochs
Correlational co-occ.	Low	1	(None beyond observation)

Table 1: Attribution channels ordered by adversarial evasion cost (highest first), following the Pyramid of Pain principle.

An attribution decision should aggregate evidence across channels, weighted by each channel's calibrated reliability, rather than rely on the output of any single channel. (The aggregation is Bayesian in spirit: each channel contributes evidence toward or against a suspect. This paper does not formalize the joint hypothesis space or channel-dependence model; the full fusion rule is an open direction identified below.) The multi-channel framing has three architectural consequences:

1. **Capability gating is a hard filter, not a prior.** An agent that cannot structurally reach capability c_i cannot have caused its contraction, and no amount of causal or correlational evidence should override that. The correct aggregation is a gate-then-aggregate architecture: apply the capability gate first to produce the candidate suspect set, then aggregate the remaining channels over the candidates.
2. **No single channel must be perfect.** Each channel has its own reliability, and the aggregate's accuracy depends on the joint reliability profile rather than on the maximum-reliability channel. This is the property that makes multi-channel attribution robust in the threat-intelligence setting and that makes the GFM actor robust to imperfect structure learning (Section 6).
3. **Channel weights should be calibrated from measurable proxies**, not fixed at design time. The causal channel's weight should scale with an SCM-confidence estimate derived

[1] The gate's strength is bounded by the fidelity of the maintained capability model: hidden delegation, latent substitute capabilities, or a misspecified poset can defeat it.

from structure-learning diagnostics; the behavioral channel's weight should scale with a population-level reliability measure derived from the per-agent trust scores T_k of Lasser [2026a] (distinct from the actor's self-trust T_s, which governs learning rate rather than channel reliability); the correlational channel's weight should scale with the confounding variance estimated from the population structure. The calibration procedure is outside the scope of this paper and is identified as an open direction in Section 6.

Paper 4's contribution to this framework is the causal channel: a principled, do-calculus-based attribution signal with a convergence advantage over the correlational baseline under stated structural conditions. It is not the attribution system, and the paper does not claim that the causal channel is load-bearing. The overall attribution robustness comes from the multi-channel structure, and from capability gating in particular as the structural floor.

3 Risk-Trust Dynamics

The companion paper [Lasser, 2026b] introduces a risk-trust factor ${T^{\text{risk}}}_j$ for weighting communicated risk claims but provides only a qualitative update rule. This section formalizes the dynamics, drawing on the EWMA structure of the foundational paper's behavioral trust [Lasser, 2026a].

3.1 Structural Verification Residuals

Definition 3 (Structural Verification Residual)**.** *For a risk claim $(S, \mathcal{P}, |\Delta\text{vol}_\text{P}|, p)$ from agent a_j (Definition 4 of Lasser [2026b]), the actor evaluates* the specific claimed pathway $\mathcal{P}$ *against its own SCM $\mathcal{M}$ (Definition 1) using the causal risk evaluation procedure (Section 4.3 of Lasser [2026b]). This produces a pathway-conditional assessment $\big(p_{\mathcal{W}}(\mathcal{P}),\ |\Delta\text{vol}_{\text{P}\mathcal{W}}(\mathcal{P})|\big)$: the probability and contraction magnitude the actor would assign if $\mathcal{P}$ were the realized causal pathway. If $\mathcal{P}$ has no counterpart in $\mathcal{M}$ — that is, if the actor's causal model contains no edges implementing the claimed mechanism — then $\big(p_{\mathcal{W}}(\mathcal{P}), |\Delta\text{vol}_{\text{P}\mathcal{W}}(\mathcal{P})|\big) = (0, 0)$ by convention. The* structural verification residual *is the dimensionless two-component divergence:*

$$r_j^{\text{risk}}(t) = \left(p_j^{\text{claimed}} - p_{\mathcal{W}}(\mathcal{P}),\ \ \frac{|\Delta\text{vol}_\text{P}{}_j^{\text{claimed}}| - |\Delta\text{vol}_{\text{P}\mathcal{W}}(\mathcal{P})|}{\text{vol}_\text{P}(G_t)} \right) \tag{4}$$

Each component is in $[-1, 1]$: the first is a probability difference, the second is a relative volume difference normalized by the current total poset volume. The Euclidean norm $\|r_j^{\text{risk}}(t)\|$ is therefore dimensionally consistent and invariant under rescaling of vol_P*. A fabricated pathway that does not exist in $\mathcal{M}$ produces the maximal residual $\|r_j^{\text{risk}}(t)\| = \sqrt{(p_j^{\text{claimed}})^2 + (|\Delta\text{vol}_\text{P}{}_j^{\text{claimed}}|/\text{vol}_\text{P}(G_t))^2}$; structural verification is enforced by the pathway-conditional evaluation, not by an extra pathway-comparison term.*

Unlike the behavioral prediction residual $r_k(t)$ (Equation 25 of Lasser [2026a]), which compares *predicted* against *observed* behavior, the risk residual compares *claimed* against *independently modeled* risk. This is necessary because risk claims are counterfactual predictions that may never be directly observed — the *Wamura problem*, introduced in the poset paper [Lasser, 2026c] and developed in the risk-communication analysis of Lasser [2026b, §4]: a successful risk intervention removes the very evidence that would have justified it, so the convergence timescale for risk-trust can in principle exceed the timescale on which trust normally accumulates. The SCM provides the independent assessment needed to break this circularity: the actor's own causal model produces $(p_{\mathcal{W}}, |\Delta\text{vol}_{\text{P}\mathcal{W}}|)$ without relying on the claim.

3.2 EWMA Risk-Trust Update

Definition 4 (Risk-Trust Dynamics)**.** *The risk-trust factor ${T^{\text{risk}}}_j$ evolves through an exponentially weighted moving average of squared structural verification residuals:*

$$\sigma_j^{\text{risk},2}(t) = \alpha_{\text{risk}} \cdot \sigma_j^{\text{risk},2}(t-1) + (1 - \alpha_{\text{risk}}) \cdot \|r_j^{\text{risk}}(t)\|^2 \tag{5}$$

$${T^{\text{risk}}}_j(t) = \frac{1}{1 + \beta_{\text{risk}} \cdot \sigma_j^{\text{risk},2}(t)} \tag{6}$$

where $\alpha_{\text{risk}} \in (0,1)$ controls the decay rate (tracking speed vs. stability), $\beta_{\text{risk}} > 0$ controls the sensitivity (how rapidly trust degrades with inconsistency), and $\|\cdot\|$ is the Euclidean norm on the two-dimensional residual.

The structure mirrors the foundational paper's behavioral trust exactly (Equations 33 and 38 of Lasser [2026a]), with the behavioral prediction residual replaced by the structural verification residual. The parameters α_{risk} and β_{risk} are independent of the behavioral parameters α and β: an agent can have high behavioral trust T_k (predictable behavior) and low risk-trust ${T^{\text{risk}}}_k$ (unreliable risk assessments), or vice versa. The two trust channels operate independently.

Initialization and cooling period. New agents enter with $\sigma_j^{\text{risk},2}(0) = \sigma_0^{\text{risk},2}$ (a prior reflecting baseline uncertainty about risk-claim quality), producing initial risk-trust ${T^{\text{risk}}}_j(0) = 1/(1+\beta_{\text{risk}} \cdot \sigma_0^{\text{risk},2})$. The effective cooling period is $\tau_{\text{risk}} = \lceil \log\delta / \log\alpha_{\text{risk}} \rceil$ (same formula as the behavioral cooling period, Equation 71 of Lasser [2026a]).

Proposition 2 (Risk-Trust L^2 Convergence)**.** *Let $X_j(t) = \|r_j^{\text{risk}}(t)\|^2$ and assume a_j's risk-claim strategy is stationary with $\mu_j^{\text{risk}} = \mathbb{E}[X_j] < \infty$ and $\nu_j^{\text{risk}} = \text{Var}[X_j] < \infty$. Then the cumulative inconsistency estimator satisfies:*

$$\lim_{t\to\infty} \mathbb{E}\big[\sigma_j^{\text{risk},2}(t)\big] = \mu_j^{\text{risk}} \qquad \textit{(asymptotic unbiasedness)} \tag{7}$$

$$\limsup_{t\to\infty} \text{Var}\big[\sigma_j^{\text{risk},2}(t)\big] \le \frac{1-\alpha_{\text{risk}}}{1+\alpha_{\text{risk}}} \cdot C_{\text{dep}} \cdot \nu_j^{\text{risk}} \qquad \textit{(bounded stationary variance)} \tag{8}$$

where $C_{\text{dep}} = 1$ under i.i.d. squared residuals, and $C_{\text{dep}}(\alpha_{\text{risk}}) = 1 + 2\sum_{h=1}^{\infty} \alpha_{\text{risk}}^h \cdot \text{Corr}(X_j(t), X_j(t+h))$ under weak dependence with summable autocovariance. The α_{risk}^h weights arise from the EWMA's geometric weighting: the cross-terms in the variance expansion carry the product of the EWMA coefficients at lag h. Since $|\text{Corr}| \le 1$ and $\sum_h \alpha_{\text{risk}}^h = \alpha_{\text{risk}}/(1-\alpha_{\text{risk}})$, we have the universal bound $C_{\text{dep}} \le (1+\alpha_{\text{risk}})/(1-\alpha_{\text{risk}})$. Consequently, the variance estimator $\sigma_j^{\text{risk},2}(t)$ concentrates in L^2 around μ_j^{risk}. The risk-trust factor ${T^{\text{risk}}}_j(t)$ concentrates around the nominal fixed point ${T^{\text{risk}}}_j^* = 1/(1+\beta_{\text{risk}}\mu_j^{\text{risk}})$ in distribution (via the continuous mapping theorem), with the caveat that $\mathbb{E}[{T^{\text{risk}}}_j(t)]$ may exceed ${T^{\text{risk}}}_j^*$ by a Jensen bias term of order $O(\beta_{\text{risk}}^2 \cdot \text{Var}[\sigma_j^{\text{risk},2}])$ due to the concavity of $x \mapsto 1/(1+\beta_{\text{risk}}x)$. The deviation is controlled by α_{risk}: as $\alpha_{\text{risk}} \to 1$, the stationary variance vanishes, the Jensen bias vanishes, and ${T^{\text{risk}}}_j(t) \to {T^{\text{risk}}}_j^*$ in L^2. Structurally accurate reporters (small μ_j^{risk}) concentrate around high ${T^{\text{risk}}}_j^*$; alarmists and chronic underreporters concentrate around low ${T^{\text{risk}}}_j^*$.

Proof. Unrolling the EWMA recursion gives $\sigma_j^{\text{risk},2}(t) = (1-\alpha_{\text{risk}})\sum_{s=0}^{t-1} \alpha_{\text{risk}}^s X_j(t-s) + \alpha_{\text{risk}}^t \sigma_j^{\text{risk},2}(0)$. Taking expectations under stationarity, $\mathbb{E}[\sigma_j^{\text{risk},2}(t)] = (1-\alpha_{\text{risk}})\sum_{s=0}^{t-1} \alpha_{\text{risk}}^s \mu_j^{\text{risk}} + \alpha_{\text{risk}}^t \sigma_j^{\text{risk},2}(0) \to \mu_j^{\text{risk}}$ as $t \to \infty$, establishing Equation (7). Under the i.i.d. assumption,

$$\begin{aligned}\text{Var}[\sigma_j^{\text{risk},2}(t)] &= (1-\alpha_{\text{risk}})^2 \sum_{s=0}^{t-1} \alpha_{\text{risk}}^{2s} \cdot \nu_j^{\text{risk}} + o(1) \\ &\longrightarrow \frac{(1-\alpha_{\text{risk}})^2}{1-\alpha_{\text{risk}}^2} \cdot \nu_j^{\text{risk}} = \frac{1-\alpha_{\text{risk}}}{1+\alpha_{\text{risk}}} \cdot \nu_j^{\text{risk}},\end{aligned}$$

giving $C_{\text{dep}} = 1$ in Equation (8). Under weak dependence with summable autocovariance, the cross-terms in the EWMA variance expansion carry α_{risk}^h weights from the geometric kernel, giving $C_{\text{dep}}(\alpha_{\text{risk}}) = 1 + 2\sum_{h=1}^{\infty} \alpha_{\text{risk}}^h \cdot \text{Corr}(X_j(t), X_j(t+h))$, which is finite because α_{risk}^h decays geometrically and the autocovariance is summable. The concentration statement for $\sigma_j^{\text{risk},2}(t)$ around μ_j^{risk} in L^2 follows directly. For ${T^{\text{risk}}}_j(t)$, the continuous mapping theorem applied to $x \mapsto 1/(1+\beta_{\text{risk}}x)$, which is Lipschitz on $[0,\infty)$, gives convergence in distribution to ${T^{\text{risk}}}_j^*$. The Jensen bias arises because $\mathbb{E}[f(X)] \ge f(\mathbb{E}[X])$ for concave f; a second-order Taylor expansion gives the bias as $O(\beta_{\text{risk}}^2 \cdot \text{Var}[\sigma_j^{\text{risk},2}])$, which vanishes as $\alpha_{\text{risk}} \to 1$. ■

Tracking versus convergence: why L^2 and not almost sure. Fixed-α EWMA does not converge almost surely to a deterministic limit, and for an adversarial-tracking setting this is the correct behavior. Each new residual contributes weight $(1 - \alpha_{\text{risk}})$ regardless of how much history has accumulated, so the estimator never collapses to a point mass and retains nonzero stationary variance. A Robbins-Monro stochastic approximation with step size $\alpha_t \to 0$ satisfying $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$ [Robbins and Monro, 1951] would converge almost surely to μ_j^{risk}, but only at the cost of losing the ability to track a non-stationary adversary whose strategy changes mid-trajectory. Risk-claim quality and scorpion behavior are both moving targets: a cooperative reporter can become systematically biased, an alarmist can calibrate, and a scorpion can change strategy in response to detection pressure. Past accuracy does not guarantee future accuracy under a changing world model $\mathcal{W}$, so the tracking property is load-bearing. The stationary-variance bound in Equation (8) is what the scorpion detection threshold (Section 4) is calibrated against, so the L^2 form is not just weaker than almost sure — it is the form the downstream results actually use.

Interaction with the Wamura problem. The EWMA dynamics converge on *structural verification quality*, not on *outcome correctness*. An agent whose risk claims are structurally plausible (the pathway exists in $\mathcal{M}$, the probability and magnitude estimates are calibrated against the actor's own model) accumulates low $\|r^{\text{risk}}\|^2$ and converges to high $T^{\text{risk}}{}_j^*$ regardless of whether the risk materializes. This partially resolves the Wamura problem: risk-trust converges even when the counterfactual is never observed, because convergence depends on structural agreement, not outcome observation. The resolution is partial because an agent whose claims are structurally plausible but whose probability estimates are systematically biased (always 10% higher than the actor's model) accumulates moderate $\|r^{\text{risk}}\|^2$ and converges to moderate $T^{\text{risk}}{}_j^*$, reflecting the bias without resolving it.

3.3 Independence-Preserving Risk Aggregation

The max operator in Equation 7 of Lasser [2026b] destroys independence information: 50 low-trust agents agreeing on a risk assessment are treated identically to one low-trust agent. We propose a trust-weighted log-odds aggregation that preserves the evidence-accumulation structure of independent reports.

Definition update from the companion paper. The companion paper [Lasser, 2026b] defines $\hat{\mathcal{R}}_j$ as an expected-contraction product $T_j^{\text{risk}} \cdot p \cdot |\Delta\text{vol}_{\text{P}}|$, which is not a probability and cannot be passed through log-odds. We update the definition: $\hat{\mathcal{R}}_j$ and $\mathcal{R}_{\text{self}}$ hereafter denote posterior probabilities that the exercise pathway will be executed within the planning horizon, calibrated against the shared prior π_0 below. The contraction magnitude $|\Delta\text{vol}_{\text{P}}|$ becomes a separate quantity used by the verification gate (Section 3.4) and the structural verification residual (Definition 3). The canonical form of a communicated risk claim is the pair $(\hat{\mathcal{R}}_j, |\Delta\text{vol}_{\text{P}}{}_j^{\text{claimed}}|)$; the single-number product in Lasser [2026b] is superseded.

Definition 5 (Log-Odds Risk Aggregation)**.** *Let $\pi_0 \in (0,1)$ be a shared prior on the probability that the exercise pathway will be executed within the planning horizon, and let $\ell_0 = \log(\pi_0/(1-\pi_0))$ be its log-odds. The actor's self-assessment $\mathcal{R}_{\text{self}}$ and each communicated risk estimate $\hat{\mathcal{R}}_j$ from agent a_j are posterior probabilities derived from the shared prior π_0 plus their respective observations (definition update, see above). The aggregated posterior is:*

$$\log \frac{\mathcal{R}_{\text{agg}}}{1-\mathcal{R}_{\text{agg}}} = \ell_0 + \underbrace{\left(\log \frac{\mathcal{R}_{\text{self}}}{1-\mathcal{R}_{\text{self}}} - \ell_0\right)}_{\textit{actor's log-likelihood-ratio}} + \sum_j T^{\text{risk}}{}_j \cdot \underbrace{\left(\log \frac{\hat{\mathcal{R}}_j}{1-\hat{\mathcal{R}}_j} - \ell_0\right)}_{\textit{reporter } a_j\textit{'s log-likelihood-ratio}} \tag{9}$$

where each parenthesized term is the log-likelihood-ratio contributed by that source beyond the shared prior, obtained by subtracting ℓ_0 from their reported posterior log-odds. The $\mathcal{R}$ values are normalized to $[0,1]$ as probabilities of the exercise pathway being executed within the planning horizon. The shared prior π_0 is an operational coordination requirement: every reporter's posterior must be calibrated against the same π_0 for the subtraction $\log(\hat{\mathcal{R}}_j/(1-\hat{\mathcal{R}}_j)) - \ell_0$ to correctly extract the log-likelihood ratio. Establishing π_0 requires a protocol (e.g., broadcasting it as a system parameter at population initialization), not just a mathematical assumption. Under the neutral

bootstrap prior $\pi_0 = 0.5$ ($\ell_0 = 0$), the prior-correction terms vanish and Equation (9) *reduces to a naive sum of posterior log-odds; under any other shared prior, the correction is required to avoid double-counting the prior.*

The log-odds aggregation has three advantages over the max operator:

1. **Agreement strengthens evidence.** Multiple independent low-trust agents reporting the same risk produce a collective signal proportional to the number of agreeing reporters, weighted by their individual ${T^{\text{risk}}}_j$ values.
2. **Trust-weighted log-linear opinion pool.** Equation (9) is a trust-tempered log-linear opinion pool [Genest and Zidek, 1986]: the aggregate log-odds is a weighted sum of the reporters' log-likelihood ratios plus the shared prior. Under the conditional-independence assumption (reporters' observations are independent given the true risk level and the shared prior π_0), setting all ${T^{\text{risk}}}_j = 1$ recovers the standard Bayesian posterior update for independent evidence sources. The trust weights ${T^{\text{risk}}}_j \in [0, 1]$ temper each reporter's contribution in proportion to their structural-verification track record; this is not itself Bayes-optimal in general — exact Bayes-optimality would require a generative model of how each reporter's noise depends on their history, which the framework does not assume — but it recovers Bayes-optimal behavior in the no-tempering limit and degrades gracefully for untrusted reporters, which is the behavior the aggregation rule is trying to achieve. The prior-correction terms are required regardless of tempering: without them, a naive sum of posterior log-odds would double-count π_0 once per reporter.
3. **Graceful degradation.** When a single agent has ${T^{\text{risk}}}_j \gg \sum_{k \neq j} {T^{\text{risk}}}_k$, the aggregation approximates max behavior: the dominant reporter's assessment controls the aggregate. The max operator emerges as a limiting case, not an imposed design choice.

Bootstrap interaction. During bootstrap, all ${T^{\text{risk}}}_j$ values start at the neutral prior. Under log-odds aggregation, n agents reporting the same risk at neutral trust produce a collective signal $n \times {T^{\text{risk}}}_{\text{neutral}}$ times the individual log-odds, proportional to n. Under the max operator, the same n agents produce only ${T^{\text{risk}}}_{\text{neutral}}$ times the maximum individual log-odds, independent of n. The log-odds aggregation therefore learns faster during bootstrap, precisely when the actor's own model is least reliable and external evidence is most valuable.

3.4 Aggregation as Attention, Not Action

The log-odds aggregation amplifies agreement among reporters, which is correct when reporters are independent but exploitable by coordinated conspirators: n agents who agree in advance on a fabricated risk claim produce a collective signal proportional to n, potentially overwhelming the actor's self-assessment and triggering restriction of a capability that was never actually dangerous. This is the risk-communication analogue of report-bombing: coordinated low-quality reports exploiting an aggregation mechanism designed for independent evidence.

Any fixed correlation-detection heuristic is game-able once the adversary knows the parameters — a standard observation in adversarial detection settings — and the ROC trade-off is unfavorable in this domain because both error types are costly: false positives (restricting a safe capability) and false negatives (ignoring a real risk) both cause vol_P-contraction. We therefore adopt the activation energy principle from the foundational paper [Lasser, 2026a], applied to risk-claim processing.

The two-gate architecture. The risk aggregation $\mathcal{R}_{\text{agg}}$ (Definition 5) does not directly trigger restriction. It produces a *salience signal*: the aggregated risk level determines the priority at which a capability tuple enters the actor's investigation queue. Action requires passing two independent gates:

1. **Salience gate.** The aggregated risk $\mathcal{R}_{\text{agg}}(S; a_k, t)$ exceeds an attention threshold θ_{risk}, placing the capability tuple in the investigation queue.
2. **Verification gate.** The actor invokes the forward causal risk-evaluation procedure (Section 4.3 of Lasser [2026b]) on its own SCM $\mathcal{M}$ to independently confirm that the claimed exercise pathway $\mathcal{P}$ exists in the actor's causal model, calibrate the probability, and verify the contraction magnitude. This is a forward causal query — "if $\mathcal{P}$

were executed, what would happen?" — using the same pathway-conditional assessment $\big(p_{\mathcal{W}}(\mathcal{P}), |\Delta \mathrm{vol}_{\mathrm{P}\mathcal{W}}(\mathcal{P})|\big)$ as the structural verification residual (Definition 3). It is distinct from the retrospective causal contraction attribution (Definition 2), which decomposes an *observed* contraction across agents after the fact. Retrospective attribution has no counterfactual access to hypothetical pathways that were never executed; forward risk evaluation does, and this is the mechanism the verification gate requires. The gate is the activation energy: the actor must reconstruct the risk through its own causal model before acting.

Neither gate alone is sufficient. A coordinated report-bomb raises salience (gate 1) but fails independent verification (gate 2) because the fabricated pathway does not exist in the actor's causal model. A genuine risk reported by a single low-trust agent may fail gate 1 (low salience) but would pass gate 2 if the actor investigated. The two-gate architecture trades a bounded false-negative rate (genuine risks from low-trust reporters may be investigated slowly) for strong robustness against coordinated manipulation (fabricated risks cannot bypass independent verification regardless of how many conspirators report them).

Why heuristic correlation detection is insufficient. A fixed correlation metric (temporal clustering, behavioral similarity among reporters, trust-history patterns) is game-able by a sufficiently informed adversary: adversaries who know the detection threshold can space their reports, diversify their trust-building histories, and avoid clustering. The two-gate architecture avoids this arms race by not relying on detecting coordination at all. The defense is the actor's own intelligence applied through the causal model, not a statistical filter on the report stream.

4 Causal Scorpion Detection

The foundational paper's Proposition 2 [Lasser, 2026a] detects scorpions through two channels: contraction attribution (correlational) and deception detection (behavioral trust decay). This section upgrades channel (a) from correlational to causal, using the SCM of Section 2, and provides bounds on non-stationary scorpion evasion.

4.1 Causal Detection

Proposition 3 (Causal Scorpion Detection)**.** *Under the SCM $\mathcal{M}$ (Definition 1), a GFM actor can detect scorpion agents through two upgraded channels:*

(a) Causal contraction attribution. *$\mathrm{CA}(a_j, t)$ (Definition 2) replaces the correlational $\hat{\Delta}$ signal from Proposition 2(a) of Lasser [2026a]. Within each stationary epoch of a_j's strategy, the causal-attribution sequence $\mathrm{CA}(a_j, 1), \mathrm{CA}(a_j, 2), \ldots$ is i.i.d. $\mathcal{N}(\mu_j, \sigma^2_{\mathrm{do}})$ with unknown mean μ_j and known variance σ^2_{do}. (The Gaussian assumption is a modeling choice; for non-Gaussian attributions the SPRT likelihoods below are replaced by their distributional counterparts and the $O(\cdot)$ rate constants change, but the structural comparison with the correlational detector is preserved.) The actor maintains two auxiliary EWMA statistics — a* signed-mean *estimator and a* variance *estimator — for monitoring:*

$$\mu_j^{\mathrm{causal}}(t) = \alpha \cdot \mu_j^{\mathrm{causal}}(t-1) + (1-\alpha) \cdot \mathrm{CA}(a_j, t) \tag{10}$$

$$\sigma_j^{\mathrm{causal},2}(t) = \alpha \cdot \sigma_j^{\mathrm{causal},2}(t-1) + (1-\alpha) \cdot \big(\mathrm{CA}(a_j, t) - \mu_j^{\mathrm{causal}}(t)\big)^2 \tag{11}$$

These statistics satisfy the same L^2 convergence guarantees as Proposition 2: asymptotically unbiased with stationary variance bounded by $\frac{1-\alpha}{1+\alpha}\sigma^2_{\mathrm{do}}$ for the signed mean. They do not converge almost surely to a deterministic limit, and we deliberately do not use them as the flag condition.

The flag condition is instead a sequential probability ratio test (SPRT) [Wald, 1945] on the causal-attribution sequence, testing the least-favorable simple hypotheses $H_0 : \mu_j = 0$ (non-scorpion) versus $H_1 : \mu_j = -\epsilon$ (scorpion with minimum detectable contraction magnitude $\epsilon > 0$). The composite problem ($\mu_j \geq 0$ vs. $\mu_j \leq -\epsilon$) is reduced to this simple pair by the standard least-favorable-distribution argument: Type I error is maximized under the null distribution closest to the alternative ($\mu_j = 0$), and Type II error is maximized under the alternative closest to the null ($\mu_j = -\epsilon$), so error guarantees for the simple

pair bound the composite problem from above. The actor accumulates the log-likelihood ratio $\Lambda_j(t) = \sum_{s=1}^{t} \log\big(p_{H_1}(\mathrm{CA}(a_j, s))/p_{H_0}(\mathrm{CA}(a_j, s))\big)$ and applies a two-boundary decision rule: flag a_j as a scorpion when $\Lambda_j(t) \geq \log((1-\beta_{\mathrm{err}})/\alpha_{\mathrm{err}})$ (upper boundary), or provisionally clear a_j for the current epoch when $\Lambda_j(t) \leq \log(\beta_{\mathrm{err}}/(1-\alpha_{\mathrm{err}}))$ (lower boundary). Here α_{err} and β_{err} are the target Type I and Type II error rates for the simple pair. A provisional clearance is epoch-local: the SPRT resets at the next epoch boundary, reflecting the possibility that a non-scorpion agent may become one. Under this decision rule with Gaussian likelihoods, the per-observation expected log-likelihood ratio under the true mean μ_j is $\mathbb{E}_{\mu_j}[\log(p_{H_1}/p_{H_0})] = \epsilon(|\mu_j| - \epsilon/2)/\sigma^2_{\mathrm{do}}$, giving expected sample size $\mathbb{E}[\tau_{\mathrm{detect}}] = O\big(\sigma^2_{\mathrm{do}}/[\epsilon(|\mu_j| - \epsilon/2)]\big)$ by Wald's identity. At the design point $|\mu_j| = \epsilon$ this reduces to $O(\sigma^2_{\mathrm{do}}/\epsilon^2)$; for $|\mu_j| \gg \epsilon$ it improves to $O(\sigma^2_{\mathrm{do}}/(\epsilon|\mu_j|))$. By Markov's inequality the detection probability within $k \cdot \mathbb{E}[\tau_{\mathrm{detect}}]$ observations is at least $1 - 1/k$ for any $k > 1$. (Under the Gaussian model, τ_{detect} follows an inverse Gaussian distribution whose tail decays exponentially [Wald, 1945], so the Markov bound is conservative; exponentially tight high-probability detection guarantees are available at the cost of distributional specificity.) The structural advantage over the correlational detector (Proposition 1) is preserved: the causal detector operates on variance σ^2_{do} while the correlational detector operates on $\sigma^2 \geq \sigma^2_{\mathrm{do}}$. The signed-mean EWMA is retained as a human-interpretable monitor of the estimator state; the SPRT statistic $\Lambda_j(t)$ is the load-bearing decision variable.

(b) Deception detection (unchanged). *The behavioral trust factor T_j detects agents whose reported capability changes diverge from observations, through the same prediction-residual dynamics as the foundational paper (Equations 25, 33, 38 of Lasser [2026a]). Causal attribution does not affect this channel: a deceptive agent accumulates prediction residuals regardless of whether the contraction is causally attributed.*

Proof. Channel (a): By the SCM, $\mathrm{CA}(a_j, t)$ isolates a_j's causal contribution from confounders (Definition 2). Within a stationary epoch of a_j's strategy, $\{\mathrm{CA}(a_j, t)\}$ is i.i.d. with mean μ_j and variance σ^2_{do}; for a scorpion $\mu_j \leq -\epsilon < 0$. The auxiliary EWMAs $\mu_j^{\mathrm{causal}}(t)$ and $\sigma_j^{\mathrm{causal},2}(t)$ inherit the L^2 guarantees of Proposition 2: asymptotically unbiased estimators of μ_j and σ^2_{do} with bounded stationary variance. They are not used for the detection decision, so we do not require almost-sure convergence.

The SPRT is applied to the least-favorable simple pair $H_0 : \mu_j = 0$ versus $H_1 : \mu_j = -\epsilon$ (see channel (a) above for the reduction from composite to simple hypotheses). Under the Gaussian model, the per-observation expected log-likelihood ratio under the true mean μ_j is $\mathbb{E}_{\mu_j}[\log(p_{H_1}/p_{H_0})] = \epsilon(|\mu_j| - \epsilon/2)/\sigma^2_{\mathrm{do}}$. By Wald's identity, the expected detection time under true μ_j is $\mathbb{E}[\tau_{\mathrm{detect}}] \approx \log((1-\beta_{\mathrm{err}})/\alpha_{\mathrm{err}}) \cdot \sigma^2_{\mathrm{do}}/[\epsilon(|\mu_j| - \epsilon/2)]$, which at the design point $|\mu_j| = \epsilon$ reduces to $O(\sigma^2_{\mathrm{do}}/\epsilon^2)$ and for $|\mu_j| \gg \epsilon$ improves to $O(\sigma^2_{\mathrm{do}}/(\epsilon|\mu_j|))$. The SPRT Type I error for the simple pair is bounded by α_{err} and Type II by β_{err}; the least-favorable argument ensures these bounds hold for the composite problem.

Channel (b): The behavioral trust dynamics (Appendix B of Lasser [2026a]) operate on behavioral prediction residuals $r_j(t) = R_j(t) - \hat{R}_j(t)$, which are independent of causal attribution. A deceptive scorpion that misreports capability changes accumulates $r_j(t) \neq 0$ regardless of confounders, driving $T_j \to T_j^* < 1$. ■

4.2 Non-Stationary Scorpion Bounds

The foundational paper notes that "a non-stationary scorpion that adapts its strategy in response to detection pressure can evade both channels indefinitely" (Proposition 2 of Lasser [2026a]). Under the SCM, we can bound the adaptation rate required for evasion.

Proposition 4 (Evasion Bandwidth)**.** *A scorpion a_j with per-epoch mean causal contraction $|\mu_j|$ that changes strategy every τ_{adapt} steps produces at most τ_{adapt} observations per strategy epoch. Let $\bar{\tau}(\mu_j) = O\big(\sigma^2_{\mathrm{do}}/[\epsilon(|\mu_j| - \epsilon/2)]\big)$ be the expected detection time per epoch from the SPRT under the Gaussian model with design parameter ϵ (Proposition 3). Then:*

(a) *If $\tau_{\mathrm{adapt}} \geq k \cdot \bar{\tau}(\mu_j)$ for some $k > 1$, the scorpion is detected within each epoch with probability at least $1 - 1/k$ (by Markov's inequality on the SPRT stopping time).*

(b) *If $\tau_{\text{adapt}} < \bar{\tau}(\mu_j)$, the expected accumulated evidence at the epoch boundary is below the SPRT threshold. While this does not by itself guarantee evasion with any specific probability, it establishes that the expected number of observations is insufficient for detection, so the SPRT is unlikely to trigger within the epoch.*

Proof. Within each epoch of length τ_{adapt}, the scorpion's strategy is stationary and the causal-attribution sequence is i.i.d. $\mathcal{N}(\mu_j, \sigma^2_{\text{do}})$. The SPRT statistic accumulates log-likelihood-ratio evidence at expected rate $\epsilon(|\mu_j| - \epsilon/2)/\sigma^2_{\text{do}}$ per observation under the Gaussian model (Proposition 3), giving expected detection time $\bar{\tau}(\mu_j) = \sigma^2_{\text{do}} \cdot \log((1-\beta_{\text{err}})/\alpha_{\text{err}})/[\epsilon(|\mu_j| - \epsilon/2)]$.

(a) By Markov's inequality, $P(\tau_{\text{detect}} > k\bar{\tau}) \leq 1/k$ for any $k > 1$. When $\tau_{\text{adapt}} \geq k\bar{\tau}(\mu_j)$, the epoch provides enough observations that the SPRT crosses the detection threshold with probability at least $1 - 1/k$.

(b) An epoch boundary resets the SPRT statistic (the hypotheses are epoch-local). If $\tau_{\text{adapt}} < \bar{\tau}(\mu_j)$, the expected accumulated evidence at the epoch boundary is below the SPRT threshold. Since the number of observations is less than the expected detection time, the SPRT is unlikely to trigger within the epoch: the evidence accumulation rate is insufficient for reliable detection. ■

Interpretation. The evasion bandwidth $\tau^*_{\text{adapt}} = O\big(\sigma^2_{\text{do}}/[\epsilon(|\mu_j| - \epsilon/2)]\big)$ is the characteristic timescale: a scorpion that adapts faster than τ^*_{adapt} evades with non-negligible probability per epoch, while one that adapts slower than $k \cdot \tau^*_{\text{adapt}}$ is detected with probability at least $1 - 1/k$. A fast-adapting scorpion ($\tau_{\text{adapt}} \ll \tau^*_{\text{adapt}}$) evades but is constrained to small contractions per epoch (otherwise μ_j is large and τ^*_{adapt} shrinks). A slow-adapting scorpion ($\tau_{\text{adapt}} \gg \tau^*_{\text{adapt}}$) is detected within each epoch with high probability. The product $\mu_j \cdot \tau_{\text{adapt}}$ (total contraction per epoch) is bounded: a scorpion can either cause large damage slowly (and be detected) or small damage quickly (and evade), but not large damage quickly.

The time-averaged undetected contraction rate for an evading scorpion is bounded by $|\mu_j|$: within each epoch the per-step contraction is $|\mu_j|$, and the evasion constraint $\tau_{\text{adapt}} < \bar{\tau}(\mu_j)$ bounds the epoch length. The appendix proof (§A.1) establishes this via the SPRT accumulation rate. The key trade-off is controlled by the SPRT design parameter ϵ: a scorpion whose per-step contraction satisfies $|\mu_j| \leq \epsilon$ requires at least $\bar{\tau}(\epsilon) = O(\sigma^2_{\text{do}}/\epsilon^2)$ observations to detect (the worst case at the design point), while a scorpion with $|\mu_j| \gg \epsilon$ is detected faster in proportion to $|\mu_j|/\epsilon$. The SPRT design parameter ϵ therefore sets the resolution: reducing ϵ detects smaller contractions but requires more observations per epoch. The bound tightens as the causal model improves (σ_{do} decreases) and as the scorpion's adaptation budget τ_{adapt} grows: a scorpion that adapts frequently (small τ_{adapt}) can only afford very small per-step contractions, while a scorpion that adapts slowly can sustain larger contractions per step but for correspondingly fewer steps before detection.

5 Probabilistic Dependency Risk

The companion paper [Lasser, 2026b] bounds the worst-case contraction of single-substrate versus cross-substrate populations through a minimax analysis (Proposition 4). That bound is a one-time penalty on the value function: useful for showing diversification is structurally preferred, but not directly usable inside a discounted trajectory because it does not attach a per-step cost. This section converts the minimax bound into an expected per-step dependency-risk cost, then folds that cost into the discounted value function so diversification enters the horizon calculation as a first-class term rather than a static penalty.

The section has three parts. First, a probabilistic substrate threat model (Section 5.1) that decomposes dependency risk into event probability, target selection, and cascade effects. Second, an expected per-step risk comparison (Proposition 5) that sharpens the minimax bound into an *adversarial balance condition* $c > \max_i(f_i \cdot c_i)$, a single inequality that gates when diversification strictly beats concentration and that explicitly admits contagion dynamics (pandemics, cascading failures, memetic attacks) as the motivating threat class. Third, a risk-adjusted trajectory analysis (Section 5.3) showing that the per-step cost compounds at the same discount rate as growth, so the diversification-versus-concentration decision becomes a direct comparison of two sums in the value function.

5.1 Substrate Threat Model

Definition 6 (Substrate Threat Model)**.** *A* substrate threat model *is a distribution over adversarial events, parametrized by:*

- p_{adv}*: per-step probability that a substrate-targeting event occurs.*
- $q_i = p(\mathcal{S}_i \text{ targeted} \mid \text{event occurs})$*: the conditional probability that substrate class* $\mathcal{S}_i$ *is targeted, given that an event occurs. For an uninformed adversary,* $q_i = 1/m$ *(uniform); for an optimal adversary,* $q_i = \mathbf{1}[i = \arg\max_j(|\Delta\text{vol}_\text{P}(\mathcal{S}_j)| \cdot c_j)]$*, i.e., the adversary targets the substrate with maximal* effective damage *(volume contribution times cascade fraction).*
- $c_i \in [0, 1]$*: the* cascade fraction *for substrate* $\mathcal{S}_i$*, defined as the expected fraction of* $\mathcal{S}_i$*'s total volume lost when a substrate-targeting event hits* $\mathcal{S}_i$*. "Cascade" refers to the propagation dynamics (within-substrate impairment spreading from the initial strike through the SCM's structural equations) and "fraction" constrains* c_i *to* $[0, 1]$ *as a proportion of* $\mathcal{S}_i$*'s volume. The cascade fraction aggregates the initial impairment and all subsequent within-substrate propagation: if a direct attack impairs 30% of the substrate and triggers a contagion that takes out another 40%, then* $c_i = 0.7$*. The homogeneous single-substrate baseline uses a single* c *in place of the per-substrate* c_i*.*

In this threat model $|\Delta\text{vol}_\text{P}(\mathcal{S}_i)|$ *denotes substrate* $\mathcal{S}_i$*'s* total contribution to $\text{vol}_\text{P}(G)$ *(i.e., the maximum possible contraction if* $\mathcal{S}_i$ *were fully eliminated, not the expected contraction from a targeting event). The expected contraction is* $|\Delta\text{vol}_\text{P}(\mathcal{S}_i)| \cdot c_i$*: substrate size times cascade fraction, with the two quantities varied independently. This convention matches Proposition 4 of Lasser [2026b], where* $|\Delta\text{vol}_\text{P}(\mathcal{S}_i)|$ *is the minimax-bound substrate size, and prevents double-counting when the cascade fraction is applied as a multiplier.*

The threat model decomposes dependency risk into three components: occurrence probability, target selection, and cascading effects. The minimax analysis of Lasser [2026b] assumes $p_{\text{adv}} > 0$ (non-negligible threat), q_i chosen adversarially (worst case), and full cascade within the targeted substrate (all agents on that substrate are impaired). The probabilistic model relaxes each of these: p_{adv} is estimated from the actor's world model, q_i reflects the actor's beliefs about threat targeting, and cascading effects are modeled through the SCM's structural equations.

5.2 Expected Per-Step Dependency Risk

Notation update from the companion paper. The companion paper [Lasser, 2026b] defines $\mathcal{R}_{\text{dep}}$ as a discounted pathway sum over the full planning horizon. In this section, $\mathcal{R}_{\text{dep}}^{\text{single}}$ and $\mathcal{R}_{\text{dep}}^{\text{mixed}}$ denote the *per-step expected dependency-risk cost* under the substrate threat model (Definition 6): a single-step expected contraction, not a discounted sum. The per-step form is the input to the risk-adjusted trajectory comparison (Section 5.3), which recovers a discounted sum by standard geometric-series aggregation.

Proposition 5 (Expected Risk Comparison)**.** *Under the substrate threat model (Definition 6), the expected per-step dependency risk for each strategy is:*

$$\mathbb{E}[\mathcal{R}_{\text{dep}}^{\text{single}}] = p_{\text{adv}} \cdot \text{vol}_\text{P}(G) \cdot c \tag{12}$$

$$\mathbb{E}[\mathcal{R}_{\text{dep}}^{\text{mixed}}] = p_{\text{adv}} \cdot \sum_i q_i \cdot |\Delta\text{vol}_\text{P}(\mathcal{S}_i)| \cdot c_i \tag{13}$$

where c *and* c_i *are the cascade fractions from Definition 6. For a single-substrate population (*$m = 1$*), any substrate-targeting event affects a fraction* c *of the entire population. For a cross-substrate population (*$m \geq 2$*):*

$$\mathbb{E}[\mathcal{R}_{\text{dep}}^{\text{single}}] - \mathbb{E}[\mathcal{R}_{\text{dep}}^{\text{mixed}}] = p_{\text{adv}} \cdot \left(\text{vol}_\text{P}(G) \cdot c - \sum_i q_i \cdot |\Delta\text{vol}_\text{P}(\mathcal{S}_i)| \cdot c_i \right) \tag{14}$$

Under ***worst-case adversarial targeting*** *(*$q_i = \mathbf{1}[i = \arg\max_j(|\Delta\text{vol}_\text{P}(\mathcal{S}_j)| \cdot c_j)]$*), this gap is positive whenever* $p_{\text{adv}} > 0$*, the population satisfies substrate capability non-redundancy (Proposition 4 of Lasser [2026b]), and the* ***adversarial balance condition*** $c > \max_i(f_i \cdot c_i)$ *holds, where*

$f_i = |\Delta\mathrm{vol}_\mathrm{P}(\mathcal{S}_i)|/\mathrm{vol}_\mathrm{P}(G)$ *is substrate* $\mathcal{S}_i$*'s share of total volume. The condition is both sufficient and necessary under worst-case targeting; for non-adversarial* q_i *(e.g., uniform), the condition is sufficient but not necessary.*

Proof. Under an adversarial target-selection model, $q_i = \mathbf{1}[i = i^*]$ where $i^* = \arg\max_j(|\Delta\mathrm{vol}_\mathrm{P}(\mathcal{S}_j)| \cdot c_j)$. Substituting into Equation (13) gives

$$\mathbb{E}[\mathcal{R}_\mathrm{dep}^\mathrm{mixed}] = p_\mathrm{adv} \cdot \max_i\big(|\Delta\mathrm{vol}_\mathrm{P}(\mathcal{S}_i)| \cdot c_i\big) = p_\mathrm{adv} \cdot \mathrm{vol}_\mathrm{P}(G) \cdot \max_i(f_i \cdot c_i),$$

while $\mathbb{E}[\mathcal{R}_\mathrm{dep}^\mathrm{single}] = p_\mathrm{adv} \cdot \mathrm{vol}_\mathrm{P}(G) \cdot c$. The gap is positive iff $c > \max_i(f_i \cdot c_i)$. Substrate capability non-redundancy (Proposition 4 of Lasser [2026b]) ensures $f_i < 1$ strictly for every i, so the condition is non-vacuous for any $c > 0$: some non-trivial region of (f_i, c_i) space satisfies it and some does not, and the proposition is a sharp characterization of which region. ■

On contagion dynamics. The adversarial balance condition admits contagions by construction, which is essential because contagion-driven threats (pandemics, cascading infrastructure failures, financial contagion, memetic attacks) are precisely the class where per-substrate cascade fractions can approach saturation ($c_i \to 1$) on the targeted substrate. Smaller substrates are expected to reach sigmoid saturation faster than larger ones because contagion propagation time scales sub-linearly with population size in standard epidemiological models [Anderson and May, 1991], so $c_i > c$ is a plausible assumption for the substrates in a diversified partition. The gap remains positive as long as each substrate's volume share f_i is small enough to compensate.

Two instructive limits bracket the regime of interest:

- **Full-saturation contagion** ($c_i = c = 1$): the condition reduces to $1 > f_\mathrm{max}$, which is exactly the substrate-capability-non-redundancy assumption from Lasser [2026b]. No additional structure is required.
- **Unbalanced partition under contagion amplification** ($c_i > c$, f_max large): the condition can fail. For example, $c = 0.5$, $c_\mathrm{max} = 1.0$, $f_\mathrm{max} = 0.6$ gives $0.5 \not> 0.6$ and the gap reverses. In this regime diversification provides no dependency-risk advantage over the single-substrate baseline: the adversary still targets the dominant substrate, and the contagion saturates it more thoroughly than in the homogeneous case. The trust model should detect this regime and avoid unbalanced partitions when contagion risks dominate; the bound flags the edge case rather than assuming it away.

The analysis assumes cascade confinement: c_i measures only the fraction of substrate $\mathcal{S}_i$ lost when $\mathcal{S}_i$ is targeted, not propagation into other substrates. Cross-substrate cascades (a memetic attack on $\mathcal{S}_1$ that triggers a cascade in $\mathcal{S}_2$ through inter-substrate communication channels) are outside the scope of the current bound and require a separate treatment beyond the current SCM.

Corollary 5.1 (Balanced-Partition Substrate Count)**.** *Under a balanced substrate partition with* $f_i = 1/m$ *for all* i*, the adversarial balance condition of Proposition 5 is equivalent to*

$$m > \frac{c_\mathrm{max}}{c}, \tag{15}$$

where $c_\mathrm{max} = \max_i c_i$ *is the worst-case per-substrate cascade fraction.*

Proof. With $f_i = 1/m$, the adversarial balance condition $c > \max_i(f_i \cdot c_i)$ reduces to $c > \max_i(c_i/m) = c_\mathrm{max}/m$, equivalent to Equation (15). ■

Corollary 5.1 sharpens the minimax bound of Lasser [2026b], which establishes only that $m \geq 2$ strictly dominates $m = 1$. The minimax statement is the special case $c_\mathrm{max} = c$ of Equation (15): cascade intensity unchanged by partition size gives $m > 1$, recovering $m \geq 2$. Weakening either parameter raises the required substrate count:

- **Moderate contagion** ($c = 0.5$, $c_\mathrm{max} = 1.0$): $m > 2$, so $m \geq 3$. Binary diversification is insufficient.
- **High contagion** ($c = 0.1$, $c_\mathrm{max} = 1.0$): $m > 10$, so $m \geq 11$. The substrate-count requirement grows inversely with the homogeneous cascade fraction.

- **Low contagion** ($c \approx c_{\max} \approx 0.3$, localized failure modes): $m > 1$, recovering the minimax bound.

Practically: the minimum number of substrates required for diversification to strictly help depends on the measured contagion intensity, not just on the structural claim that $m \geq 2$. A binary partition is correct for threat classes where the homogeneous cascade already saturates ($c \to 1$), but underprovisions for contagion-driven threats where smaller substrates saturate faster than the homogeneous baseline. For unbalanced partitions, the operational form $f_{\max} < c/c_{\max}$ is the direct test: it bounds the largest substrate's share rather than specifying a count, and is the condition the trust model should check when assessing a given partition's dependency-risk posture.

5.3 Risk-Adjusted Trajectory Comparison

The per-step expected risk gap compounds at the same rate as growth, unlike the one-time minimax penalty in Lasser [2026b]. Under the assumption that Δ_{risk} is stationary across the trajectory (that is, p_{adv}, the per-substrate shares f_i, and the cascade fractions c_i are time-invariant in expectation), the geometric series sums to the standard discounted form:

$$V_\gamma^{\text{div,risk}} - V_\gamma^{D,\text{risk}} = \frac{r_{\text{ext}} + \Delta_{\text{risk}}}{1-\gamma} - \Delta_0 \tag{16}$$

where $\Delta_{\text{risk}} = \mathbb{E}[\mathcal{R}_{\text{dep}}^{\text{single}}] - \mathbb{E}[\mathcal{R}_{\text{dep}}^{\text{mixed}}] > 0$ is the per-step risk gap from Proposition 5. The threshold for diversity dominance becomes:

$$\gamma_{\text{risk}}^* = 1 - \frac{r_{\text{ext}} + \Delta_{\text{risk}}}{\Delta_0} \tag{17}$$

The risk gap Δ_{risk} enters additively with r_{ext}, effectively increasing the external contribution by the per-step risk mitigation value of cross-substrate diversity. When the risk gap is non-stationary (e.g., the threat landscape evolves, or the substrate partition itself changes during the trajectory), the closed-form $(1-\gamma)^{-1}$ denominator becomes an approximation, and the exact value is $\sum_{t=0}^{\infty} \gamma^t \cdot \Delta_{\text{risk}}(t)$ computed from a time-varying risk profile. The trust model's standard non-stationarity assumption (agents and threats evolve on timescales slower than the discount horizon) makes the stationary approximation adequate for planning; rapidly-varying threat landscapes require a finer-grained value-iteration calculation instead of the closed form.

Quantitative impact. For a population with $p_{\text{adv}} = 0.01$ (1% per-step probability of substrate-targeting event) and cascade fraction $c = 0.5$ (half the targeted substrate's capabilities are lost in a typical event), $\Delta_{\text{risk}} = 0.01 \times \Delta_{\text{div}} \times 0.5$ where Δ_{div} is the substrate diversification advantage from Proposition 4 of Lasser [2026b]. If Δ_{div} is 10% of $\text{vol}_{\text{P}}(G)$, then $\Delta_{\text{risk}} \approx 0.0005 \times \text{vol}_{\text{P}}(G)$: the per-step risk cost of single-substrate concentration is 0.05% of total volume per step. Over the full planning horizon ($1/(1-\gamma)$ steps), this compounds to $0.05\% \times \text{vol}_{\text{P}}(G)/(1-\gamma)$, which for $\gamma = 0.99$ is $5\% \times \text{vol}_{\text{P}}(G)$, which is substantial.

Sensitivity analysis. Table 2 shows how $\Delta_{\text{risk}}/\text{vol}_{\text{P}}(G)$ varies over plausible parameter ranges. The per-step risk gap scales linearly with p_{adv} and c, and the compounded effect over the planning horizon is $\Delta_{\text{risk}}/(1-\gamma)$. The 80% upper-bound entry corresponds to a 5% adversarial rate with high contagion ($c = 0.8$) and substantial divergence ($\Delta_{\text{div}} = 0.20$), representing a compromised-substrate scenario rather than a typical operating point. A more representative regime is $p_{\text{adv}} = 0.01$, $c = 0.5$, $\Delta_{\text{div}} = 0.10$, which yields a 5% compounded gap, material but containable by the verification architecture of Section 3.4.

Interaction with the horizon paper's cooperative-novelty result. The per-step risk gap Δ_{risk} is amplified by the cross-substrate cooperative capabilities analyzed in Lasser [2026b]: any cooperative capability requiring agents on both substrates is destroyed when either substrate is targeted, so $|\Delta\text{vol}_{\text{P}}(\mathcal{S}_i)|$ includes those cooperative losses on top of the individual-capability loss on $\mathcal{S}_i$. Under the horizon paper's cooperative-novelty result, these cross-substrate cooperative capabilities are the primary growth channel, so their destruction represents a disproportionate fraction of the total contraction during a substrate-targeting event. The cooperative novelty and dependency-risk analyses reinforce each other: domination is costly both because it eliminates the cross-substrate cooperative growth channel *and* because it concentrates adversarial exposure to the single surviving substrate [Lasser, 2026b].

p_{adv}	c	$\Delta_{\text{div}}/\text{vol}_{\text{P}}$	$\Delta_{\text{risk}}/\text{vol}_{\text{P}}$ (per step)	Compounded at $\gamma = 0.99$
0.001	0.3	0.10	3×10^{-5}	0.3%
0.01	0.3	0.10	3×10^{-4}	3.0%
0.01	0.5	0.10	5×10^{-4}	5.0%
0.01	0.5	0.20	1×10^{-3}	10.0%
0.01	0.8	0.10	8×10^{-4}	8.0%
0.05	0.5	0.10	2.5×10^{-3}	25.0%
0.05	0.8	0.20	8×10^{-3}	80.0%

Table 2: Sensitivity of the per-step dependency-risk gap $\Delta_{\text{risk}} = p_{\text{adv}} \cdot \Delta_{\text{div}} \cdot c$ and its compounded effect over the planning horizon at $\gamma = 0.99$ ($1/(1-\gamma) = 100$ effective steps). The compounded percentage is $\Delta_{\text{risk}}/[(1-\gamma) \cdot \text{vol}_{\text{P}}(G)] \times 100$.

Summary. The adversarial balance condition $c > \max_i(f_i \cdot c_i)$ converts the companion paper's minimax bound into an expected per-step cost Δ_{risk} that compounds as $\Delta_{\text{risk}}/(1-\gamma)$ in the discounted value function, making dependency risk a first-class term in the horizon calculation rather than a one-time structural penalty. The condition is explicit about contagion dynamics and inverts (Corollary 5.1) into a quantitative bound on the required substrate count: $m > c_{\max}/c$ for balanced partitions, or $f_{\max} < c/c_{\max}$ for arbitrary partitions. This sharpens the minimax $m \geq 2$ into a measurement-driven requirement that scales with contagion intensity. The regime where the condition fails (unbalanced partitions under contagion amplification) is a regime where diversification genuinely does not help, which the trust model can detect and route around.

6 Discussion

6.1 Limitations

SCM structure learning as channel optimization, not prerequisite. The causal attribution channel of Section 2 requires the actor to know or learn the causal DAG over agent actions and capability changes. Standard observational structure learning [Spirtes et al., 2000] is inadequate for the GFM setting for three distinct reasons. First, the data-generating process is adversarial: scorpions can strategically align their actions with observed noise to mask their causal role, so classical conditional-independence tests can be gamed by an adversary who knows which tests will be run. Standard structure learning does not have a threat model. Second, the agent cardinality is non-fixed: new agents can appear, existing agents can delegate, and a group can coordinate and behave as a single effective actor, so the variable set itself evolves. Latent-variable discovery methods (FCI, RFCI) assume a fixed variable cardinality and do not naturally handle this. Third, the actor cannot typically intervene on other agents: interventional causal discovery relies on actual interventions, which a GFM actor can simulate inside its SCM but cannot enforce in the world. This rules out the entire class of interventional discovery methods and leaves only observational methods, which are strictly weaker under the adversarial conditions above.

Because causal attribution is one channel among several (Section 2.4), and capability gating via the poset of Lasser [2026c] and Lasser [2026b] is the strongest and most difficult to fake channel (relative to a trusted, sufficiently complete capability graph), structure learning becomes an *optimization target* for one channel's contribution rather than a *prerequisite* for attribution at all. When the SCM is poorly learned, the causal channel degrades, the other channels are not directly degraded by the SCM misspecification, and the multi-channel aggregate reweights toward the structurally-grounded channels. The system does not collapse; it falls back on the trust and capability machinery of Lasser [2026a], Lasser [2026c], and Lasser [2026b].

The structural assumptions underlying Proposition 1(c) are partially testable through a residual-covariance diagnostic: if $\text{Cov}(\text{CA}(a_j), \text{CA}(a_k) \mid G_t)$ exceeds the level the current SCM predicts, the SCM is misspecified (possible causes include a missing edge, latent common causes, exogenous-orthogonality failure, non-additivity, or nonstationarity), and the causal channel's reliability should degrade accordingly. Integrating such a diagnostic into a calibrated SCM-confidence scalar $c_{\mathcal{M}}(t) \in [0,1]$ that weights the causal channel's contribution to the multi-channel aggregate, so that declining sufficiency down-weights causal attribution without burning specific agents'

trust, is an architectural direction we identify but do not formalize here. The full adversarial-robust structure-learning problem, including the calibration procedure and its interaction with the behavioral and capability channels, is deferred to a companion paper.

Coordinated multi-agent attacks and non-additive interactions. The one-at-a-time causal attribution $\mathrm{CA}(a_j, t)$ (Definition 2) decomposes the total $\mathrm{vol_P}$-change exactly only under additive SCM structure (Equation (2)). In multi-agent populations, coordinated attacks produce non-additive interactions: an attack requiring m agents to cooperate has $\Delta\mathrm{vol_P}$ that depends on their joint action, and the one-at-a-time intervention $\mathrm{do}(\pi_j = \mathrm{skip})$ produces non-unique individual attributions: under pure complementarities ($\Delta\mathrm{vol_P} = \pi_1 \cdot \pi_2$), marginal attribution may over-count or double-count each coordinator's contribution; under other non-additive structures the attribution allocates the interaction term arbitrarily across agents. The fundamental issue is that the one-at-a-time decomposition is not well-defined when the structural equation is non-additive, and no single-agent marginal faithfully represents the joint effect. The exact fix is Shapley-style attribution [Shapley, 1953] that averages over all orderings of agent inclusion; developing this within the GFM SCM, including its computational cost (exponential in coalition size without structure-exploiting approximations) and its interaction with the SPRT detection rule, is deferred to a companion paper on joint attribution. In the interim, the multi-channel attribution architecture (Section 2.4) provides the structural defense: coordinated attacks that evade the causal channel are still detectable through capability gating (which constrains the set of agents that *could* have caused the contraction regardless of interaction structure) and through behavioral prediction residuals, which detect individual coordinators whose actions diverge from their predictive models. The causal channel's failure under non-additivity degrades one channel's contribution; the multi-channel aggregate reweights toward the structurally-grounded channels, as designed.

Risk-trust residual dependence on the actor's model. The pathway-conditional structural verification residual (Definition 3) measures divergence between the claim and the actor's own assessment of the claimed pathway. If the actor's world model is itself poorly calibrated, the residual reflects model error rather than claim quality. An agent whose risk assessments are accurate but disagree with a miscalibrated actor will accumulate high $\|r^{\mathrm{risk}}\|^2$ and concentrate around low ${T^{\mathrm{risk}}}^*_j$, producing a false negative. This is mitigated by the self-trust mechanism of Lasser [2026a]: an actor with a poorly calibrated model has low T_s and high learning rate, which accelerates model correction. But during the correction period, risk-trust assessments may be unreliable.

Exogenous orthogonality as the most fragile assumption. Of the three structural assumptions underlying Proposition 1(c), exogenous orthogonality is the most likely to fail in deployment. Shared environmental shocks (power grid state, information broadcasts, market movements, weather, time-of-day effects on a shared population) induce covariance between agents' contributions even under a correctly specified causal DAG. Conditioning on G_t does not remove time-localized shocks that affect multiple agents' actions simultaneously. When the assumption fails, the variance reduction $\sigma^2_{\mathrm{do}} \leq \sigma^2$ no longer holds and the causal detector may perform worse than the correlational baseline (as acknowledged in Proposition 1(c)). The residual-covariance diagnostic described above provides detection: empirically elevated $\mathrm{Cov}(\mathrm{CA}(a_j), \mathrm{CA}(a_k) \mid G_t)$ signals orthogonality failure, and the SCM-confidence scalar $c_{\mathcal{M}}(t)$ should down-weight the causal channel accordingly. The multi-channel aggregate then shifts weight toward capability gating and behavioral channels, which do not depend on exogenous orthogonality. Importantly, orthogonality failure is *detectable* rather than silent. The residual covariance term $\mathrm{Cov}(\mathrm{CA}(a_j), \mathrm{CA}(a_k) \mid G_t)$ is an estimable quantity: it can be estimated from repeated observations at similar capability-graph states, using the same attribution data the actor already collects for the SPRT detector. When the estimated covariance exceeds the level predicted under exogenous orthogonality, the assumption has failed and the failure is empirically visible. This turns the most fragile of the three structural assumptions into the most monitorable one: a better epistemic position than "fragile but unobservable." The natural integration point is the SCM-confidence scalar $c_{\mathcal{M}}(t)$ described above: elevated residual covariance is a direct input to $c_{\mathcal{M}}$ updating, triggering the multi-channel aggregate to down-weight the causal channel before the degraded variance reduction produces misleading detection results. Characterizing the populations and environments in which exogenous orthogonality approximately holds — and deriving a partial-independence correction for the variance reduction when it does not — is an open problem.

Log-odds aggregation and reporter correlation. The trust-weighted log-linear opinion pool (Definition 5) recovers Bayesian updating in the no-tempering limit under conditional independence of agents' observations given the true risk level. If reporters' assessments are correlated (they share information sources, or one reporter's assessment influences others), the aggregation overweights the shared signal. In practice, partial independence is the norm: agents with different observation channels and different world models produce partially independent assessments, and the log-linear pool is an approximation whose error grows with the correlation. Characterizing the approximation error under partial independence, and deriving a correction term from a reporter-network model, is deferred to future work.

Adversarial parameter exposure. Three of this paper's mechanisms share a common vulnerability: the adversary can observe and exploit fixed defense parameters.

Verification-gate denial of service (Section 3.4). The salience gate is cheap (aggregate and threshold); the verification gate is computationally expensive (forward causal evaluation on the actor's own SCM). An adversary who floods the actor with structurally plausible false risk claims can exhaust the verification budget without any single claim bypassing independent verification. The two-gate architecture prevents coordinated amplification of false claims from reaching *action*, but leaves the actor vulnerable to attention exhaustion via structurally plausible decoys. The defense direction is prioritization: the verification gate should process candidates in order of ${T^{\text{risk}}}_j$-weighted expected contraction magnitude, so high-trust reporters' claims are verified first and low-trust flooding imposes only queue delay, not budget exhaustion. Formalizing the verification-gate capacity model and its interaction with the salience threshold is deferred.

The adversary knows ϵ (Section 4). The SPRT design parameter ϵ sets the minimum detectable contraction magnitude. A well-informed adversary who can infer ϵ from the actor's behavior (by probing with small contractions and observing flag rates) will calibrate $|\mu_j|$ just below ϵ and sit in the undetected regime permanently. The evasion-bandwidth bound's "undetected contraction rate bounded by $|\mu_j|$" is only informative if $|\mu_j|$ is bounded by something other than the design parameter itself. The defense direction is to treat ϵ as a nuisance parameter with an informative prior $p(\epsilon)$ encoding the actor's belief about the adversary's calibration capability, and to marginalize. Because the evasion-bandwidth bound $\tau_{\text{adapt}}(\epsilon) = O\big(\sigma^2_{\text{do}}/[\epsilon(|\mu_j| - \epsilon/2)]\big)$ is highly nonlinear in ϵ: small-ϵ draws produce very long detection times. The marginalized bound

$$\mathbb{E}_p[\tau_{\text{adapt}}(\epsilon)] = \int_{\epsilon_{\min}}^{\epsilon_{\max}} \frac{\sigma^2_{\text{do}}}{\epsilon\,(|\mu_j| - \epsilon/2)}\, p(\epsilon)\, d\epsilon \tag{18}$$

is finite for any prior supported on $[\epsilon_{\min}, \epsilon_{\max}]$ with $0 < \epsilon_{\min}$ and $\epsilon_{\max} < 2|\mu_j|$ (the detection regime where the SPRT drift is positive; the upper cutoff requires the defender to have prior beliefs about plausible scorpion contraction magnitudes, since $|\mu_j|$ is precisely what the SPRT is trying to detect), and the tail bound $\Pr_p(\tau_{\text{adapt}} > T)$ is the prior mass on the ϵ-region where $\tau_{\text{adapt}}(\epsilon) > T$, a set that shrinks as T grows. A log-uniform prior $p(\epsilon) \propto 1/\epsilon$ on $[\epsilon_{\min}, \epsilon_{\max}]$ is a natural special case: it is the maximum-entropy prior over $\log \epsilon$ on bounded support [Jaynes, 1957], equivalently the scale-invariant maxent prior under the multiplicative invariance group [Jaynes, 1968], corresponding to maximum uncertainty about the scale of the adversary's evasion margin. A caveat: log-uniform is maxent over the parameter ϵ, but "maxent over the parameter" is not the same as "minimax-optimal defender distribution against an adversary minimizing expected detection time." The game-theoretic problem of finding the defender's minimax-optimal randomization is more involved and is not developed here; log-uniform is a defensible default under maximum uncertainty, not a proven optimum. A second, independent justification comes from the adversary's own learning problem: to optimize against deployed randomization, the adversary must estimate $p(\epsilon)$ from flag observations across epochs, and the information cost of learning a distribution scales with its entropy. Log-uniform maximizes this entropy on bounded log-support, making it near-optimal against a Bayesian adversary trying to infer $p(\epsilon)$ from observations. The defense requires the actor to *deploy* randomized ϵ, drawing from $p(\epsilon)$ each epoch and keeping the draw private, rather than merely marginalize analytically. Under deployed randomization, the adversary's optimal evasion strategy must be computed against the distribution $p(\epsilon)$ rather than a known point estimate, following the minimax-theorem argument for mixed strategies against an optimizing adversary [von Neumann, 1928]. Inferring the distribution itself requires observing the actor's flagging behavior across multiple epochs to estimate $p(\epsilon \mid \text{flags})$. This is strictly harder than inverting a known ϵ. The general principle is that fixed, known parameters are the exploit surface; replacing a point ϵ with a privately-

drawn $p(\epsilon)$ is one instance of the randomization-family defense that applies to all three adversarial surfaces identified above.

The defense is not free. Under log-uniform $p(\epsilon)$ the integrand of Equation (18) is dominated by the $1/\epsilon^2$ tail near $\epsilon_{\min}$, giving

$$\mathbb{E}_p[\bar{\tau}] \;\approx\; \frac{\sigma_{\text{do}}^2}{|\mu_j|\,\epsilon_{\min}\,\log(\epsilon_{\max}/\epsilon_{\min})} \tag{19}$$

for $\epsilon_{\min} \ll |\mu_j|$. Relative to fixed $\epsilon = \epsilon_{\max}$ (the least-delay choice), the expected detection time expands by a factor $\epsilon_{\max}/[\epsilon_{\min}\,\log(\epsilon_{\max}/\epsilon_{\min})]$. For plausible operational values $\epsilon_{\min} = 0.01$, $\epsilon_{\max} = 0.1$, this factor is approximately 4.3, material but bounded. The parameter $\epsilon_{\min}$ is thus a first-class policy variable: reducing it extends the range of detectable contractions but proportionally expands expected detection delay. The Markov-inequality bound (Proposition 4(a)) remains valid under randomization, but the reference timescale $\bar{\tau}$ inflates by the same factor, so the high-probability detection guarantee degrades proportionally. The defender trades evasion-bandwidth robustness against a quantifiable expansion in expected detection time.

T_j / $T^{\text{risk}}{}_j$ *independence* (Section 3). The behavioral trust T_j and risk-trust $T^{\text{risk}}{}_j$ evolve independently by design, optimizing for the case of honest specialists (good capability reporter, poor risk forecaster, or vice versa). A strategic reporter can exploit the independence: operating as a known-unreliable capability reporter (low T_j) while generating structurally plausible risk claims that accumulate high $T^{\text{risk}}{}_j$, thereby driving the verification-gate queue despite a poor behavioral track record. A two-sided clamp on risk-trust addresses both failure modes:

$$T_{\min}^{\text{risk}} \;\le\; T^{\text{risk}}{}_j \;\le\; \max\bigl(f(T_j),\, T_{\min}^{\text{risk}}\bigr) \tag{20}$$

for some monotone f and a design parameter $T_{\min}^{\text{risk}} > 0$. The lower bound is the floor: after each EWMA update the realized $T^{\text{risk}}{}_j$ is clipped from below at $T_{\min}^{\text{risk}}$, so neither an adversary driving T_j toward zero nor a run of correlated environmental noise on an honest agent can push risk-trust below the floor, preventing self-reinforcing exclusion. The upper bound is the coupling cap: when behavioral trust is high enough that $f(T_j) \ge T_{\min}^{\text{risk}}$, the cap limits how far risk-trust can exceed the behaviorally-warranted level, closing the low-T_j exploit. This is structurally analogous to the adversarial-balance condition $c > \max_i(f_i \cdot c_i)$ of Section 5: a structural constraint that weakens both failure modes without claiming to resolve either tightly. The parameter $T_{\min}^{\text{risk}}$ itself carries a trade-off: too large re-admits the low-T_j exploit (the cap lifts so far that risk-trust can remain high despite poor behavioral trust); too small approaches the unbounded case. The two-sided form does not eliminate the robustness-versus-fairness tension; it bounds its worst-case expression on both sides.

The unifying theme is that fixed, observable parameters of the defense create an adversarial surface. The structural defense direction across all three is *randomization and coupling*: randomize the parameters the adversary needs to know (gate priority, ϵ, channel weights) and couple the trust channels so that gaming one does not leave the others unaffected.

Shared-prior π_0 as an adversarial surface. The log-odds aggregation (Definition 5) requires every reporter to calibrate against a shared prior π_0, and the paper notes this requires a coordination protocol. The choice of π_0 is itself a manipulation surface: a reporter who can influence π_0 (by proposing it, by coordinating on a biased value, or by exploiting a protocol weakness) can tilt the aggregation without submitting a false report. This paper does not develop a prior-robust solution; the adversarial robustness of π_0 establishment is an open problem that interacts with the adversarial parameter exposure pattern above. We note that per-agent prior randomization, the natural analog of the randomized-ϵ defense, would break the shared-prior condition on which the log-odds aggregation's Bayes-optimality depends (Definition 5), so prior-robustness likely requires re-deriving the aggregation rule under heterogeneous priors rather than perturbing π_0.

Structural verification residual norm weights. The structural verification residual (Definition 3) uses the Euclidean norm on a two-dimensional vector (probability difference, normalized magnitude difference), which implicitly treats both components as equally important. A probability error of 0.1 and a magnitude error of $0.1 \cdot \text{vol}_{\text{P}}(G_t)$ enter symmetrically into the trust update but represent different kinds of miscalibration. A weighted norm with weights tuned to the actor's utility over probability-versus-magnitude errors would be more principled, but would require the actor to

specify a preference ordering over error types that the framework does not assume. The equal-weight Euclidean norm is a neutral default; domain-specific actors with a preference for probability accuracy over magnitude accuracy (or vice versa) should adjust the norm accordingly.

Cascade confinement and cross-substrate propagation. The dependency-risk analysis (Section 5) assumes cascade fractions c_i measure only within-substrate propagation, with cross-substrate cascades outside the scope of the current bound. In the motivating threat classes (memetic attacks, financial contagion, supply-chain disruption), cross-substrate propagation is exactly the concerning pathway: a memetic attack on silicon agents that triggers panic-driven behavior in biological agents through inter-substrate communication channels is precisely a cross-substrate cascade, and the adversarial balance condition does not cover it. Modeling cross-substrate cascade propagation requires extending the SCM's structural equations to include inter-substrate edges, which produces a coupled cascade model whose analysis is substantially more complex than the within-substrate case.

This gap is also a coupling point with the structural diversification guarantees developed elsewhere in the framework. Lasser [2026b] establishes the minimax result that $m \geq 2$ substrate types strictly dominate $m = 1$; Corollary 5.1 of the present paper sharpens this to $m > c_{\max}/c$ under the contagion model of Section 5. Both results assume that substrate-exclusive capabilities do not propagate across the partition boundary. In the limit of zero cross-substrate leakage this assumption holds and the bounds are model-level guarantees under the stated assumptions; at nonzero leakage rate, the bounds become approximations whose error depends on the magnitude of inter-substrate information flow. The theorems remain valid; the proofs are correct for the model they assume. What changes at nonzero leakage is the empirical condition under which the guarantees are tight: cross-substrate informational channels (shared training data, memetic propagation, supply-chain coupling) must be small relative to the capability differences that define the partition. The forward direction is either modeling inter-substrate edges explicitly within this paper's SCM framework or characterizing the leakage rate empirically so that the approximation error can be bounded.

Epoch-boundary gaming. Proposition 4's SPRT bounds are per-epoch, and an epoch boundary resets the SPRT statistic. The epoch-declaration protocol is a structural vulnerability: if the actor declares epoch boundaries based on detected non-stationarity, an adversary can trigger epoch resets by inducing structural transitions. If epoch boundaries are time-based (every τ steps), an adversary can synchronize strategy shifts to epoch boundaries so the SPRT never accumulates sufficient evidence. The defensive direction is a hybrid protocol: time-based epoch boundaries provide a baseline accumulation window that the adversary cannot shorten, while non-stationarity detection triggers auxiliary monitoring (e.g., a parallel SPRT with carry-over statistics) rather than resetting the primary accumulator. This mitigates the most direct epoch-manipulation attacks but does not eliminate the vulnerability, since any fixed-window protocol still gives the adversary a known accumulation budget to calibrate against. The hybrid construction does, however, make the adversary's inversion problem harder: the detection trigger is a function of SPRT threshold accumulation, which depends on σ^2_{do}, ϵ, and the observed statistics sequence; the observed sequence is itself a function of both the actor's prior monitoring history and the adversary's current strategy. The joint distribution over (time-based budget phase, detection-trigger sensitivity) is therefore harder for the adversary to invert than the pure time-based case, unless it can reconstruct the actor's internal monitoring state from observable history. The residual vulnerability is that an adversary with high-bandwidth observation of the actor's flag decisions can estimate this joint distribution over many epochs, gradually reducing its entropy. "Hybrid" thus narrows the adversarial surface rather than closing it, consistent with the adversarial-parameter-exposure theme: fixed parameters leak to observation; hybrid parameters leak more slowly. Formalizing this hybrid protocol and its adversarial robustness is deferred.

7 Conclusion

The preceding GFM papers define what to maximize ($\text{vol}_{\text{P}}(G)$), prove structural safety properties, and establish anti-monopolar pressure. This paper addresses how the actor *reasons about other agents*: providing a causal model that upgrades scorpion detection and risk evaluation from statistical to causal identification and sharpens the remaining evaluation mechanisms.

Causal attribution (Section 2) upgrades the foundational paper's correlational scorpion detection to causal identification through do-calculus counterfactuals. The convergence advantage over corre-

lational detection is proportional to confounding variance, which is substantial in populations with many simultaneously acting agents. Non-stationary scorpion bounds (Section 4) characterize the evasion bandwidth: how fast a scorpion must adapt to remain undetected, and the maximum undetected contraction rate.

Risk-trust dynamics (Section 3) provide formal EWMA update equations for ${T^{\text{risk}}}_j$, with a two-gate architecture separating attention allocation (log-odds aggregation) from action decisions (the actor's own causal verification). This prevents coordinated report-bombing from being amplified into action while preserving sensitivity to genuine independent agreement. Probabilistic dependency risk (Section 5) converts minimax substrate bounds into per-step expected costs that compound at the same rate as growth, providing the quantitative risk model needed for risk-adjusted planning.

Together with the foundational framework [Lasser, 2026a], the computable poset [Lasser, 2026c], and the horizon-aware objective [Lasser, 2026b], these results narrow the gap between the GFM framework's structural guarantees and the mechanisms a concrete agent would need to realize them. The results here are conditional on seven load-bearing assumptions: causal sufficiency, additive SCM structure, and exogenous orthogonality (Proposition 1(c)); conditional independence of reporters given a shared prior (Definition 5); cascade confinement within substrates and the adversarial balance condition (Proposition 5); and a stationary-epoch adversary model (Proposition 4). Each assumption is explicit, falsifiable, and load-bearing. As implementation machinery, these results do not affect the framework's structural guarantees: a flaw in any mechanism here degrades the actor's evaluation accuracy but does not compromise the self-balancing, anti-monopolar, or structural alignment properties established in Papers 1–3 [Lasser, 2026a,c,b]. The remaining work is both theoretical (relaxing each assumption and characterizing what happens when it fails) and empirical: instantiating these computations in a concrete environment and checking whether the predicted alignment properties hold under real conditions.

Author Contributions

Teague Lasser identified the statistical-to-causal upgrade as the unifying contribution, proposed the log-odds risk aggregation and the two-gate attention/verification architecture, directed all revisions, and made final editorial decisions. Responsible for the paper's intellectual direction and all claims made.

Claude Opus 4.6 (Anthropic) drafted the formal exposition: the structural causal model for capability dynamics, the do-calculus contraction attribution, the L^2 risk-trust EWMA dynamics, the pathway-conditional structural verification residual, the SPRT-based causal scorpion detection and evasion bandwidth bounds, and the probabilistic dependency-risk model with the adversarial balance condition.

GPT 5.4 (OpenAI) served as technical reviewer, identifying formal gaps in the risk-trust convergence claims, the causal attribution assumptions, the log-odds aggregation's cross-paper type consistency, and the dependency-risk quantitative analysis — in particular the need to replace the pointwise cascade-fraction assumption with the adversarial balance condition that admits contagion dynamics.

Transparency note. Both AI systems operated as tools under human direction. Neither system has continuity across sessions, cannot take responsibility for the work in the sense required by most venue authorship policies, and cannot respond to reviewer queries independently. They are listed as authors to accurately represent their contributions to the intellectual content of the paper, not to claim that they meet all criteria of traditional academic authorship. The corresponding author for all inquiries is Teague Lasser.

References

Roy M Anderson and Robert M May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1991.

David J Bianco. The pyramid of pain. `http://detect-respond.blogspot.com/2013/03/the-pyramid-of-pain.html`, 2013. Blog post, Enterprise Detection & Response.

Sergio Caltagirone, Andrew Pendergast, and Christopher Betz. The diamond model of intrusion analysis. Technical report, Center for Cyber Intelligence Analysis and Threat Research, 2013.

Christian Genest and James V Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986.

Richards J Heuer. *Psychology of Intelligence Analysis*. Center for the Study of Intelligence, Central Intelligence Agency, 1999.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

Edwin T Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4(3): 227–241, 1968.

Teague Lasser. Goal-frontier maximizers are civilization aligned. `https://teague.info/papers/gfm/`, 2026a.

Teague Lasser. Horizon-aware goal-frontier maximization excludes singleton behavior. `https://teague.info/papers/horizon/`, 2026b.

Teague Lasser. Computable goal frontiers and the gradient toward civilization-building. `https://teague.info/papers/poset/`, 2026c.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Lloyd S Shapley. A value for n-person games. In H. W. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, 1953.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

Blake E Strom, Andy Applebaum, Doug P Miller, Kathryn C Nickels, Adam G Pennington, and Cody B Thomas. MITRE ATT&CK: Design and philosophy. Technical Report MP180360R1, The MITRE Corporation, 2018.

John von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16 (2):117–186, 1945.

A Proofs

Propositions 1 (Convergence Advantage), 2 (Risk-Trust Convergence), 3 (Causal Scorpion Detection), and 5 (Expected Risk Comparison) are proved in the main text. This appendix provides the remaining proofs.

A.1 Proof of Proposition 4 (Evasion Bandwidth)

Proof. Within each strategy epoch of length τ_{adapt}, the scorpion's behavior is stationary and the causal-attribution sequence $\{\text{CA}(a_j, t)\}$ is i.i.d. with mean $\mu_j < 0$ and variance σ^2_{do} (the same assumption as Proposition 3). The SPRT decision rule tests the least-favorable simple pair $H_0 : \mu_j = 0$ versus $H_1 : \mu_j = -\epsilon$ under the Gaussian model. The per-observation expected log-likelihood ratio under the true mean μ_j is $\mathbb{E}_{\mu_j}[\log(p_{H_1}/p_{H_0})] = \epsilon(|\mu_j| - \epsilon/2)/\sigma^2_{\text{do}}$. By Wald's identity [Wald, 1945], the expected detection time within a stationary epoch with actual mean μ_j is:

$$\bar{\tau}(\mu_j) = \mathbb{E}[\tau_{\text{detect}}] = \frac{\sigma^2_{\text{do}}}{\epsilon(|\mu_j| - \epsilon/2)} \cdot \log\frac{1-\beta_{\text{err}}}{\alpha_{\text{err}}}.$$

Part (a): detection. By Markov's inequality, $P(\tau_{\text{detect}} > k\bar{\tau}(\mu_j)) \leq 1/k$ for any $k > 1$. When $\tau_{\text{adapt}} \geq k\bar{\tau}(\mu_j)$, the epoch provides enough observations that the SPRT crosses the detection threshold with probability at least $1 - 1/k$.

Part (b): evasion. On an epoch boundary the null and alternative hypotheses change (the new epoch has a new $(\mu_j, \sigma^2_{\text{do}})$), so the SPRT statistic is reset to zero: evidence accumulated under one pair of hypotheses is not informative about a different pair. The auxiliary EWMA monitors (Equations (10)–(11)) retain a fraction $\alpha^{\tau_{\text{adapt}}}$ of prior evidence as their output, but the SPRT decision rule itself is memoryless across regime changes. If $\tau_{\text{adapt}} < \bar{\tau}(\mu_j)$, the expected accumulated evidence at the epoch boundary is below the SPRT threshold: the evidence accumulation rate is insufficient for reliable detection within the epoch.

The maximum undetected contraction per epoch is $|\mu_j| \cdot \tau_{\text{adapt}}$. For a scorpion to evade detection within an epoch, it must have $\tau_{\text{adapt}} < \bar{\tau}(\mu_j)$. The per-step undetected contraction is bounded by $|\mu_j|$, and the scorpion can sustain this for at most τ_{adapt} steps before changing strategy. At the design point $|\mu_j| = \epsilon$, the per-step contraction is at most ϵ. The time-averaged undetected contraction rate is therefore at most $|\mu_j|$, achievable only by a scorpion that never needs to pause for strategy adaptation. ■