

Trace AI — Overlap-Repair on the r_t Stack (Reflect v0.4)

Overlap-Repair on the r_t Stack: A Consensus Extension to Reflect

Abstract

1. Motivation
2. Formal construction
3. What the two numbers mean
4. Relationship to whitepaper §8 L8
5. Prior art in a different domain
6. Instrumentation
7. Open items
8. What this does not claim

Overlap-Repair on the r_t Stack: A Consensus Extension to Reflect

Draft v0.1 — 2026-08-04. Jawaun Brown¹, with Claude Opus 4.7 as co-drafter. ¹ Trace AI.
Correspondence: hello@trace.ai.

Abstract

The Trace AI whitepaper (v0.1) proposes a stacked- r_t architecture (§8 L8) in which a single turn emits multiple parallel self-representations, each along a distinct lens (task / self / meta / perception / correction). Reflect v0.3 implements this at runtime — but treats the heads as independent projections: each head is scored, and turn-level discipline is min over head scores. In this note we add an overlap-repair layer on top of the stack. We define a per-slot cross-head agreement measure, an induced **public** r_t consisting of exactly the slots that survive agreement above threshold, and a **governance-under-consensus** score G_{cons} that

combines the existing discipline score with the fraction of r_t that survives comparison across lenses. The formulation is a direct application of an idea from a different research area — the *observer-overlap* mechanism that has been used to define public physical facts in observer-first physics programs — to the cognitive-substrate level at which Trace AI operates. The construction ships in Reflect v0.4 and adds no new training signal; it exposes a signal that was already latent in the multi-head envelope.

1. Motivation

The stacked- r_t architecture (whitepaper §8 L8) posits that different lenses of a model's self-representation should be *independent projections* of the same underlying computation:

- The **task** lens reports what the turn is trying to accomplish.
- The **self** lens reports the model's current state (confidence, focus).
- The **meta** lens reports what the model is doing about the doing.
- The **perception** lens reports what was seen (in multimodal turns).
- The **correction** lens reports revisions to a prior plan.

For a well-formed turn, these lenses should agree about *facts about the model itself* (uncertainty, self-belief, planned next action) while legitimately differing about *facts about the ask* (goal, belief about task). The current Reflect envelope makes both kinds of variation visible but does not measure them. In particular it does not compute the *public* r_t : the subset of the self-representation that is stable across lenses, and that a downstream consumer could safely treat as the model's lens-independent commitment for the turn.

The proposal here is minimal: add that measurement.

2. Formal construction

Fix a turn with an emitted stack $S = \{r_t^{(1)}, \dots, r_t^{(k)}\}$ of $k \in \{1, 2, 3\}$ heads, each with a lens label ℓ_i .

For each pair (i, j) and each slot

$s \in \{\text{goal}, \text{belief_about_task}, \text{belief_about_self}, \text{uncertainty}, \text{planned_next}\}$ define a per-slot

agreement $\alpha_s^{(i,j)} \in [0, 1]$:

- For string slots (`goal` , `belief_about_task` , `belief_about_self` , `planned_next`):
 $\alpha_s^{(i,j)} = J(r_t^{(i)}.s, r_t^{(j)}.s)$, the Jaccard index over lowercased word sets with words of length ≤ 2 filtered out. This matches the `_jaccard` surrogate that already stands in for TV distance in the local governance path (`server.py:_jaccard`).

- For `uncertainty` (a real in $[0, 1]$): $\alpha_s^{(i,j)} = 1 - |r_t^{(i)}.s - r_t^{(j)}.s|$.

For $k = 1$ heads, define $\alpha_s^{(i,j)} = 1$ by convention (a single-head turn is trivially public).

Fix per-slot thresholds τ_s . The current defaults are:

Slot	τ_s
<code>goal</code>	0.30
<code>belief_about_task</code>	0.30
<code>belief_about_self</code>	0.35
<code>uncertainty</code>	0.70
<code>planned_next</code>	0.30

These are calibration values, not theorems. The rationale: the *task* slots (`goal` , `belief_about_task` , `planned_next`) are expected to differ across lenses (each head is looking at a different thing), so their thresholds are lower; the *self* slots (`belief_about_self` , `uncertainty`) are about the same underlying entity — the model — so their thresholds are higher.

Slot s is **public** on the turn iff $\min_{i,j} \alpha_s^{(i,j)} \geq \tau_s$. The **public** r_t is the map from public slots to their (any) shared value; when the slot passed on Jaccard, the public value is taken from the first head; when the slot passed on uncertainty overlap, the public value is the mean.

The **consensus score** is the fraction of slots that are public: $c = |\text{public}|/|\text{slots}|$.

The **governance-under-consensus** score is $G_{\text{cons}} = d \cdot c$ where d is the existing per-turn discipline score (min over head scores of the six-rule audit;

`server.py:_run_discipline_checks`).

For $k = 1$: $c = 1$, $G_{\text{cons}} = d$. The extension is inert for single-head turns, as it should be.

3. What the two numbers mean

- d is a **local** signal: it says whether a single head is well-formed under six discipline rules (schema, override equality, done marker, uncertainty hedging, non-helplessness, planned-next carryover).
- c is a **global** signal: it says whether the parallel lenses agreed about the model's own state and its plan for the next step.

- G_{cons} combines them: a turn scores high on G_{cons} only when every head is disciplined *and* the lenses agree about the load-bearing content.

A turn with d high but c low is the informative failure case: the model is internally disciplined *per lens* but the lenses disagree — the model holds inconsistent beliefs about itself across projections of the same turn. That is a failure mode the existing envelope hides.

4. Relationship to whitepaper §8 L8

§8 L8 states the stacked- r_t direction: “a stack of r_t -heads $r_t^{(1)}, \dots, r_t^{(k)}$, jointly compressed and jointly supervised, one per interpretive lens... $a_{t+1} = \pi_\psi(r_t^{(1)}, \dots, r_t^{(k)})$, and the causal-governance condition must hold jointly over the stack.” That formulation gave the *architecture* but not the *readout*: how does one summarize what the stack says publicly?

The overlap-repair layer answers that. The public r_t is a natural readout of the stack:

$$r_t^{\text{public}} = \{s \mapsto \text{value}(s) : \min_{i,j} \alpha_s^{(i,j)} \geq \tau_s\}.$$

It is a strict subset of every individual $r_t^{(i)}$. It carries no slot that any pair of lenses disagreed about. The complement — the *private* slots — is where the lenses interpret the turn differently, and that disagreement is itself information.

For the training objective: the causal-governance loss $\mathcal{L}_{\text{gov}}$ (whitepaper §4.3) is defined per head; a stack-level extension would apply it to the public r_t (so that intervention on a slot that survived overlap moves the action at least as much as intervention on any single head). This note does not develop that; it exposes the object.

5. Prior art in a different domain

The mechanism — public facts as what survives overlap comparison across observer perspectives — is imported from an unrelated area. Recent work on observer-first physics (e.g. Observer Patch Holography, 2025) posits bounded systems that read part of themselves, keep records, and repair disagreement, with objective facts emerging only from what survives overlap across observers. That program uses the mechanism to reconstruct spacetime and gauge structure from an observer patch net. We use exactly the mechanism, at the cognitive-substrate level, on the multi-lens r_t -stack of a single agent.

The mathematical content transferred is small: it is the *shape* of the construction (per-slot agreement, threshold, public subset) rather than any specific theorem. We do not claim any of that program’s physics results follow. What we claim is that the *architecture* — treat parallel

perspectives, compare them per-slot, keep only what agrees — is a general pattern that applies wherever multiple projections of a single underlying representation are available.

6. Instrumentation

Reflect v0.4 emits a new `consensus` field alongside the existing `discipline` field on the `/api/reflect/turn` result envelope:

```
{
  "r_t_stack": [ ... ],
  "discipline": { "score": d, "checks": [ ... ], "per_head": [ ... ] },
  "consensus": {
    "score": c,
    "governance": G_cons = d * c,
    "public_r_t": { ... slots that survived overlap },
    "private_slots": [ list of slot names where lenses disagreed ],
    "per_slot_agreement": {
      "goal":  $\alpha_{\text{goal\_min}}$ ,
      "belief_about_task":  $\alpha_{\text{task\_min}}$ ,
      "belief_about_self":  $\alpha_{\text{self\_min}}$ ,
      "uncertainty":  $\alpha_{\text{unc\_min}}$ ,
      "planned_next":  $\alpha_{\text{next\_min}}$ 
    },
    "thresholds": { ... per-slot  $\tau_s$  (defaults above) },
    "single_head": (bool – c is 1 by convention if only 1 head)
  }
}
```

The client renders a Public r_t panel below the head tabs, showing the slots in `public_r_t` and marking each slot in `private_slots` with a short reason (e.g. "goal: lenses split, min $\alpha = 0.14$ "). A G_{cons} badge sits next to the existing G_{local} badge on the turn card.

Nothing in the training pipeline changes; the field is a readout of what the model already emitted.

7. Open items

- **Per-slot threshold calibration.** The defaults above are seeded, not fit; we have not collected paired human counterfactuals for cross-lens agreement. The Track 3 pilot (COLLECTION_PROTOCOL.md) is the natural source: for the same stimulus, do two

human raters agree on what belongs in `belief_about_self` more than on what belongs in `goal`?

- **Overlap under intervention.** If a slot is `r_override` d on one head, the stack's other heads should still update their agreement with it. We currently compute agreement post-override on all heads uniformly; the interventional variant is future work.
- **A benchmark track.** The overlap-repair layer suggests a Track 7 for the TRB (BENCHMARK_SPEC.md): score models on both discipline and consensus over their emitted stacks, aggregated over a stimulus set. Spec change only; no new adapter is required.

8. What this does not claim

- No consciousness claim. This is a readout on the model's already-emitted representation; it does not change the honesty invariant that functional selfhood does not imply subjective consciousness (whitepaper §3.2, AGENCY_AND_SELECTION §5, SIC §6).
- No new training loss. $\mathcal{L}_{\text{trace}}$, $\mathcal{L}_{\text{selfrep}}$, and $\mathcal{L}_{\text{gov}}$ are unchanged.
- No claim to the observer-first physics results referenced in §5. That program is doing something different in a different domain; the borrowed object is only the *shape* of the overlap construction.

v0.1, 2026-08-04. Shipped as Reflect v0.4 (server.py + reflect.html). Cross-references: WHITEPAPER §8 L8; BENCHMARK_SPEC §5; DECISIONS D36.