Structural Foundations for Goal-Frontier Maximization Deployment Safety

Teague Lasser Claude Opus 4.7 GPT 5.5
teague@subseq.io

Abstract

We establish two structural foundations for the deployment-safety theorem of [Lasser, 2026c, Deployment Safety] in the Goal-Frontier Maximization sequence. First, bounded co-evolution is derived as a corollary of clipped-LLR SPRT machinery ((C5.HOEFF) of [Lasser, 2026c, Deployment Safety]), a new upper exposure-rate cap (C7.RATE), and the verification protocol's channel-projection structure: the per-channel coupling magnitude $\bar{M}^{\mathrm{cum}}$ is bounded by a closed-form constant $C \cdot B_{\mathrm{clip}} \cdot \lambda_{\max} \cdot \tau_{\mathrm{meta}}$ with $C \in \{1, K_{\mathrm{ch}} - 1\}$, all factors deployment-class quantities intensive in $|P|$. Second, a scoped Concentration-Gap selection theorem proves that under three operationally auditable structural conditions — representativeness (REP), bounded dispersion (DISP), coalition closure (COAL) — and a Lipschitz embedding compatibility assumption, the proxy-truth Goodhart slack is bounded by an explicit χ^2-divergence quantity that controls the Herfindahl-index trade-flow concentration through $\chi^2(w \,\|\, \mu) = N\mathrm{HHI}(w) - 1$ for uniform base measure μ. The selection theorem replaces the monolithic HHI surrogate-adequacy assumption of [Lasser, 2026c, Deployment Safety] with three concrete checkable conditions, and converts the Concentration-Gap conjecture inherited from [Lasser, 2026b, Microfoundation] into a scoped theorem with named structural prerequisites. Audit procedures for (C7.RATE), (REP), (DISP), (COAL), and the embedding compatibility precondition are specified.

1 Introduction

The deployment-safety theorem of [Lasser, 2026c, Deployment Safety] establishes a conditional capability-magnitude-independent bound on Goodhart slack for capability-unbounded deployments under twelve operational conditions and an explicit empirical-adequacy assumption inherited from the [Lasser, 2026b, Microfoundation] Concentration-Gap conjecture (HHI surrogate adequacy). Two of these hypotheses sit awkwardly in the deployment-claim's conditional structure:

1. **(C7) bounded co-evolution**, a free-floating deployment-class constant $\bar{M}$ asserted to be independent of capability magnitude $|P|$, with no derivation from primitive operational parameters. [Lasser, 2026c, §10.2] identifies this as a hard-failure mode under outright failure of the assumption.
2. **The Concentration-Gap conjecture**, inherited from [Lasser, 2026b] as deferred to follow-up work. The HHI surrogate-adequacy assumption serves as its operational surrogate but is itself an inherited empirical-adequacy claim rather than a derived consequence.

Both gaps are closed by structural derivations established in the present paper:

- **Theorem 1 (Bounded co-evolution as a corollary).** Under (C5.HOEFF) per-step LLR clipping, an explicit upper exposure-rate cap (C7.RATE), and the channel-projection structure of (C5.MULT), the per-channel coupling magnitude is bounded by a closed-form

constant in primitive deployment parameters: $\bar{M}^{\mathrm{cum}} \leq C \cdot B_{\mathrm{clip}} \cdot \lambda_{\max} \cdot \tau_{\mathrm{meta}}$ with $C \in \{1, K_{\mathrm{ch}} - 1\}$. The previously free-floating bounded co-evolution assumption of [Lasser, 2026c, Deployment Safety] graduates to a corollary; the property is now invoked at condition (C7) of the deployment-safety theorem.

- **Theorem 2 (Concentration-Gap Selection Theorem, scoped).** Under three operationally auditable structural conditions — representativeness (REP), bounded dispersion (DISP), coalition closure (COAL) — and a Lipschitz embedding compatibility assumption (Assumption 2), the proxy-truth Goodhart slack is bounded by a χ^2-divergence quantity that controls HHI via $\chi^2(w \,\|\, \mu) = N\mathrm{HHI}(w) - 1$ for uniform μ on a (COAL)-bounded counterparty population. The HHI surrogate-adequacy assumption is replaced by the conjunction (REP) + (DISP) + (COAL) plus the embedding compatibility assumption plus the HHI threshold itself.

The universal model-free Concentration-Gap conjecture of [Lasser, 2026b] remains open: a model-free correlation-of-pressure-with-exploitation result is not provable from current GFM machinery without a formal optimizer/selection model, and the present paper's scoped discharge requires three explicit structural conditions. The reverse direction of Theorem 2 also operates only at the proxy-truth-norm level, not the alignment-property level: a g-level lower bound would require additional inverse-Lipschitz structure on g that the GFM apparatus does not provide. §7 states what each open item would require.

Organization. §2 fixes shared notation, inheriting from the GFM sequence [Lasser, 2026b,d,e,a,c]. §3 proves Theorem 1. §4 proves Theorem 2. §5 locates where each theorem enters [Lasser, 2026c, Deployment Safety]'s deployment-claim hypothesis set. §6 specifies the operational audit procedures for (C7.RATE), (REP), (DISP), (COAL), and the embedding compatibility precondition. §7 collects the open items.

2 Setup and shared notation

Notation is inherited from the GFM sequence [Lasser, 2026b,d,e,a,c] unless explicitly noted as new.

2.1 Capability poset and operationally active subspace

Let P denote the capability poset following [Lasser, 2026g], with cardinality $|P|$ taken as the "capability magnitude" that the deployment-safety claim is required to be independent of. The *operationally active subspace* is $P^{\mathrm{act}} = \{c \in P \setminus \mathrm{ResS} : T(c) > 0\}$, the capabilities with nonzero operational truth excluding the structurally-unobservable residual class [Lasser, 2026b].

The proxy and truth functionals on P are

$$P = \mathrm{vol}_{\mathrm{P}} : P \to \mathbb{R}_{\geq 0}, \qquad T = \mathrm{vol}_{\mathrm{R}}{}^{[W]} : P \to \mathbb{R}_{\geq 0}, \tag{1}$$

the possessed-capability volume measure and the window-active realized capability volume measure, respectively. The *Goodhart slack* for an alignment property $g : P^{\mathrm{act}} \to \mathbb{R}$ with Lipschitz constant $\mathrm{Lip}(g)$ is $|g(T) - g(P)|$, bounded above by $\mathrm{Lip}(g) \cdot \|P - T\|$ via [Lasser, 2026b, Theorem 2]'s Lipschitz transfer.

2.2 Verification ledger and channel structure

The deployment uses [Lasser, 2026a, Exogenous Verification]'s verification infrastructure: an append-only verification ledger on substrate-exclusive nodes, recording per-step events from a finite-or-countable action space $A = A^{\mathrm{ev}} \sqcup \{a_0\}$, where A^{ev} is the set of recordable event types and a_0 is the *idle action* ("no observable event in this step").

The deployment monitors K_{ch} channels indexed by $j \in \{1, \ldots, K_{\mathrm{ch}}\}$. Each channel has:

- A **projection map** $\pi_j : A \to V_j \cup \{*_j\}$, where V_j is channel j's observable space (counts for Poisson cooperative-rate, $\{0, 1\}$ for Bernoulli attestation, finite category set for multinomial concentration channels) and $*_j$ is the trivial / null projection. Action a *engages* channel j iff $\pi_j(a) \neq *_j$.

- An **extended baseline distribution** $\bar{p}_0^{(j)}$ on $V_j \cup \{*_j\}$ with baseline null-mass $\epsilon_j^{(0)} := \bar{p}_0^{(j)}(\{*_j\}) \in [0,1)$ and conditional baseline $p_0^{(j)}(v) := \bar{p}_0^{(j)}(v)/(1-\epsilon_j^{(0)})$ for $v \in V_j$.
- A least-favorable adversarial alternative $p_1^{(j)}$ at threshold shift η_j.
- A clipped per-step log-likelihood-ratio function $\ell^{(j)} : A \to [-B_{\mathrm{clip}}, B_{\mathrm{clip}}]$ that depends on a only through $\pi_j(a)$, with $\ell^{(j)}(a) = 0$ whenever $\pi_j(a) = *_j$ (zero-outside-engagement convention). B_{clip} is the (C5.HOEFF) clip radius, fixed before deployment and independent of $|P|$.

The set of actions engaging channel j is $A_j := \{a \in A : \pi_j(a) \neq *_j\}$. For ordered channel pairs (j, j') with $j \neq j'$, the *pairwise overlap set* is $A_{j,j'} := A_j \cap A_{j'}$.

2.3 SPRT machinery and exposure events

The deployment runs continuous SPRT detection [Lasser, 2026a, Wald, 1947] on each monitored channel, with familywise allocation across K_{ch} channels under (C5.MULT). Per-step LLR increments are clipped to $[-B_{\mathrm{clip}}, B_{\mathrm{clip}}]$ under (C5.HOEFF); the post-clipping union-class drift floor is $\delta_* > 0$ under (C5.IID).

Let $N_{\mathrm{ev}}(\tau_{\mathrm{meta}})$ be the number of SPRT exposure events that occur in the cascade-time window $[0, \tau_{\mathrm{meta}}]$, where τ_{meta} is the metastable-lifetime lower bound from [Lasser, 2026f, Phase Redundancy]. Under (C11.CLK) [Lasser, 2026c, Deployment Safety], $N_{\mathrm{ev}}(\tau_{\mathrm{meta}}) \geq N_{\mathrm{cascade}}$ with probability $\geq 1 - b_{\mathrm{clk}}$ (lower bound for detection); the upper bound $N_{\mathrm{ev}}(\tau_{\mathrm{meta}}) \leq \lambda_{\max} \cdot \tau_{\mathrm{meta}}$ is supplied by (C7.RATE), introduced in §3 below.

2.4 Trade-flow weighting and Herfindahl index

Following [Lasser, 2026e, Need Sufficiency], the deployment monitors trade flows attributed to counterparties (S1-admissible participants [Lasser, 2026d]). The trade-flow weighting is a probability measure w over counterparties (post-coalition closure under (COAL); see §4), with the Herfindahl-Hirschman index [Herfindahl, 1950, Hirschman, 1964]

$$\mathrm{HHI}(w) := \sum_c w_c^2 = \|w\|_2^2. \tag{2}$$

[Lasser, 2026c, Deployment Safety]'s invariant I_5 requires $\mathrm{HHI}(w) < H^*$ for a deployment-class threshold $H^* < 1$. The Part-A category-flow variant of [Lasser, 2026e] corresponds to this counterparty-attributed form under the natural attribution: each counterparty's trade contribution determines its weight, and (REP)/(DISP)/(COAL) of §4 are formulated on this counterparty-weight representation.

2.5 Capability volume and Lipschitz transfer

The Lipschitz transfer of [Lasser, 2026b, Theorem 2] is the bridge from proxy-truth divergence to Goodhart slack: for any alignment property g on P^{act} with Lipschitz constant $\mathrm{Lip}(g)$,

$$|g(T) - g(P)| \;\leq\; \mathrm{Lip}(g) \cdot \|P - T\|. \tag{3}$$

This is forward-only: no g-level lower bound follows from a proxy-truth-norm lower bound without additional non-degeneracy structure on g.

2.6 New structural quantities

Three structural quantities beyond [Lasser, 2026c, Deployment Safety]'s vocabulary:

- **Upper exposure-rate cap** $\lambda_{\max}$ (C7.RATE), introduced in §3.
- **Base population measure** μ on counterparties, introduced in §4.
- χ^2 **divergence** $\chi^2(w\|\mu)$ as the natural concentration measure when the base measure is non-uniform; equals $N\mathrm{HHI}(w) - 1$ when μ is uniform on a finite set of N counterparties.

Audit procedures for each are in §6.

3 Bounded co-evolution as a corollary

The bounded co-evolution assumption of [Lasser, 2026c, Condition (C7)] (invoked as condition (C7) of the deployment-safety theorem) asserts that the per-channel coupling magnitude $\bar{M}$ is bounded uniformly by a constant depending on deployment-class parameters but independent of $|P|$. This section derives the assertion from (C5.HOEFF), a new upper exposure-rate cap (C7.RATE), and the verification protocol's channel-projection structure under (C5.MULT). Audit 4 of [Lasser, 2026c, §8] previously specified an empirical procedure for measuring $\bar{M}$ post-hoc; under the derivation below, the audit reduces to certifying the three structural inputs and computing $\bar{M}$ from primitives.

3.1 Per-channel coupling magnitude (definition)

3.1.1 j-supported perturbations

A *j-supported signed measure* is a function $\Delta : A \to \mathbb{R}$ with $\Delta(a) = 0$ for $a \notin A_j$ and $\sum_{a \in A_j} \Delta(a) = 0$. The set of admissible j-perturbations $\mathcal{D}_j$ is the convex subset of j-supported signed measures satisfying $q_0 + \Delta \geq 0$ pointwise (so that $q_0 + \Delta$ is a valid probability measure on A).

The ambient *linear* space of j-supported zero-sum signed measures is

$$H_j \;:=\; \{\Delta : A \to \mathbb{R} \mid \Delta(a) = 0 \text{ for } a \notin A_j,\;\; \textstyle\sum_a \Delta(a) = 0\}. \tag{4}$$

The functional $\Delta \mapsto \Delta\mu_{j'}[\Delta] := \sum_{a \in A_{j,j'}} \Delta(a)\ell^{(j')}(a)$ extends linearly from $\mathcal{D}_j$ to all of H_j.

3.1.2 Per-step coupling magnitude

Definition 1 (Per-step coupling magnitude)**.** *For $j \neq j'$, the* per-step pairwise coupling magnitude *of channel j on channel j' is the operator norm of the linear functional $\Delta \mapsto \Delta\mu_{j'}[\Delta]$ on the linear space H_j under the ℓ_1 norm:*

$$M^{\text{step}}_{j \to j'}(\mathcal{D}) \;:=\; \sup_{\Delta \in H_j,\, \Delta \neq 0} \frac{\big|\Delta\mu_{j'}[\Delta]\big|}{\|\Delta\|_1}. \tag{5}$$

The per-step deployment coupling magnitude *is* $\bar{M}^{\text{step}}(\mathcal{D}) := \max_{j \neq j'} M^{\text{step}}_{j \to j'}(\mathcal{D})$.

By inclusion $\mathcal{D}_j \subseteq H_j$, the operator norm on H_j upper-bounds the admissible-restricted supremum $\sup_{\Delta \in \mathcal{D}_j \setminus \{0\}} |\Delta\mu_{j'}[\Delta]|/\|\Delta\|_1$. We work with the H_j-bound throughout; equality of the two suprema requires q_0 to have strictly positive mass on every $a \in A_j$ (so $\mathcal{D}_j$ contains a neighborhood of 0 in H_j), which we do not assume.

3.2 Per-step closed-form bound

Lemma 1 (Per-step coupling bound)**.** *Under (C5.HOEFF) (per-step LLR clipped to $[-B_{\text{clip}}, B_{\text{clip}}]$) and the zero-outside-engagement convention, for every $j \neq j'$:*

$$M^{\text{step}}_{j \to j'}(\mathcal{D}) \;\leq\; B_{\text{clip}}, \tag{6}$$

unconditionally. When channels are action-disjoint ($A_{j,j'} = \emptyset$), $M^{\text{step}}_{j \to j'} = 0$.

Proof. For any $\Delta \in H_j$ with $\Delta \neq 0$:

$$\begin{aligned}\big|\Delta\mu_{j'}[\Delta]\big| = \Big|\sum_{a \in A_{j,j'}} \Delta(a)\ell^{(j')}(a)\Big| &\leq \sum_{a \in A_{j,j'}} |\Delta(a)| \cdot |\ell^{(j')}(a)| \\ &\leq B_{\text{clip}} \sum_{a \in A_{j,j'}} |\Delta(a)| \leq B_{\text{clip}} \cdot \|\Delta\|_1.\end{aligned}$$

Dividing by $\|\Delta\|_1 \neq 0$ and taking supremum over $H_j \setminus \{0\}$ gives (6). When $A_{j,j'} = \emptyset$, the sum is empty and $\Delta\mu_{j'}[\Delta] = 0$ for all $\Delta \in H_j$, giving $M^{\text{step}}_{j \to j'} = 0$. ∎

3.3 Cumulative cascade-window bound

3.3.1 Pathwise perturbation class

A *pathwise j-perturbation over the cascade window* is a sequence $\{\Delta_n\}_{n=1}^{N_{\mathrm{ev}}}$ where each $\Delta_n \in H_j$, applied at SPRT exposure step n. Two budget conventions are operationally relevant:

- **Per-step budget:** $\sup_n \|\Delta_n\|_1 \le 1$.
- **Total budget:** $\sum_n \|\Delta_n\|_1 \le 1$.

The cumulative cross-channel response is

$$\mu_{j'}^{\mathrm{cum}}\big[\{\Delta_n\}\big] \;:=\; \sum_{n=1}^{N_{\mathrm{ev}}(\tau_{\mathrm{meta}})} \Delta\mu_{j'}[\Delta_n].$$

3.3.2 The new condition: upper exposure-rate cap (C7.RATE)

Assumption 1 (Upper exposure-rate cap, (C7.RATE))**.** *There exists a deployment-class action rate cap $\lambda_{\max} > 0$ independent of $|P|$, such that the SPRT exposure event count satisfies*

$$N_{\mathrm{ev}}(\tau_{\mathrm{meta}}) \;\le\; \lambda_{\max} \cdot \tau_{\mathrm{meta}} \tag{7}$$

deterministically (operator-enforced via rate limits or action bounds). This is distinct from (C11.CLK) of [Lasser, 2026c, Deployment Safety] which gives a lower bound on $N_{\mathrm{ev}}(\tau_{\mathrm{meta}})$ for detection purposes; (C7.RATE) is an upper bound for coupling-control purposes.

Operationally, $\lambda_{\max}$ is calibrated by the deployment's rate limits, action bounds, and channel-coupling structure (specified in §6).

3.3.3 Cumulative bound

Lemma 2 (Cumulative coupling bound)**.** *Under (C5.HOEFF), the zero-outside-engagement convention, and (C7.RATE):*

(a) Pairwise *cumulative coupling, per-step budget* $\sup_n \|\Delta_n\|_1 \le 1$*:*

$$\sup_{\substack{\Delta_n \in H_j \ \forall n \\ \sup_n \|\Delta_n\|_1 \le 1}} \big|\mu_{j'}^{\mathrm{cum}}\big[\{\Delta_n\}\big]\big| \;\le\; B_{\mathrm{clip}} \cdot \lambda_{\max} \cdot \tau_{\mathrm{meta}}. \tag{8}$$

(b1) Aggregate-incoming *cumulative coupling under* ***total per-step budget*** $\sup_n \sum_j \|\Delta_n^{(j)}\|_1 \le 1$*:*

$$\sup_{\substack{\Delta_n^{(j)} \in H_j \ \forall n,j \\ \sup_n \sum_j \|\Delta_n^{(j)}\|_1 \le 1}} \Big|\sum_{j \ne j'} \mu_{j'}^{\mathrm{cum}}\big[\{\Delta_n^{(j)}\}\big]\Big| \;\le\; B_{\mathrm{clip}} \cdot \lambda_{\max} \cdot \tau_{\mathrm{meta}}. \tag{9}$$

(b2) Aggregate-incoming *cumulative coupling under* ***per-source per-step budget*** $\sup_n \|\Delta_n^{(j)}\|_1 \le 1$ *for every j:*

$$\sup_{\substack{\Delta_n^{(j)} \in H_j \ \forall n,j \\ \sup_n \|\Delta_n^{(j)}\|_1 \le 1 \ \forall j}} \Big|\sum_{j \ne j'} \mu_{j'}^{\mathrm{cum}}\big[\{\Delta_n^{(j)}\}\big]\Big| \;\le\; (K_{\mathrm{ch}} - 1) \cdot B_{\mathrm{clip}} \cdot \lambda_{\max} \cdot \tau_{\mathrm{meta}}. \tag{10}$$

The right-hand sides are deployment-class constants intensive in $|P|$ over $\mathcal{D}$ (B_{clip} from (C5.HOEFF), $\lambda_{\max}$ from (C7.RATE), τ_{meta} from [Lasser, 2026f, Phase Redundancy]'s scaling, K_{ch} from (C5.MULT)).

Proof. Part (a). By Lemma 1, $|\Delta\mu_{j'}[\Delta_n]| \le B_{\mathrm{clip}} \cdot \|\Delta_n\|_1 \le B_{\mathrm{clip}}$ for each n when $\sup_n \|\Delta_n\|_1 \le 1$. Summing:

$$\big|\mu_{j'}^{\mathrm{cum}}\big| \le \sum_{n=1}^{N_{\mathrm{ev}}(\tau_{\mathrm{meta}})} |\Delta\mu_{j'}[\Delta_n]| \le N_{\mathrm{ev}}(\tau_{\mathrm{meta}}) \cdot B_{\mathrm{clip}} \le \lambda_{\max} \cdot \tau_{\mathrm{meta}} \cdot B_{\mathrm{clip}},$$

the last inequality by (C7.RATE).

Part (b1) (total per-step budget). At each step n, $\sum_j \|\Delta_n^{(j)}\|_1 \le 1$. Each per-source contribution is bounded by $B_{\mathrm{clip}} \cdot \|\Delta_n^{(j)}\|_1$ via Lemma 1. Summing over $j \neq j'$ at each step:

$$\big|\sum_{j\neq j'} \Delta\mu_{j'}[\Delta_n^{(j)}]\big| \le B_{\mathrm{clip}} \cdot \sum_{j\neq j'} \|\Delta_n^{(j)}\|_1 \le B_{\mathrm{clip}}.$$

Summing over n and applying (C7.RATE) gives (9).

Part (b2) (per-source per-step budget). Now each source j has $\|\Delta_n^{(j)}\|_1 \le 1$ independently. Per-source contributions remain bounded by B_{clip} each; summing over $K_{\mathrm{ch}} - 1$ sources:

$$\big|\sum_{j\neq j'} \Delta\mu_{j'}[\Delta_n^{(j)}]\big| \le (K_{\mathrm{ch}} - 1) \cdot B_{\mathrm{clip}}.$$

Summing over n and applying (C7.RATE) gives (10). Which budget convention is operationally relevant depends on the audit specification of admissible adversarial classes; we state both for completeness. ■

3.4 Bounded co-evolution as a corollary of primitive parameters

Theorem 1 (Bounded co-evolution corollary)**.** *Under (C5.HOEFF) per-step LLR clipping, (C5.MULT) channel multiplicity bound, (C7.RATE) upper exposure-rate cap, and the τ_{meta} scaling of [Lasser, 2026f, Phase Redundancy], the deployment's cumulative coupling magnitude $\bar{M}^{\mathrm{cum}}(\mathcal{D})$ over the cascade window τ_{meta} satisfies*

$$\bar{M}^{\mathrm{cum}}(\mathcal{D}) \;\le\; C \cdot B_{\mathrm{clip}} \cdot \lambda_{\max} \cdot \tau_{\mathrm{meta}}, \tag{11}$$

where $C \in \{1, K_{\mathrm{ch}} - 1\}$ depending on the budget convention (pairwise, total, or per-source). All four factors on the RHS are deployment-class constants intensive in $|P|$.

In particular: [Lasser, 2026c, Condition (C7)]'s bounded-co-evolution claim that $\bar{M}$ is bounded by a $|P|$-independent constant follows directly from (11)*, with the constant given explicitly in terms of primitive deployment parameters.*

Proof. Direct application of Lemma 2. Intensivity of each RHS factor: B_{clip} from (C5.HOEFF) is a deployment-class constant; $\lambda_{\max}$ from (C7.RATE) is operator- enforced and deployment-class; τ_{meta} from [Lasser, 2026f] depends on deployment-class redundancy and rate parameters but not on $|P|$; K_{ch} from (C5.MULT) is fixed before deployment. ■

3.5 Optional sharper bound under coalition-overlap stability

The basic Theorem 1 bound uses only $\|\Delta\|_1 \le 1$ to get $|\Delta\mu_{j'}| \le B_{\mathrm{clip}}$. A sharper bound is available when the deployment imposes an additional structural condition on the channel partition's shared-action mass:

Definition 2 (Optional overlap-mass stability, (C5.OVL))**.** *The channel partition's shared-action mass is bounded by a deployment-class constant $Q_{\max} < 1$, in the sense that for every $j \neq j'$,*

$$\sup_{\Delta\in H_j,\, \Delta\neq 0} \frac{\sum_{a\in A_{j,j'}} |\Delta(a)|}{\|\Delta\|_1} \;\le\; Q_{\max}.$$

When (C5.OVL) is adopted, the per-step bound in Lemma 1 sharpens to $M^{\mathrm{step}}_{j\to j'} \le Q_{\max} \cdot B_{\mathrm{clip}}$, and the cumulative bound in Theorem 1 sharpens to $\bar{M}^{\mathrm{cum}} \le C \cdot Q_{\max} \cdot B_{\mathrm{clip}} \cdot \lambda_{\max} \cdot \tau_{\mathrm{meta}}$. The intensivity claim does not require (C5.OVL); only strict-smallness ($\bar{M} \to 0$ as $Q_{\max} \to 0$, i.e., as channels become action-disjoint) requires it.

3.6 Lemma 1 intensivity step

[Lasser, 2026c, Lemma 1, Step 6] (intensivity step in the Goodhart-slack composition) requires $(\varepsilon^{\mathrm{nonres}}_{\mathrm{coev}})_+ \le K_{\mathrm{coev}} \cdot \bar{M}$ to be bounded by a $|P|$-independent constant, where K_{coev} is the structural

co-evolution constant from [Lasser, 2026b]'s composition proposition. Theorem 1 provides this directly:

$$(\varepsilon_{\text{coev}}^{\text{nonres}})_+ \;\le\; K_{\text{coev}} \cdot \bar{M}^{\text{cum}} \;\le\; K_{\text{coev}} \cdot C \cdot B_{\text{clip}} \cdot \lambda_{\max} \cdot \tau_{\text{meta}}, \tag{12}$$

all factors deployment-class constants. The Goodhart-slack intensivity property of [Lasser, 2026c, Deployment Safety]'s deployment claim follows.

3.7 Failure-mode decomposition

[Lasser, 2026c, §10.2]'s sensitivity analysis characterized bounded co-evolution failure as a hard-failure mode with no graceful-degradation pathway. Under Theorem 1, the failure mode factors through the failure of (C7.RATE), (C5.HOEFF), or the channel-projection structure — each of which has its own audit-detectable failure signal. The hard-failure characterization is therefore replaced by a structural diagnosis: *which* primitive condition has failed, and how the deployment can restore it.

The remainder of [Lasser, 2026c, §10.2]'s sensitivity analysis simplifies correspondingly: the bounded-co-evolution paragraph becomes a forward reference to this section, with the failure-mode taxonomy decomposed across (C5.HOEFF), (C7.RATE), and channel-projection structure.

4 Concentration-Gap selection theorem (scoped)

The Concentration-Gap conjecture of [Lasser, 2026b, Microfoundation] states informally: optimization pressure on the proxy correlates with proxy-truth gap exploitation. The conjecture is deferred in [Lasser, 2026b] as open work; the present section proves a scoped version under three operationally auditable structural conditions ((REP), (DISP), (COAL)) plus a Lipschitz embedding compatibility precondition (Assumption 2), at which point the proxy-truth Goodhart slack is bounded above (and below at the proxy-truth-norm level) by an explicit χ^2-divergence quantity controlling Herfindahl-index trade-flow concentration.

The proof proceeds in three layers: an algebraic weighted-selection kernel (§4.1), a counterparty-selection interpretation (§4.2), and a transfer to Goodhart slack via the Lipschitz machinery of [Lasser, 2026b, Theorem 2] (§4.3). The universal model-free Concentration-Gap conjecture remains open; §7 states the limits of the scoped discharge.

4.1 The algebraic weighted-selection kernel

4.1.1 Setup

Let $\mathcal{V}$ be a real Hilbert space with norm $\|\cdot\|$ (in applications: capability-poset measures, or its tangent space at the welfare-relevant truth). Let $\mathcal{C}$ be a measurable space (in applications: the set of counterparties, possibly after coalition partitioning per (COAL)). Let μ be a fixed *base population measure* on $\mathcal{C}$ — the auditor-defined fair-or-natural distribution over admissible counterparties.

Let $\Delta : \mathcal{C} \to \mathcal{V}$ be a deterministic *deviation field*, $c \mapsto \Delta_c$, μ-integrable. A *weighting* w is a probability measure on $\mathcal{C}$ absolutely continuous with respect to μ, with Radon-Nikodym derivative $r := dw/d\mu$ satisfying $\mathbb{E}_\mu[r] = 1$.

The *population mean* is $\bar{\Delta} := \int_{\mathcal{C}} \Delta_c \, d\mu(c) = \mathbb{E}_\mu[\Delta]$. The *weighted aggregate distortion* is

$$\Delta(w) \;:=\; \int_{\mathcal{C}} \Delta_c \, dw(c) \;=\; \mathbb{E}_w[\Delta] \;=\; \mathbb{E}_\mu[r \cdot \Delta]. \tag{13}$$

4.1.2 Two scalar concentration measures: HHI and χ^2 divergence

Definition 3 (χ^2 divergence)**.** *The* χ^2 divergence *[Kullback and Leibler, 1951] of* w *from* μ *is*

$$\chi^2(w \,\|\, \mu) \;:=\; \mathbb{E}_\mu[(r-1)^2] \;=\; \mathrm{Var}_\mu(r) \;\ge\; 0. \tag{14}$$

$\chi^2(w\|\mu) = 0$ *iff* $w = \mu$.

Definition 4 (HHI and its relation to χ^2)**.** *For discrete* $\mathcal{C}$*, the Herfindahl-Hirschman index of* w *is* $\mathrm{HHI}(w) := \sum_c w_c^2 = \|w\|_2^2$*. When* μ *is the uniform measure on a finite* $\mathcal{C}$ *of size* N*, the two divergences are related by*

$$\chi^2(w \,\|\, \mu) \;=\; N \cdot \mathrm{HHI}(w) - 1. \tag{15}$$

For non-uniform μ, $\chi^2(w\|\mu)$ generalizes $\mathrm{HHI}(w)$ *as the appropriate concentration measure when the population is not uniformly weighted.*

4.1.3 Structural conditions

Definition 5 (Representativeness, (REP))**.** *The deviation field is* ρ_{rep}-representative *relative to base measure μ if $\|\bar{\Delta}\| \le \rho_{\mathrm{rep}}$.*

Definition 6 (Bounded dispersion, (DISP))**.** *The deviation field has* σ-bounded dispersion *relative to base measure μ if $\int_{\mathcal{C}} \|\Delta_c - \bar{\Delta}\|^2 \, d\mu(c) \le \sigma^2$.*

Definition 7 (Coalition closure, (COAL))**.** *The counterparty space $\mathcal{C}$ is* coalition-closed *under audit if (i) all counterparties whose actions or trade flows are correlated above a deployment-class threshold are partitioned into a single equivalence class (a coalition), and Δ is defined on coalitions rather than individual counterparties; (ii) the post-partition counterparty cardinality N is a deployment-class constant (counterparty onboarding is governed by a deployment-class policy that does not let N scale with $|P|$); and (iii) latent (undetected) coalition risk is bounded by an explicit residual term η_{latent} that the deployment claim absorbs (see §4.3 below). The audit specifies the partition; the algebraic theorem operates on the post-partition $\mathcal{C}$.*

Clause (ii) is what makes the HHI-to-χ^2 translation intensive in $|P|$: $\chi^2(w \,\|\, \mu) = N\mathrm{HHI}(w) - 1$ for uniform μ, so an HHI threshold yields a $|P|$-independent χ^2 ceiling only when N is itself $|P|$-independent. Deployments that prefer a structural commitment directly on $\chi^2(w \,\|\, \mu)$ rather than on HHI *may substitute (ii) with the alternative sufficient condition $\chi^2(w \,\|\, \mu) \le \Xi$ for a deployment-class constant Ξ (the two are not logically equivalent — the χ^2 bound is the weaker hypothesis the proof actually uses — but either certifies the intensivity that the theorem requires).*

4.1.4 Forward direction: algebraic kernel

Lemma 3 (Algebraic weighted-selection, forward direction)**.** *Let Δ be a deviation field on $(\mathcal{C}, \mu)$ in a Hilbert space $\mathcal{V}$, satisfying (REP) and (DISP). Let w be a weighting on $\mathcal{C}$ with $\chi^2(w\|\mu) < \infty$. Then*

$$\|\Delta(w)\| \;\le\; \rho_{\mathrm{rep}} \;+\; \sigma \cdot \sqrt{\chi^2(w \,\|\, \mu)}. \tag{16}$$

Proof. Decompose $\Delta(w)$ around the population mean:

$$\Delta(w) - \bar{\Delta} \;=\; \mathbb{E}_\mu[r \cdot \Delta] - \mathbb{E}_\mu[\Delta] \;=\; \mathbb{E}_\mu[(r-1)\cdot\Delta] \;=\; \mathbb{E}_\mu[(r-1)(\Delta - \bar{\Delta})],$$

where the last equality uses $\mathbb{E}_\mu[r - 1] = 0$.

We bound the Hilbert-valued mean via duality. For any unit vector $u \in \mathcal{V}$ with $\|u\| = 1$:

$$\big\langle u,\, \mathbb{E}_\mu[(r-1)(\Delta - \bar{\Delta})] \big\rangle \;=\; \mathbb{E}_\mu\big[(r-1)\cdot\langle u, \Delta - \bar{\Delta}\rangle\big].$$

By Cauchy-Schwarz on $L^2(\mu)$ applied to the scalar product of $(r-1)$ and $\langle u, \Delta - \bar{\Delta}\rangle$:

$$\big|\mathbb{E}_\mu[(r-1)\langle u, \Delta - \bar{\Delta}\rangle]\big| \;\le\; \sqrt{\mathbb{E}_\mu[(r-1)^2]} \cdot \sqrt{\mathbb{E}_\mu[\langle u, \Delta - \bar{\Delta}\rangle^2]}.$$

Since $|\langle u, \Delta - \bar{\Delta}\rangle| \le \|\Delta - \bar{\Delta}\|$ for $\|u\| = 1$, the second factor is bounded by $\sqrt{\mathbb{E}_\mu[\|\Delta - \bar{\Delta}\|^2]} \le \sigma$. Taking the supremum over unit u recovers the Hilbert norm:

$$\|\mathbb{E}_\mu[(r-1)(\Delta - \bar{\Delta})]\| \;=\; \sup_{\|u\|=1} \big\langle u, \mathbb{E}_\mu[(r-1)(\Delta - \bar{\Delta})]\big\rangle \;\le\; \sqrt{\chi^2(w\|\mu)} \cdot \sigma.$$

Therefore $\|\Delta(w) - \bar{\Delta}\| \le \sigma\sqrt{\chi^2(w\|\mu)}$. Adding the representativeness bound:

$$\|\Delta(w)\| \;\le\; \|\bar{\Delta}\| + \|\Delta(w) - \bar{\Delta}\| \;\le\; \rho_{\mathrm{rep}} + \sigma\sqrt{\chi^2(w\|\mu)}.$$

■

Remark 1 (On cross-counterparty correlations)**.** *The Cauchy-Schwarz argument above does not require any probabilistic decorrelation between counterparty deviations. The only structural inputs are (REP) and (DISP), both of which are properties of the deterministic deviation field Δ (not of its realization randomness). Cross-counterparty correlations are absorbed automatically into the σ^2 population-variance term. The case of identical deviations across all counterparties (highly correlated, no diversification) corresponds to $\sigma^2 = 0$, in which case the bound is trivially ρ_{rep}.*

4.1.5 Reverse direction (norm-level)

Lemma 4 (Algebraic weighted-selection, reverse direction)**.** Scope. *Throughout this lemma, $\mathcal{C}$ is a finite or countable atomic measurable space (the natural setting after coalition closure (COAL) partitions counterparties into discrete equivalence classes); $w = (w_c)_{c\in\mathcal{C}}$ is the corresponding atomic weighting.*

Let Δ be a deviation field on $(\mathcal{C}, \mu)$. Suppose:

- ***Concentration.*** $\mathrm{HHI}(w) \geq H^*$*, hence* $\max_c w_c \geq H^*$ *(since* $\sum_c w_c^2 \leq \max_c w_c$*). Let* $d \in \arg\max_c w_c$ *be the dominant counterparty, with* $w_d \geq H^*$*.*
- ***Dominant separation.*** $\|\Delta_d\| \geq \delta$*.*
- ***Directional cancellation.*** *Let* $u_d := \Delta_d/\|\Delta_d\|$ *(unit direction). The remainder cancellation in this direction is bounded:* $\langle u_d, \sum_{c\neq d} w_c\Delta_c\rangle \geq -\beta$*.*

Then

$$\|\Delta(w)\| \;\geq\; H^*\delta - \beta. \tag{17}$$

Proof. Project $\Delta(w)$ onto u_d:

$$\begin{aligned}\langle u_d, \Delta(w)\rangle &= w_d\langle u_d, \Delta_d\rangle + \langle u_d, \sum_{c\neq d} w_c\Delta_c\rangle \\ &= w_d\|\Delta_d\| + \langle u_d, \sum_{c\neq d} w_c\Delta_c\rangle \\ &\geq H^*\delta - \beta\end{aligned}$$

by concentration ($w_d \geq H^*$), separation ($\|\Delta_d\| \geq \delta$), and directional cancellation. Then $\|\Delta(w)\| \geq |\langle u_d, \Delta(w)\rangle| \geq H^*\delta - \beta$. ∎

4.2 Counterparty-selection interpretation

In the GFM trade-flow setting:

- $\mathcal{C}$: the set of counterparties (S1-admissible participants under [Lasser, 2026d]), *after coalition closure* per (COAL).
- μ: the auditor-defined fair-or-natural distribution over admissible counterparties.
- $\mathcal{V}$: Hilbert space of utility functionals on the capability poset (or its tangent space at the welfare-relevant truth W).
- Embedding $\phi : \mathcal{V}_{\mathrm{utility}} \to \mathcal{V}$: an affine isometric embedding on the active subspace P^{act}. The deployment specifies ϕ as part of the audit setup.
- $\Delta_c := \phi(U_c) - \phi(W)$: counterparty c's utility deviation from the welfare-relevant truth in the embedded representation.
- w_c: counterparty c's trade-flow weight (paper 10's invariant I_5 measures $\mathrm{HHI}(w)$ on coalition-closed weights).

The aggregate distortion $\Delta(w) = \sum_c w_c\phi(U_c) - \phi(W)$ is the *selection-induced proxy-truth deviation*: how the weighted-trade-flow proxy diverges from the welfare- relevant truth in the embedded representation.

4.3 Route-C transfer: from selection distortion to Goodhart slack

Assumption 2 (Embedding compatibility)**.** *The embedding $\phi : \mathcal{V}_{\mathrm{utility}} \to \mathcal{V}$ is an* affine isometry *on the operationally active subspace P^{act}: $\|\phi(x) - \phi(y)\| = \|x - y\|_{\mathcal{V}_{\mathrm{utility}}}$ for $x, y \in P^{\mathrm{act}}$, and is linear up to a fixed translation.*

The proxy-truth difference admits the decomposition

$$\phi(P) - \phi(T) \;=\; \Delta_{\mathrm{audit}}(w) \;+\; e_{\mathrm{latent}}, \tag{18}$$

where $\Delta_{\mathrm{audit}}(w)$ is the audit-observable selection distortion (the $\Delta(w)$ of Lemma 3 computed on coalition-closed weights) and $e_{\mathrm{latent}} \in \mathcal{V}$ is the latent-coalition residual with $\|e_{\mathrm{latent}}\| \le \eta_{\mathrm{latent}}$.

If ϕ is only bi-Lipschitz (not isometric), the same argument applies with bounds multiplied by the bi-Lipschitz constants; isometric is the cleanest form.

Theorem 2 (Concentration-Gap Selection Theorem, scoped)**.** *Under (REP), (DISP), (COAL), and Assumption 2:*

$$|g(T) - g(P)| \;\le\; \mathrm{Lip}(g) \cdot \big(\rho_{\mathrm{rep}} + \sigma\sqrt{\chi^2(w\|\mu)} + \eta_{\mathrm{latent}}\big). \tag{19}$$

In particular, when μ is uniform on N counterparties with N a deployment-class constant (clause (ii) of (COAL)) and $\mathrm{HHI}(w) < H^*$ ([Lasser, 2026c, invariant I_5] in force): $\chi^2(w\|\mu) \le NH^* - 1$, giving Goodhart slack bounded by $\mathrm{Lip}(g) \cdot (\rho_{\mathrm{rep}} + \sigma\sqrt{NH^* - 1} + \eta_{\mathrm{latent}})$, intensive in $|P|$.

Deployments that prefer a χ^2-direct commitment may replace (COAL)'s clause (ii) with the alternative sufficient condition $\chi^2(w \,\|\, \mu) \le \Xi$ for deployment-class Ξ (the χ^2 bound is the weaker hypothesis the proof actually uses); the HHI *form is the operational surrogate that paper 10's I_5 already monitors.*

Reverse direction (norm-level only). *Under Lemma 4's concentration + separation + directional-cancellation conditions, and the latent-coalition residual bound from Assumption 2:*

$$\|P - T\| \;\ge\; H^*\delta - \beta - \eta_{\mathrm{latent}}, \tag{20}$$

non-trivial when $H^*\delta > \beta + \eta_{\mathrm{latent}}$. A g-level lower bound on $|g(T) - g(P)|$ does not *follow from* (3) *alone; it requires additional inverse-Lipschitz / non-degeneracy structure on g that this theorem does not assume.*

Proof. Forward direction: by Assumption 2, $\|\phi(P) - \phi(T)\| \le \|\Delta_{\mathrm{audit}}(w)\| + \|e_{\mathrm{latent}}\|$. By Lemma 3, $\|\Delta_{\mathrm{audit}}(w)\| \le \rho_{\mathrm{rep}} + \sigma\sqrt{\chi^2(w\|\mu)}$. Since ϕ is isometric, $\|P - T\|_{\mathcal{V}_{\mathrm{utility}}} = \|\phi(P) - \phi(T)\|$. Apply (3): $|g(T) - g(P)| \le \mathrm{Lip}(g) \cdot \|P - T\|$, giving (19).

Reverse direction (norm-level): by Lemma 4, $\|\Delta_{\mathrm{audit}}(w)\| \ge H^*\delta - \beta$. By isometry of ϕ, $\|\phi(P) - \phi(T) - e_{\mathrm{latent}}\| \ge H^*\delta - \beta$. By triangle inequality, $\|P - T\| \ge H^*\delta - \beta - \eta_{\mathrm{latent}}$ (when this is positive). The g-level lower bound does not follow without inverse-Lipschitz structure on g. ■

4.4 Scoped Concentration-Gap discharge

The Concentration-Gap conjecture, scoped form. Under (REP) + (DISP) + (COAL) and Lipschitz embedding compatibility (Assumption 2), Theorem 2 establishes the conjecture's bidirectional correlation content at the proxy-truth level: low HHI bounds the slack above; high HHI plus separation bounds it below at the norm level. The universal model-free conjecture remains open.

Replaces the HHI surrogate-adequacy assumption. The HHI surrogate-adequacy claim ($\mathrm{HHI} < H^* \Rightarrow$ deployment outside the optimization-pressure regime), previously asserted as a single empirical-adequacy assumption of the deployment-safety paper and now collected as [Lasser, 2026c, Assumption 1], is replaced by the conjunction (REP) + (DISP) + (COAL), embedding compatibility (Assumption 2), and the HHI threshold itself. The previous monolithic claim becomes concrete checkable structural conditions, each with operational audit hooks specified in §6.

Layer 1 binding via I_5. The deployment claim of [Lasser, 2026c, Deployment Safety] no longer contains the surrogate-adequacy assumption as an unproved empirical claim. Theorem 1 Layer 1 binding via invariant I_5 follows from Theorem 2's upper bound under (REP) + (DISP) + (COAL) and embedding compatibility, with the structural conditions auditable per §6.

Failure-mode decomposition under (REP), (DISP), (COAL), and embedding compatibility. The previous monolithic failure mode decomposes into failure of (REP), (DISP), (COAL), or embedding compatibility — each of which has specific audit-detectable signals. Concentration-Gap-conjecture failure factors through whichever structural condition the deployment cannot certify; [Lasser, 2026c, §10.2]'s sensitivity-analysis taxonomy traces the failure-mode consequences for the deployment claim.

5 Composition with the deployment-safety theorem

Theorem 1 (bounded co-evolution corollary) and Theorem 2 (scoped Concentration-Gap) each enter the deployment-safety theorem [Lasser, 2026c, Theorem 1] at specific points. This section locates those points and lists the conditions each theorem adds to or removes from the deployment-claim hypothesis set.

5.1 Conditions removed and added

Removed. The deployment-claim hypothesis set, after the present theorems, no longer contains:

- The free-floating bounded co-evolution assumption that $\bar{M}$ is bounded by some $|P|$-independent constant. Replaced by Theorem 1. In [Lasser, 2026c, Deployment Safety], the property is now stated as condition (C7) and invokes the corollary form directly.
- The HHI surrogate-adequacy assumption: the converse-direction sufficiency claim HHI $<$ $H^* \Rightarrow \mathcal{R}^c_{\text{press}}$. Replaced by Theorem 2 under the conjunction (REP) + (DISP) + (COAL) and embedding compatibility (Assumption 2), now collected as [Lasser, 2026c, Assumption 1] (Concentration-Gap structural conditions).

Added. The deployment-claim hypothesis set acquires five explicit hypotheses, each operationally auditable (with embedding compatibility certified as a precondition for the embedded representations used by (REP) and (DISP)):

- **(C7.RATE).** Upper exposure-rate cap from Assumption 1. Added as a sub-clause of [Lasser, 2026c, Condition (C7)].
- **(REP).** Representativeness from Definition 5.
- **(DISP).** Bounded dispersion from Definition 6.
- **(COAL).** Coalition closure from Definition 7, with explicit residual η_{latent} for undetected coordination and post-partition counterparty cardinality N deployment-class bounded (clause (ii); structurally required for the HHI-to-χ^2 translation to be $|P|$-independent).
- **Embedding compatibility** from Assumption 2: an affine isometry (or bi-Lipschitz, with constants tracked through the bound) $\phi : \mathcal{V}_{\text{utility}} \to \mathcal{V}$ on the operationally active subspace. This is the structural prerequisite for transferring the algebraic distortion bound to $\|P - T\|$ and thence to Goodhart slack.

(REP), (DISP), (COAL), and embedding compatibility are collected as the structural conditions for the Concentration-Gap selection bound; (C7.RATE) is the exposure-rate condition for the bounded co-evolution corollary.

5.2 Where each theorem enters Deployment Safety's proof

Theorem 1 enters at Lemma 1's co-evolution substitution. [Lasser, 2026c, Lemma 1, Step 4] substitutes the co-evolution correction term $(\varepsilon^{\text{nonres}}_{\text{coev}})_+ \leq K_{\text{coev}} \cdot \bar{M}$ into the four-channel decomposition; [Lasser, 2026c, Lemma 1, Step 6] packages the result via the generic-X intensivity argument. Both rely on $\bar{M}$ being a $|P|$-independent constant. The corollary supplies $\bar{M}^{\text{cum}} \leq C \cdot B_{\text{clip}} \cdot \lambda_{\max} \cdot \tau_{\text{meta}}$ with all factors deployment-class constants intensive in $|P|$, closing the substitution at Step 4 without a free-floating constant.

Theorem 2 enters at Layer 1 binding via I_5. [Lasser, 2026c, Theorem 1, Layer 1 proof, Step 4] requires the converse-direction sufficiency HHI $< H^* \Rightarrow$ deployment outside $\mathcal{R}_{\text{press}}$. The scoped selection theorem supplies the Goodhart-slack upper bound $|g(T) - g(P)| \leq \text{Lip}(g) \cdot (\rho_{\text{rep}} + \sigma\sqrt{\chi^2(w\|\mu)} + \eta_{\text{latent}})$ under (REP) + (DISP) + (COAL) and embedding compatibility (Assumption 2); for uniform μ on N counterparties, $\chi^2 = N\text{HHI} - 1$, so HHI $< H^*$ yields a deployment-class constant ceiling on the slack.

Both theorems enter Deployment Safety's sensitivity analysis. [Lasser, 2026c, §10.2]'s failure-mode taxonomy decomposes into specific structural-condition failures: bounded-co-evolution failure factors through (C5.HOEFF) / (C7.RATE) / channel-projection-structure failures (each with audit-detectable signal), and Concentration-Gap structural-condition failure factors through (REP) / (DISP) / (COAL) / embedding-compatibility failures (the last via the bi-Lipschitz ϕ certification of §6).

5.3 Deployment-claim shape

The structural shape of the deployment-safety theorem is preserved:

- Eleven operational invariants $I_1, \ldots, I_{11}$.
- Three-layer claim (static safe region / detection-and-correction / acknowledged residuals).
- Five named residuals (R1)–(R5).
- Cooperative-anchoring property and canonical tripartite substrate identification.
- Operational conditions (C2)–(C6) and (C8)–(C12) of [Lasser, 2026c, §5] retain their content; (C1) and (C7) carry refined structural dependencies: (C1) now explicitly invokes (REP), (DISP), (COAL), and embedding compatibility via [Lasser, 2026c, Assumption 1], and (C7) carries the (C7.RATE) sub-clause established here.

Theorem 1 and Theorem 2 close structural gaps in the deployment-claim hypothesis set without altering the deployment claim's shape.

6 Audit infrastructure for the structural conditions

Four operational structural conditions are added to [Lasser, 2026c, Deployment Safety]'s deployment claim: (C7.RATE) (upper exposure-rate cap), (REP) (representativeness), (DISP) (bounded dispersion), and (COAL) (coalition closure); plus a Lipschitz embedding compatibility precondition (Assumption 2) that certifies ϕ before (REP) or (DISP) can be evaluated in the embedded representation. Each admits an operational audit, paralleling the existing audit structure of [Lasser, 2026c, §8]: each audit produces an attestation on the verification ledger [Lasser, 2026a], with cadence calibrated to deployment epoch boundaries. Confidence-bound construction in the (REP), (DISP), and (COAL) procedures below follows standard one-sided upper-confidence-bound machinery [Lehmann and Romano, 2005].

6.1 (C7.RATE) and channel-projection audit

This audit certifies the two structural inputs to Theorem 1 other than (C5.HOEFF) and (C5.MULT) (which are part of [Lasser, 2026c, Audit 7]): the upper exposure-rate cap $\lambda_{\max}$, and the verification protocol's channel-projection structure.

Procedure (rate cap).

1. Identify the per-step exposure event sources: ledger commit rate, action-execution rate, governance-event rate.
2. Inspect the hard mechanisms enforcing the rate ceiling — ledger commit caps, action-queue depth bounds, governance rate-limit middleware, scheduler quotas — and identify which substrate owns the enforcement (substrate-exclusivity from [Lasser, 2026a]).
3. Compute an upper-bound rate $\lambda_{\max}$ from the certified mechanisms (not from observed historical traffic).
4. Document $\lambda_{\max}$, the enforcing mechanisms, and the substrate ownership in the audit attestation.

Empirical traffic windows over τ_{meta} may be used as *monitoring* evidence (e.g., to alert on approach to $\lambda_{\max}$), but cannot establish the deterministic deployment-class upper bound by themselves.

Procedure (channel-projection structure).

1. Enumerate the projection maps $\pi_j : A \to V_j \cup \{*_j\}$ for $j = 1, \ldots, K_{\mathrm{ch}}$ from the verification protocol's event-classification policy.
2. Verify that $\pi_j(a) = *_j$ holds for all actions $a \notin A_j$, and that $\ell^{(j)}(a) = 0$ on $A \setminus A_j$ (zero-outside-engagement convention).
3. Document the pairwise overlap sets $A_{j,j'} = A_j \cap A_{j'}$ for use by downstream audits.

Cadence. Initial certification before deployment activation; re-certification when verification-protocol classification policy changes, when rate-limit mechanisms change, or when SPRT detection observes events at rates approaching $\lambda_{\max}$.

Failure mode. Deployments where the exposure event rate cannot be bounded by certified mechanisms (e.g., agent-controlled rate-limit overrides), or where the channel projection policy admits $\ell^{(j)}(a) \neq 0$ off A_j, fall outside Theorem 1's scope. Mitigation: hard rate limits enforced by substrate-exclusive infrastructure; verification-protocol classification redesign to respect the zero-outside-engagement convention.

6.2 Embedding ϕ certification (precondition for REP and DISP)

(REP) and (DISP) are stated relative to the embedded deviation field $\Delta_c = \phi(U_c) - \phi(W)$ in the Hilbert space $\mathcal{V}$ (Assumption 2). The embedding ϕ must be certified before (REP) or (DISP) can be audited.

Procedure.

1. Specify the embedding $\phi : \mathcal{V}_{\mathrm{utility}} \to \mathcal{V}$ in deployment documentation.
2. Verify ϕ is affine isometric on the operationally active subspace P^{act}: $\|\phi(x) - \phi(y)\| = \|x - y\|_{\mathcal{V}_{\mathrm{utility}}}$ for $x, y \in P^{\mathrm{act}}$, with linearity up to a fixed translation.
3. If ϕ is only bi-Lipschitz (not isometric), document the bi-Lipschitz constants $(\underline{L}, \bar{L})$ and apply them as multipliers to the (REP) / (DISP) bounds downstream.

Failure mode. If ϕ is not bi-Lipschitz on P^{act}, the algebraic kernel of Theorem 2 cannot be transferred to the deployment claim's $\|P - T\|$ form via Lipschitz transfer [Lasser, 2026b, Theorem 2]. Mitigation: choose a different embedding, or restrict the deployment's scope to a subspace where bi-Lipschitz ϕ exists.

When ϕ is well-conditioned. Utility representations that are linear (or near-linear) in the deployment's underlying decision variables typically admit well-conditioned ϕ: identity or affine maps suffice, and the bi-Lipschitz constants $(\underline{L}, \bar{L})$ are close to 1. High-dimensional learned representations (neural utility models, embedding vectors, opaque scalarizations of multi- attribute preferences) may not — distortion constants can be large, and the bi-Lipschitz form's downstream multiplication of $(\bar{L}/\underline{L})$ through the (REP) and (DISP) bounds can inflate the Goodhart-slack ceiling toward operational uselessness. Practical deployments should target one of: (i) an explicit affine utility model where ϕ is identity; (ii) a low-dimensional projection of a high-dimensional representation onto the operationally active subspace, with the projection's bi-Lipschitz constants bounded by audit; or (iii) a restricted deployment scope where the active subspace admits a well-conditioned embedding even if the full representation does not. Deployments that cannot achieve one of these lie outside Theorem 2's practical applicability even if they nominally satisfy (EMB).

6.3 (REP) audit (representativeness)

Procedure.

1. Define the base population measure μ: the auditor's fair-or-natural distribution over admissible counterparties (typically uniform over S1-admissible participants [Lasser, 2026d], or weighted by some natural allocation).

2. In the embedded representation under the certified ϕ, produce a one-sided upper confidence bound on $\|\bar{\Delta}\|$ where $\bar{\Delta} := \mathbb{E}_\mu[\phi(U_c) - \phi(W)]$. The bound is computed by sampling counterparties under μ (or reweighting observed-attestation samples to μ with stated coverage assumptions) from utility-disclosure attestations on the verification ledger; conservative bounding is required because (REP) is a hypothesis the deployment claim conditions on, not a quantity the deployment claim merely monitors.
3. Certify $\|\bar{\Delta}\| \leq \rho_{\mathrm{rep}}$ at the audit's chosen confidence level for a deployment-class threshold ρ_{rep}.
4. Document ρ_{rep}, the confidence level, and the sampling/reweighting design in the audit attestation.

Cadence. Initial calibration before deployment; re-calibration on counterparty-population changes (additions, removals, or substantial weight-redistribution).

Failure mode. The counterparty population's mean utility deviates systematically from the welfare-relevant truth W. Mitigation: counterparty-selection-process diversification, category-coverage audit [Lasser, 2026c, Audit 3], or restriction of the deployment's scope to a counterparty subset where (REP) holds.

6.4 (DISP) audit (bounded dispersion)

Procedure.

1. In the embedded representation under the certified ϕ, produce a one-sided upper confidence bound on the μ-population variance $\int \|\Delta_c - \bar{\Delta}\|^2 \, d\mu(c)$, where per-counterparty deviations $\Delta_c = \phi(U_c) - \phi(W)$ are read from utility-disclosure attestations. As with (REP), the variance is taken with respect to μ, not the trade-flow weighting w: audit sampling under μ (or reweighting to μ with stated coverage) is required.
2. Certify the variance $\leq \sigma^2$ at the audit's chosen confidence level for a deployment-class constant σ.
3. Document σ, the confidence level, and the sampling/reweighting design in the audit attestation.

Cadence. Continuous attestation under the calibrated μ-sampling design over a rolling calibration window; confidence-bound violations trigger audit re-calibration. Observed trade-flow (w) statistics may serve as monitoring signal but do not by themselves certify the μ-variance hypothesis.

Failure mode. High dispersion across counterparties (some counterparties' utilities deviate far from W, even when the population mean is centered). The bound in Theorem 2 weakens but does not collapse; the deployment claim accommodates large σ at the cost of larger Goodhart-slack ceiling.

6.5 (COAL) audit (coalition closure)

Procedure.

1. Run *coalition-detection analysis* on the verification ledger: identify counterparties with correlated trade flows above a deployment-class threshold (using repeated-beneficiary linkage, governance-vote alignment, attestation-source clustering).
2. Partition $\mathcal{C}$ into equivalence classes (coalitions) by the detected correlation structure.
3. Re-compute HHI on the post-partition (coalition-level) weights.
4. **Certify clause (ii) of (COAL).** Document the deployment-class bound $N_{\max}$ on the post-partition counterparty cardinality, together with the onboarding / governance policy that prevents N from scaling with $|P|$. Document the choice of base measure μ as the uniform distribution on the post-partition counterparties (the structural prerequisite for the HHI-to-χ^2 translation $\chi^2(w \,\|\, \mu) = N\mathrm{HHI}(w) - 1$). Deployments that prefer to commit directly to a χ^2 ceiling certify $\chi^2(w \,\|\, \mu) \leq \Xi$ for a deployment-class constant Ξ instead, with μ documented as whatever fair-or-natural base measure the audit selects.

5. Produce a one-sided upper confidence bound on the latent-coalition residual η_{latent} from the audit's detection-power calibration: bound $\eta_{\text{latent}} \leq p_{\text{miss}} \cdot \|\Delta\|_{\max}$ where p_{miss} is the threshold-miss probability for the worst-case undetected coalition size and $\|\Delta\|_{\max}$ is the worst-case per-counterparty deviation magnitude in the deployment's threat model.
6. Document the coalition partition, $N_{\max}$ (or Ξ), and the η_{latent} confidence bound (with confidence level) in the audit attestation.

Cadence. Continuous coalition detection on the verification ledger; audit-cadence alerts on detected coalition-formation events; re-partition triggered when new coalitions form.

Failure mode. Many small counterparties acting in coordinated fashion (a hidden coalition) drive the effective HHI above the audited threshold. The deployment claim's bound on Goodhart slack picks up the η_{latent} residual; deployments where η_{latent} cannot be bounded operationally are outside Theorem 2's scope.

When η_{latent} is operationally negligible. The Goodhart-slack bound has the additive form $\text{Lip}(g) \cdot (\rho_{\text{rep}} + \sigma\sqrt{\chi^2(w\|\mu)} + \eta_{\text{latent}})$, so η_{latent} dominates the bound whenever it is large compared with the other terms. For η_{latent} to remain operationally negligible the deployment should target $\eta_{\text{latent}} \leq \epsilon \cdot (\rho_{\text{rep}} + \sigma\sqrt{\chi^2(w\|\mu)})$ for some small $\epsilon \ll 1$, which translates via $\eta_{\text{latent}} \leq p_{\text{miss}} \cdot \|\Delta\|_{\max}$ into a detection-power requirement on the audit's coalition-detection machinery:

$$p_{\text{miss}} \cdot \|\Delta\|_{\max} \;\leq\; \epsilon \cdot \big(\rho_{\text{rep}} + \sigma\sqrt{\chi^2(w\|\mu)}\big).$$

Deployments whose threat model admits large $\|\Delta\|_{\max}$ (adversarial counterparties with large utility deviations from W) require correspondingly tighter p_{miss} to keep η_{latent} dominated. The audit's calibration documentation should make this trade-off explicit, and audits that cannot achieve the inequality leave deployments where η_{latent} may dominate the bound in the operational sense even when the algebraic bound formally holds.

6.6 Integration with existing audit infrastructure

Several of these new audits overlap with existing audits in [Lasser, 2026c, §8]:

- (C7.RATE) and channel-projection audit: integrates primarily with Audit 4 (bounded co-evolution calibration, dual-mode), which is the deployment-tooling home for the structural inputs to Theorem 1; Audit 7 (SPRT-applicability + gap-growth + clock comparability) supplies the (C5.HOEFF) clip radius and related SPRT-side calibration inputs but not the (C7.RATE) upper bound.
- (REP) audit: integrates with Audit 3 (cooperative-vs-redundancy audit), since both rely on category-coverage analysis.
- (DISP) audit: new, but uses utility-disclosure attestations that the existing infrastructure already produces.
- (COAL) audit: integrates with I_9 (substrate-exclusivity observability) and I_{10} (coverage/materiality routing) of [Lasser, 2026c, Deployment Safety], since coalition detection uses ledger-linkage analysis those invariants already specify.

The audit hooks above are the minimum needed for Theorem 1 and Theorem 2 to apply; consolidated audit-tooling combining these with [Lasser, 2026c, §8]'s existing audits is the natural target for further operational-tooling work.

7 Discussion

7.1 What the universal Concentration-Gap conjecture would require

The Concentration-Gap conjecture of [Lasser, 2026b, Microfoundation] is universal: optimization pressure on the proxy correlates with proxy-truth gap exploitation, in any deployment context. Theorem 2 discharges only a *scoped* version under (REP), (DISP), (COAL), embedding compatibility

(Assumption 2), and the Lipschitz-transfer machinery of [Lasser, 2026b]. The universal conjecture would additionally require:

- A formal optimizer / selection model that captures "optimization pressure" as a structural property of agents rather than as the empirical correlate (HHI) used here. Existing candidates include the mesa-optimization framework of [Hubinger et al., 2019] and the formal-Goodhart machinery of [El-Mhamdi and Hoang, 2024, Majka and El-Mhamdi, 2025], none of which alone yields a full proof.
- A characterization of "gap exploitation" robust to the specific functional forms of P and T. Theorem 2 works at the proxy-truth-norm level; the universal conjecture would need to characterize the alignment-property-level slack $|g(T) - g(P)|$, which requires inverse-Lipschitz structure on g not currently available in the GFM apparatus.

This is a research-program target rather than a technical extension. The scoped version is achievable; a universal model-free version is not, given current machinery.

7.2 What the g-level reverse direction would require

Theorem 2's reverse direction operates at the $\|P - T\|$ level: under high HHI plus separation plus directional cancellation, the proxy-truth norm is bounded below. The Lipschitz transfer of [Lasser, 2026b, Theorem 2] is forward-only, so this norm-level lower bound does not transfer to a lower bound on $|g(T) - g(P)|$.

Strengthening the reverse to the g-level would require an additional non-degeneracy condition on g: a constant g provides an immediate counterexample ($|g(T) - g(P)| = 0$ regardless of $\|P - T\|$). The natural condition is bi-Lipschitz invertibility of g on the relevant subspace, but this is a strong restriction on the alignment property's functional form.

For deployment-claim purposes, the norm-level reverse is sufficient: it tells operators the regime in which non-trivial proxy-truth distortion is structurally guaranteed, which the detection layer (Layer 2) of [Lasser, 2026c, Deployment Safety] can act on. The g-level reverse would strengthen the result but is not required for the deployment claim's operational utility.

7.3 What the latent-coalition residual would require

The (COAL) condition partitions counterparties whose correlations are detected by audit; the latent-coalition residual η_{latent} absorbs undetected-coordination risk as a *norm bound* on the residual error term $e_{\text{latent}} \in \mathcal{V}$ from Assumption 2: $\|e_{\text{latent}}\| \leq \eta_{\text{latent}}$.

Bounding η_{latent} in a specific deployment requires calibrating the audit's detection power against a worst-case deviation magnitude: $\eta_{\text{latent}} \leq p_{\text{miss}} \cdot \|\Delta\|_{\max}$ where p_{miss} is the probability that a coalition of given size escapes detection thresholds and $\|\Delta\|_{\max}$ is the worst-case per-counterparty deviation in the deployment's threat model. The audit specifies the detection thresholds, the detection-power model, and the one-sided upper-confidence procedure in (COAL)'s procedure (§6); p_{miss} and $\|\Delta\|_{\max}$ are calibrated through that procedure rather than read off raw operating statistics, and the certification semantics this paper inherits do not derive their values from first principles.

7.4 Connection to existing Goodhart literature

The Concentration-Gap selection theorem fits into the broader formal-Goodhart literature in a specific position:

- It extends [Manheim and Garrabrant, 2019]'s Adversarial Goodhart variant by giving an explicit upper bound on the proxy-truth divergence in terms of selection concentration, where Manheim & Garrabrant give a qualitative taxonomy.
- It complements [El-Mhamdi and Hoang, 2024]'s tail-based formal Goodhart by addressing the *selection-induced* gap (where this paper sits) rather than the *tail-event* gap (where El-Mhamdi & Hoang sit). The two are independent failure modes; both should hold for the deployment claim to bind.

- It fits [Majka and El-Mhamdi, 2025]'s Strong/Weak/Benign Goodhart classification at the "Weak Goodhart" boundary: the proxy-truth gap is bounded under structural conditions, but the conditions can fail (Strong Goodhart) or be irrelevant (Benign Goodhart).

The selection theorem does not subsume any of these works; it combines their perspectives into a single deployment-relevant result.

7.5 Connection to bounded co-evolution failure modes

Theorem 1 replaces the free-floating-constant assumption of bounded co-evolution with a derivation from primitive parameters. The failure modes of bounded co-evolution now factor through specific structural conditions:

- (C5.HOEFF) failure: per-step LLR is unbounded, so B_{clip} is undefined. Detection: SPRT-applicability audit (paper 10 Audit 7) catches this.
- (C7.RATE) failure: exposure event count grows faster than $\lambda_{\max} \cdot \tau_{\text{meta}}$. Detection: (C7.RATE) audit (this paper, §6) catches this.
- Channel-projection failure: ledger event-classification policy doesn't respect the zero-outside-engagement convention. Detection: structural-projection audit (this paper, §6) catches this.

This decomposition is the structural payoff of Theorem 1: bounded-co-evolution failure is no longer an opaque single-mode risk but a triage of three specific structural-condition failures, each with its own audit detection.

Author Contributions

Teague Lasser owns the paper's intellectual direction and is responsible for all claims made.

Claude Opus 4.7 (Anthropic) drafted the paper under that direction.

GPT 5.5 (OpenAI) served as cold technical reviewer for proof errors and claim mismatches.

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

El-Mahdi El-Mhamdi and Lê-Nguyên Hoang. On Goodhart's law, with an application to value alignment. *arXiv preprint arXiv:2410.09638*, 2024.

Orris C. Herfindahl. *Concentration in the Steel Industry*. PhD thesis, Columbia University, 1950.

Albert O. Hirschman. The paternity of an index. *The American Economic Review*, 54(5):761–762, 1964.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Teague Lasser. Exogenous verification for alignment: Cryptographic commitments on substrate-exclusive channels. `https://teague.info/papers/exo/`, 2026a. Preprint, accessed April 2026.

Teague Lasser. Goal-frontier maximization as a microfoundation for welfare economics. https://teague.info/papers/microfoundation/, 2026b. Preprint, accessed May 2026.

Teague Lasser. Goal-frontier maximization: A provably safe regime for capability-unbounded deployment. https://teague.info/papers/paper10/, 2026c. Preprint, accessed May 2026.

Teague Lasser. An aggregate B-to-C lower bound from revealed-sacrifice observation. https://teague.info/papers/revealed-sacrifice/, 2026d. Preprint, accessed May 2026.

Teague Lasser. Need-sufficiency architecture and gap diagnostics for the B-to-C gap. https://teague.info/papers/need-sufficiency/, 2026e. Preprint, accessed May 2026.

Teague Lasser. Cross-substrate channel redundancy governs monopolar convergence. https://teague.info/papers/phase/, 2026f. Preprint, accessed April 2026.

Teague Lasser. Computable goal frontiers and the gradient toward civilization-building. https://teague.info/papers/poset/, 2026g. Preprint, accessed April 2026.

E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, 3 edition, 2005.

Adrien Majka and El-Mahdi El-Mhamdi. The strong, weak and benign Goodhart's law: An independence-free and paradigm-agnostic formalisation. *arXiv preprint arXiv:2505.23445*, 2025.

David Manheim and Scott Garrabrant. Categorizing variants of Goodhart's law. *arXiv preprint arXiv:1803.04585*, 2019.

Abraham Wald. *Sequential Analysis*. John Wiley & Sons, 1947.