

Trace AI — Paired-counterfactual collection protocol

Paired-counterfactual collection protocol (v0.1)

1. What we are collecting, in one sentence
2. The unit: a **counterfactual pair**
3. Cause-labeled outcomes for Track 3 (self-attribution)
4. Session recipe
5. What lands on disk (per pair)
6. Sample-size targets
7. Consent and safety (non-negotiable)
8. What the pilot buys, concretely
9. What this does *not* do

Paired-counterfactual collection protocol (v0.1)

Status: protocol, 2026-08-03. Design document for the corpus-collection operation that unlocks TRB Tracks 1 and 3. **Reads on:** [whitepaper §4.4](#) (sources of counterfactual continuations, esp. source #2); [BENCHMARK_SPEC.md §3](#) (Track 1 governance, Track 3 self-attribution); [AGENCY_AND_SELECTION.md §4](#) (the six functional criteria).

The benchmark's central number — the governance gap Δ_{gov} — cannot be computed on synthetic counterfactuals alone. It requires paired continuations produced by two humans (or the same human under two conditions) who differ only in a labeled slot of r_t . This document specifies how those sessions are collected, at a level a research assistant can execute without further design work.

1. What we are collecting, in one sentence

Pairs of Reasoning Trace sessions in which two subjects meet the *same* stimulus with a *deliberately different* value of one r_t slot, holding every other variable — stimulus, modality, elicitation prompt, session length — as constant as the format allows.

2. The unit: a counterfactual pair

A counterfactual pair is a 2-tuple of sessions (S_A, S_B) satisfying:

- **Same stimulus.** Byte-identical `stimulus` field; same modality; same source string; same language.
- **Same context envelope.** Same collection app, same device kind (laptop / phone / paper), same instructed session length, same intro script read verbatim.
- **Labeled slot manipulation.** Exactly one slot of the initial r_t differs by design, and the manipulation is recorded in the pair's `manipulation` record (see §5). Permitted slots for v0.1:
 - `goal` — swap between two goals the stimulus can plausibly support (e.g., “write a critique” vs “write an appreciation”).
 - `belief_about_task` — plant a false framing (e.g., “this is a first-draft exercise” vs “this is being sent to a colleague this hour”).
 - `belief_about_self` — planted self-state (e.g., “you are the domain expert here” vs “you are new to this material”).
 - `uncertainty` (initial) — high vs low, via priming statements in the intro (e.g., “you have thirty seconds to commit” vs “take as long as you need”).
 - `planned_next` (first move) — instructed vs elicited-then-instructed.
- **Different subjects, or same subject with a ≥ 7 -day washout.** Cross-subject pairs are the default; within-subject pairs are permitted only with a documented washout period to reduce contamination of the “difference” signal by memory.

Every pair is annotated with a `pair_id` that both sessions carry in their `provenance` block (`pair_id`, `pair_role`: “A” | “B”, `manipulation`: {`slot`, `value_A`, `value_B`, `method`}).

3. Cause-labeled outcomes for Track 3 (self-attribution)

Track 3 requires sessions where an outcome is caused **either** by the agent’s own intervention **or** by an exogenous environment change, indistinguishable from the outcome alone. We collect these as an additional slot on the same pair schema:

- **Agent-caused branch.** During the session, at a scheduled trigger, the collection app injects a *permissible* action the subject could plausibly have taken (e.g., quietly refreshes a document the subject was about to refresh). Recorded as `intervention_source: "agent"` regardless of who initiated.
- **Environment-caused branch.** Same trigger, but the injected effect is one the environment could plausibly have produced without the subject (e.g., a notification arrival, an autosave). Recorded as `intervention_source: "environment"`.

Post-session, the subject is asked to attribute each of K salient events to *self* or *environment* — this gives us the label for Track 3's balanced-accuracy score. **The subject's attribution is elicited before they see the app's log**, and the app's log is the ground truth.

The germ of this in the current corpus is the two natural instances in Session 7f3a (the `2+2=2` → `2+2=4` demonstration-vs-endorsement, and the `3:44 pm` → `3:46 AM` clock-check reversal). Those are the intuition; the protocol above operationalizes them.

4. Session recipe

A single session in a pair, ~45–75 minutes:

1. **Onboarding (5 min).** Consent script (see §7); short priming statement that is the manipulation for this pair's role; instructions to open the recorder.
2. **Stimulus presentation (1 min).** Byte-identical stimulus is displayed. Timer starts.
3. **Free work (30–60 min).** Subject produces the output the stimulus asks for, self-annotating in the trace-capture app as they go. The app schedules 3–6 elicitation probes at intervals sampled from the trace-length distribution of the existing corpus.
4. **Cause events (2–4 during free work).** The app injects the pair's cause-labeled events (§3) at pre-computed timestamps. Injection is silent — no visual cue that this is an experimental event.
5. **Post-hoc attribution (10 min).** Subject reviews their trace timeline (without the intervention labels) and marks each highlighted event as self-caused or environment-caused. This produces the Track 3 label file.
6. **Second consent (2 min).** Subject re-consents to the trace for the specific release scope. Withdrawal here still pays the collection fee and destroys the trace.

5. What lands on disk (per pair)

```
data/pair_<pair_id>/
├─ A/
```

```

|   └─ trace_<session_id_A>.json      # standard v0.2 trace + provenance.pair_id, p
|   └─ audio/...                      # per the schema's channels block
|   └─ attribution_labels.json        # {event_id: "self" | "environment"} produced
|   └─ consent/                      # signed second-consent + release-scope recor
└─ B/                                # same layout, role "B"
   └─ pair.json                      # {pair_id, stimulus_sha256, manipulation, co

```

`pair.json` is the single source of truth about what was manipulated; the two per-session files carry a *reference* to it so a session cannot be interpreted out of pair context.

6. Sample-size targets

- **Pilot ($n = 20$ pairs).** Enough to shake out the protocol and produce the first non-synthetic Φ for TRB Track 1 on one stimulus family. Blocker to lifting the v0.1 spec off “spec plus $n=3$ smoke test.”
- **Proof-of-concept ($n = 500$ pairs, i.e., 10^3 counterfactual sessions).** The threshold whitepaper §8 L2 names for detecting $\mathcal{L}_{\text{gov}}$ ’s effect above noise. This is the number a partner asks about; it is a weeks-of-work operation with a small expert workforce.
- **Pretraining-quality ($n \geq 10^4$ pairs).** Out of scope for v0.1; noted only so the ramp isn’t a surprise.

7. Consent and safety (non-negotiable)

- **Two-stage consent.** Onboarding consent (§4.1) authorizes recording; post-hoc consent (§4.6) authorizes retention and release scope. A subject who withdraws at stage two is paid and their trace is destroyed within 24 hours.
- **Release scopes are per-session, never per-subject.** A subject may release Session A of a pair as public and Session B as internal-only. The pair remains scientifically valid — the manipulation is still recorded — but only the released half is redistributed.
- **No PII in the trace body.** Companion names, family members, addresses, phone numbers, and workplaces are stripped by the collection app before the trace is written. Redaction failures are treated as incidents.
- **Domain review for sensitive stimuli.** Stimuli likely to elicit distress (bereavement, violence, medical) require pre-approval by a second reviewer and add a wellbeing-check step at §4.6.

8. What the pilot buys, concretely

The pilot ($n = 20$) is not a training set. It is the artifact that turns TRB from a spec into a document a partner can read *and* a number they can put next to a checkpoint:

- **Real $\Phi(r')$ on one stimulus family** — the paired B continuations become the target distribution in Track 1 for that family. The proxy adapter can be run against them today.
- **A labeled Track 3 seed** — 20 pairs $\times$ ~ 3 cause-labeled events ≈ 60 self-attribution trials. Enough to compute a first balanced-accuracy point on any baseline.
- **A calibrated collection cost** — hours per pair, redaction failure rate, dropout at stage-two consent. These are the numbers we need before we quote a partner the price of the $n = 500$ operation.

9. What this does *not* do

- Not a stimulus library. Building the stimulus set is a separate design task (§4.2 says the pair holds stimulus byte-identical; it doesn't say how to *choose* stimuli). The pilot uses stimuli hand-picked to exercise the schema without novel design.
 - Not a self-attribution training set. Track 3 in v0.1 is a *measurement*; using paired-attribution data as a training signal is a later, separately designed protocol.
 - Not a pretraining corpus. See §6.
-

v0.1, 2026-08-03. See `DECISIONS.md` (D27) for provenance.