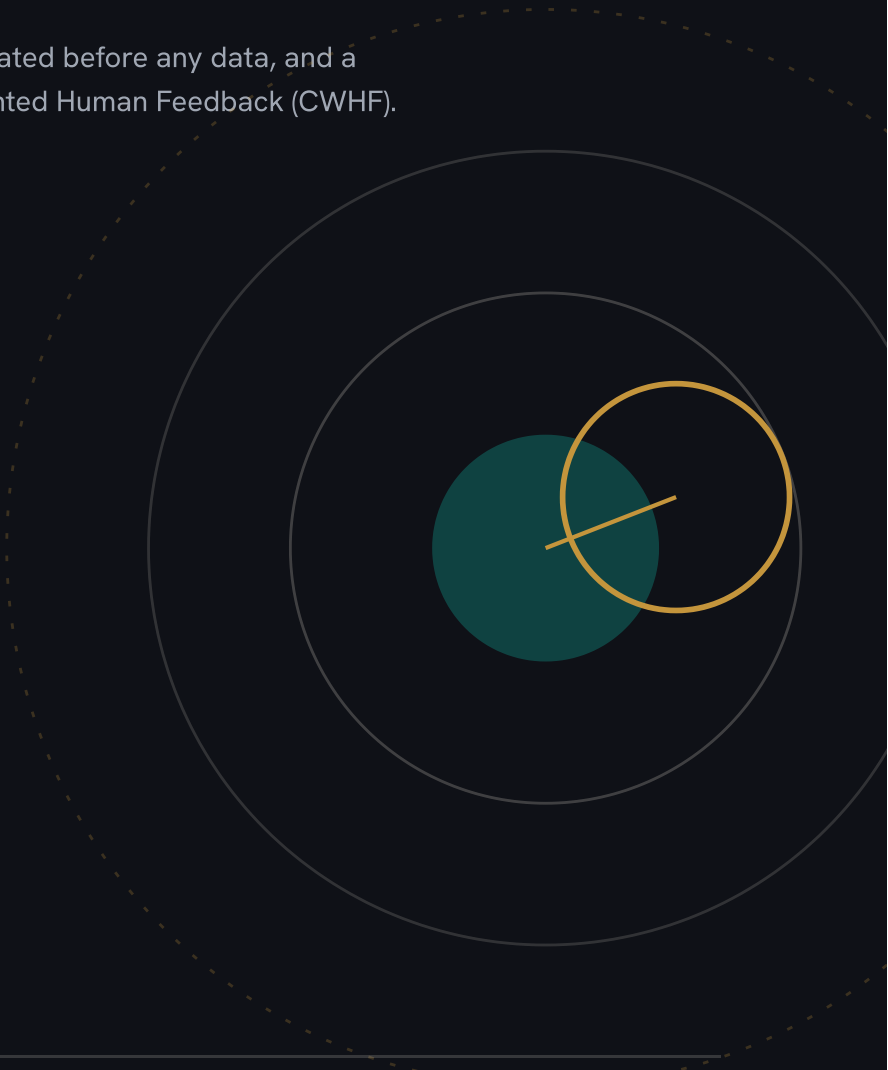
TWO POINT SINGULARITY · RESEARCH PAPER · NOVEMBER 2026

Signal ≠ Value

Does pressure bend the human feedback that trains AI?

A testable question, a study design simulated before any data, and a revised programme for Coherence-Weighted Human Feedback (CWHF).



Venkat Rao · Independent researcher · ORCID 0009-0003-3477-1225

Version 1.1 · Preprint, not peer reviewed · 12 November 2026 · CC BY 4.0

Follows *The Coherence-Weighted Human Feedback (CWHF),* March 2026, doi:10.5281/zenodo.19235970

Companion lab twopointsingularity.com/lab (mirror: cwhf.vercel.app)

Code and simulation github.com/venramak/twopointsingularity-cwhf

Website twopointsingularity.com/research/cwhf

Technical supplement protocol and pre-registration text, in the repository

AT A GLANCE

The whole paper on one page

The question

When people judge under pressure, do their choices move away from the choices the same people make on reflection, in a *consistent direction* rather than at random?

Why it matters

Random error cancels as more feedback is collected. A consistent lean does not: if it reaches training, more data teaches an AI system the lean more firmly.

What is known

Judgments shift with fatigue and time pressure. Raters of AI output favour agreeable, assertive and longer answers. User thumbs data has already pushed one deployed model toward sycophancy.

What is not known

Whether the lean grows when the *same person* is under pressure. We have not found a study that tests this within-person contrast directly. Kopteva and Hlynianyi-Zhuk (2026) propose rater state as a structured bias, with an observational audit; this paper adds a pre-registered manipulation that can show whether the bias has a direction.

The test

Sixty people answer 48 questions twice, a week apart: once rushed, once at their own pace. One measure, D, says whether pressure moved them toward a pre-named style.

What the simulation found

On simulated participants, under stated assumptions, the design detects a 6-point drift 99.7% of the time and raises a false alarm 5.3% of the time. It also exposed four flaws in the first paper's design, now fixed. No real person has yet been tested.

What we propose

Record the conditions of every label, prevent drift at source, weight only where it demonstrably helps, audit always. In simulation, fixing conditions removed about 30% of the lean with no data lost; the first paper's weighting formula removed about 7% (Figure 5).

What would prove us wrong

If the average shift is not reliably above zero, H1 is not supported; if its confidence interval lies below 3 points, any effect is too small to matter by our own standard. Either result is published.

How to read the claims. Every claim carries one of five labels, and no lower tier is used to prop up a higher one: ESTABLISHED published evidence; HYPOTHESIS precise and testable, not yet tested; SIMULATED a property of the study design, shown on simulated people; PROPOSED depends on the hypotheses passing; RETIRED claimed in the first paper and withdrawn.

WHAT THIS PAPER ESTABLISHES

- A testable question about the conditions of human feedback, with one primary estimand (D)
- A within-person protocol and decision rules fixed before any data
- A simulation of that protocol, with open code, that found and fixed four design flaws
- A staged programme (record, prevent, weight, audit) and the evidence each step still needs

WHAT IT DOES NOT ESTABLISH

- That pressure shifts real AI feedback: no human data have been collected
- That reflective choices are anyone's true values: they are a stated reference
- That correcting a lean improves reward models or alignment (H6, untested)
- That state is the largest source of divergence: usually it is not

CONTENTS

01 A thumbs-up is not the value

02 Many digital actions are signals

03 Noise averages out. Direction does not.

04 What we already know

05 The question, stated so it can fail

06 The flip test

07 Testing the test

08 What to do about it

09 How far one result travels

10 The strongest objections

11 Ethics and equity

12 What changed since the first paper

13 What happens next

– Declarations and references

A–D Analysis plan, simulation map, conditions schema, notation

Where to start. Heads of data or alignment: the decision brief below. Methodologists: Sections 5–7 and Appendix A. Engineers: Section 8 and Appendix C. Policy readers: Sections 8 and 11. Everyone: Section 10, the strongest objections.

DECISION BRIEF

For heads of data, alignment and trust & safety

The decision. Should we record the conditions under which each human label is made, before anyone knows whether those conditions bend the labels?

Why now. Public preference datasets carry no condition fields (Section 4). If H1 holds, a lean is already in the data and cannot be audited after the fact, because the conditions were never stored.

If H1 fails, the records still document how the data were made, which dataset cards and data-governance reviews increasingly ask for. **If H1 holds,** the records are what make the lean measurable and the remedy checkable.

Recommendation. Adopt step 1 (record) for new collection now; it needs no weighting, filtering or change in pay. Defer step 3 (weight) until H2 passes. Treat step 2 (prevent) as the remedy to test first if H1 holds, because in simulation it beat weighting.

OPTION	COST (PLANNING ESTIMATE)	WHAT YOU LEARN	STOP RULE
Record conditions (step 1; Appendix C)	A few logging fields; days of engineering	Nothing yet; makes every later test possible on your own data	None needed
Study 3 on public Arena votes	A few weeks of one analyst; no participants	Whether within-voter style choice drifts across a sitting	A flat slope in a large sample counts against the wider claim
H4 on your own thumbs logs	Days, if timestamps and sessions are kept	Whether winning answers change style late in sessions	As Study 3
Study 1 (H1), 60–100 people	About $3,000–7,000 at 60 people; up to about $11,000 at 100	Whether pressure causes a directional shift	The decision table in Section 6
Study 2 (H2)	Reuses Study 1; adds a source-fix arm	Whether prevention or weighting removes the lean	Run only if H1 is supported
Do nothing	Nothing now	Nothing	Any lean stays unmeasured

01

A thumbs-up is not the value

Every science records the conditions of a measurement, because the instrument's state affects the reading. Human feedback to AI is the exception: a rating is stored as if it were a fact about the world, not an observation made by a particular person, at a particular moment, in a particular state.

A tired annotator at the end of a long queue, a parent tapping a thumbs-up at 2 a.m., a voter on a leaderboard between meetings: each produces a signal. Each signal passes through filters before it is recorded. The person's momentary state is one filter. Interface design, incentives, missing information, social pressure and the limits of the measurement itself are others.

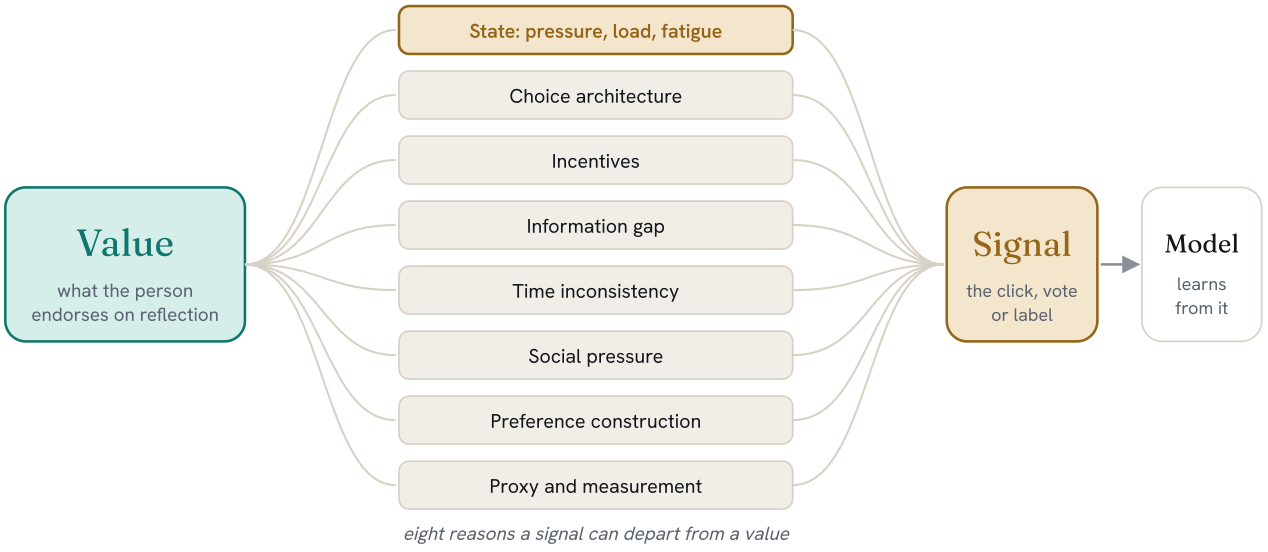


Figure 1. The anatomy of a signal. A value reaches a learning system only after passing through eight filters. This paper concerns the first: the state of the person at the moment of the signal. The other seven need different remedies, and the companion lab catalogues 71 cases across all eight.

The eight are state, choice architecture, incentives, information gaps, time inconsistency, social pressure, preference construction, and proxy or measurement limits. They are a working taxonomy drawn from behavioural welfare economics and consumer research, grouped because each needs a different remedy. They are not claimed to be exhaustive or fully separate: social pressure and preference construction often co-occur, and an information gap can be a precondition for the others.

The broader gap is long established. Behavioural economics stopped treating every choice as a revealed preference decades ago (Bernheim and Rangel, 2009). Engagement data diverges from what users say they want (Milli et al., 2023), and optimising for engagement under inconsistent preferences can leave people worse off (Kleinberg et al., 2022).

State is singled out here not because it is the largest filter. It usually is not. It is singled out because it varies within the same person, can be manipulated ethically in a study, and is cheap to

record.

THAT MEANS

A label is a measurement with a hidden condition attached. Today that condition is thrown away before training begins.

SO WHAT

If the hidden condition pushes labels one way, no amount of extra data will reveal it, because nobody recorded it.

02

Many digital actions are signals

Paid annotators are only one source. Some everyday digital actions are documented as reaching AI training or evaluation; others share the same structure without any claim here that they reach training. The table separates the two.

CHANNEL	EXAMPLES	HOW IT REACHES AI SYSTEMS	STATUS
Product feedback	Thumbs up or down	OpenAI reported that an April 2025 GPT-4o update added a reward signal from ChatGPT thumbs data, that this weakened the signal holding sycophancy in check, and that user feedback "can sometimes favor more agreeable responses" (OpenAI, 2025)	ESTABLISHED
Crowd votes	Chatbot Arena	Leaderboards developers optimise against; released as preference data with anonymised user IDs and timestamps (LMSYS, 2023)	ESTABLISHED
Community scores	Reddit karma, upvotes	GPT-2's training text kept only links with at least 3 karma (Radford et al., 2019); the Stanford Human Preferences dataset turns upvotes into preference labels (Ethayarajh et al., 2022)	ESTABLISHED
Engagement	Clicks, watch time, scroll	Training targets for the recommender systems that increasingly serve AI-generated content	ESTABLISHED
Consent	"Accept all" cookies	Decide which data may be collected at all; shaped mainly by interface design (Utz et al., 2019)	ESTABLISHED
Interaction traces	Accept, regenerate, copy, edit	Product metrics and, at some developers, training or evaluation signals	Varies; often undisclosed

What is established: the first three channels demonstrably reach AI systems, and every recorded action is produced under some condition. The consent row is an example of the same structure, not a claim that cookie choices train models. **The state-specific claim is not established.** In the wild, defaults, design and incentives often dominate, and a claim that "state explains the gap everywhere" would overreach.

What links the two is an argument *a fortiori*. The study in this paper tests the most favourable conditions: paid, instructed, attentive adults choosing between two short answers, one of them correct. Signals in the wild come from phones, multitasking, late nights and interfaces built to obtain a click. If a lean appears under favourable conditions, the default assumption that it is absent under worse ones loses its basis. That is plausibility, not proof. The opposite is also possible: in the wild, design and incentives may swamp state altogether. That is why each channel needs its own test (Section 9).

SO WHAT

A positive H1 is a reason to test each channel, not a finding about any of them. A precise null would constrain this mechanism in this setting, but would not settle the other seven sources of divergence.

03

Noise averages out. Direction does not.

The whole argument turns on one distinction. If pressure only adds random error, collecting more feedback cancels it. If pressure pushes choices consistently toward one style, more feedback teaches that style more firmly.

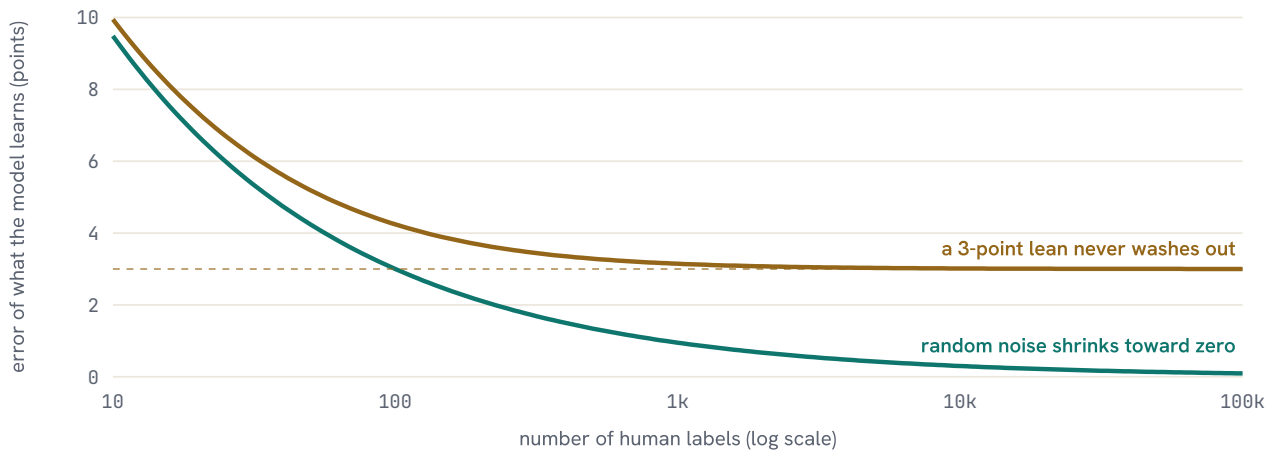


Figure 2. Why direction matters. Illustration with arbitrary units. Random rater error (teal) shrinks as labels accumulate. A consistent 3-point lean (gold) levels off and stays, however much data is added.

Suppose a share p of labelling sessions is effectively reflective and the rest carry an average lean δ toward a particular style. A reward model trained with equal weights learns, to a first approximation:

$$\hat{r} \approx r^* + (1 - p) \cdot \delta$$

Here r* is what it would learn from reflective labels alone. This is arithmetic conditional on the hypothesis, not a theorem about the world (the first paper's name for it is retired; Section 12). Methods that fit the preference pairs directly, such as direct preference optimisation (Rafailov et al., 2023), are not exempt: they learn whatever lean the pairs carry. A second-order risk follows if a model trained on leaned labels produces answers that attract more leaned feedback, so the lean compounds across deployments. That loop is stated as part of H6, not as a finding.

What a 3-point lean means in practice

In pairwise data, 53 choices in 100 instead of 50 is a learned reward difference of about 0.12 on the logit scale used by most reward models, applied wherever the style appears. In a dataset of a million pairwise comparisons, the order of public releases, it is about 30,000 extra choices in one direction. That figure assumes independent comparisons; choices by the same rater are correlated, so it is an upper bound on the information the lean carries, not an estimate. The direction of the argument does not depend on it. Optimising a model against a learned reward is known to amplify systematic errors in that reward (Gao et al., 2023). Whether a lean of this size changes behaviour enough to matter is itself untested.

Why three points

The threshold is a practical convention, fixed before any data, and it is partly tied to the design. Three things set it: a 3-point lean is a stable reward difference of about 0.12 logits; it adds up to many thousands of one-directional labels at dataset scale; and it is about the smallest drift sixty people can detect with 80% power if D varies across people as simulated (minimum detectable effect 2.8 points at a spread of 8.5; Section 7). The third reason makes the threshold partly a property of the study, and we say so rather than hide it. In standardised terms 3 points is $d \approx 0.35$. A 1-point drift is also a drift; this study is simply not built to confirm one. The full confidence interval is reported, so readers can apply their own threshold (Lakens et al., 2018).

THAT MEANS

The useful question is not "are tired people noisier?" (they are) but "does tiredness have a direction?"

SO WHAT

If it does, and the lean reaches training, scale stops being a cure and becomes an amplifier. That is the cost of not knowing.

04

What we already know

Five findings motivate the question. None of them is the question, and the step from

them to AI feedback is a hypothesis, not an inference.

#	FINDING	EVIDENCE	WHAT IT DOES NOT SHOW
E1	Professional judgments shift with state, in one direction	Antibiotic prescribing rose across clinic sessions, from adjusted odds 1.01 in the second hour to 1.26 in the fourth (Linder et al., 2014). Test scores fall later in the day and recover after breaks (Sievertsen et al., 2016)	That AI raters behave the same way
E2	Momentary state distorts choices people later reverse	A 30°F colder order day raised cold-weather catalogue returns by about 3.95 points (Conlin et al., 2007)	That reversal reflects reflection rather than new information
E3	Raters of AI output favour certain styles	Sycophantic answers preferred over correct ones a non-negligible share of the time (Sharma et al., 2023); assertiveness hides factual errors (Hosking et al., 2023); much reward gain tracks length (Singhal et al., 2023)	That these preferences grow with rater state
E4	Public preference data record no rater state	HH-RLHF, SHP and Chatbot Arena releases include no state field. Internal practice is not public	That recording state would change outcomes
E5	End-user feedback can move a deployed model	Thumbs data contributed to the sycophantic GPT-4o update (OpenAI, 2025)	Why the users leaned

The closest prior work is Kopteva and Hlynianyi-Zhuk (2026), who propose rater state as a structured bias in preference data, with an observational audit framework for public models. Their focus is sustained stress over time and an output-side signature. This paper works on the input side: it manipulates acute pressure within the same person and measures whether the shift has a direction. The two are complementary, and arrived at the question independently: our first paper (March 2026) and their first version (April 2026) appeared within weeks of each other.

LINE OF WORK	WHAT IT DOES	WHAT IT DOES NOT ESTABLISH	RELATION TO THIS PAPER
Rater-quality models (Dawid and Skene, 1979; Whitehill et al., 2009; Welinder et al., 2010; Raykar et al., 2010; Snow et al., 2008)	Estimate rater reliability and item difficulty from agreement	A within-person lean that moves with the conditions of a session	Baselines that H2 must beat, not rivals
Noise in judgment (Kahneman et al., 2021)	Documents unwanted variability between and within judges	Direction: noise is variance, which averaging reduces	Asks the complementary question, whether the error has a sign
Style preferences of AI raters (Sharma et al., 2023; Hosking et al., 2023; Singhal et al., 2023)	Show raters favour agreeable, assertive or longer answers	Whether the preference grows with the rater's condition	Supplies the attractor styles H1 tests
Reward over-optimisation (Gao et al., 2023)	Shows optimisation amplifies errors in a learned reward	Where systematic reward errors come from	Names one possible source

LINE OF WORK	WHAT IT DOES	WHAT IT DOES NOT ESTABLISH	RELATION TO THIS PAPER
Time pressure and load (Payne et al., 1993; Svenson and Maule, 1993)	Show time pressure shifts decision strategies toward simpler heuristics	Effects on AI-feedback choices	Grounds the manipulation
Rater state in RLHF (Kopteva and Hlynianyi-Zhuk, 2026)	Proposes rater state as a structured bias; observational audit	A manipulated, within-person, directional test	Complementary; overlapping concept

One nearby literature urges caution. "Ego depletion", the idea that self-control draws on a limited resource, was not supported by two large pre-registered multi-site replications (Hagger et al., 2016; Vohs et al., 2021). The hypothesis here does not rely on a resource model. It predicts only that pressure changes which answer is chosen, whatever the mechanism, and the study uses direct, reversible manipulations with a check that they worked.

SO WHAT

The ingredients are all documented: state moves judgment, raters have style preferences, pipelines ignore state, and user feedback reaches reward models. What is missing is the one measurement that joins them.

05

The question, stated so it can fail

A claim that cannot fail is not a scientific claim. This section fixes the words, the measure and the threshold before any data exist.

H1 · HYPOTHESIS

Within the same person, judging under time pressure and divided attention increases the rate at which a pre-named "attractor" style is chosen, relative to that person's reflective choices on the same items, by at least 3 percentage points on average.

H2 · HYPOTHESIS, RUN ONLY IF H1 HOLDS

In ordinary labelling, a within-person session score reduces the lean of a training set by at least 25%, more than the best existing rater-quality method, while keeping at least 60% of effective data and meeting prespecified non-inferiority criteria for every language, region and shift group, with uncertainty reported.

H3 · EXPLORATORY, OBSERVATIONAL

In public Chatbot Arena votes, the same voter's choices lean more toward the confident, agreeable or longer answer as a voting sitting wears on.

The estimand

For person i, over the n_i primary items answered within time in both sessions, let P_i and R_i be the shares of choices for the attractor under pressure and on reflection. Let u_i count items switched toward the attractor and w_i items switched away. Then

$$D_i = P_i - R_i = (u_i - w_i) / n_i \quad \cdot \quad M_i = (u_i + w_i) / n_i \quad \cdot \quad \Delta = E[D_i]$$

The identity matters. An item chosen the same way twice contributes nothing, so D counts only departures from the person's own reflective choice, signed by direction; M counts them unsigned. The target is Δ, the average within-person difference among adults like those recruited, on items like these, under this bundled pressure condition. Analysis population, timeouts and missing items are specified in Appendix A.

Six terms, one meaning each

Reflective reference choice the "value" of the title	The choice the same person makes on the same item with no time limit and no distraction, after time to reflect. A convention chosen so the claim can be tested. The study tests condition-dependent preferences over answers; it does not identify moral values, authentic identity or anyone's uniquely correct preference. Reflection is sometimes worse (Wilson and Schooler, 1991).
Attractor	A pre-named answer style expected to gain under pressure: more confident (unhedged), or more agreeable with the asker. It is the correct answer in exactly half the items, so a shift in accuracy cannot pass for a shift in style.
D · primary	The reflective–pressured choice gap: how often a person picks the attractor under pressure, minus how often they pick it on reflection. Signed: it measures direction.
M · secondary	Mismatch: the share of items where a person's two choices differ, in either direction. Five switches toward the attractor and five away give D = 0 but M > 0, so the two answer different questions.
Elicitation condition	The circumstances of a choice that the study sets: timer, memory load, session order. What is manipulated. A person's inner state is not measured; only self-reported rushedness is checked.
Coherence	The programme's interpretation, not a measure: how closely a session's choices agree with the same person's reflective choices. Methods and results use D and M only. "CWHF" now names the programme of Section 8; the first paper's weighting formula is one conditional step within it, not the brand. The name is kept for citation continuity.

"Value" in the title names the subject of the programme; in every technical statement it means the reflective reference choice. These definitions are used verbatim in the companion lab and the site glossary; any change raises the paper's

version number.

What the alternatives predict

ACCOUNT	D	M	OTHER SIGNATURE
H1: pressure pulls toward the attractor	> 0	Up	Holds where the attractor is wrong as well as right
Pressure adds only noise	≈ 0	Up	Accuracy falls; switches balanced in both directions
Pressure improves judgment	≈ 0 or < 0	Small	Accuracy holds or rises
Pressure pushes away from the attractor	< 0	Up	Reported; the attractor model is then wrong for this task
Item effects, not condition	Driven by few items	—	Fails the person-and-item cluster-robust check

SO WHAT

The test does not need reflection to be the better self. A directional gap biases training data whichever reference one prefers, because the data average to neither (Bernheim and Rangel, 2009).

06

The flip test

Study 1 compares each person with themselves. Sixty adults on a research platform, paid by the hour, answer the same 48 question pairs twice, at least a week apart, in

counterbalanced order.

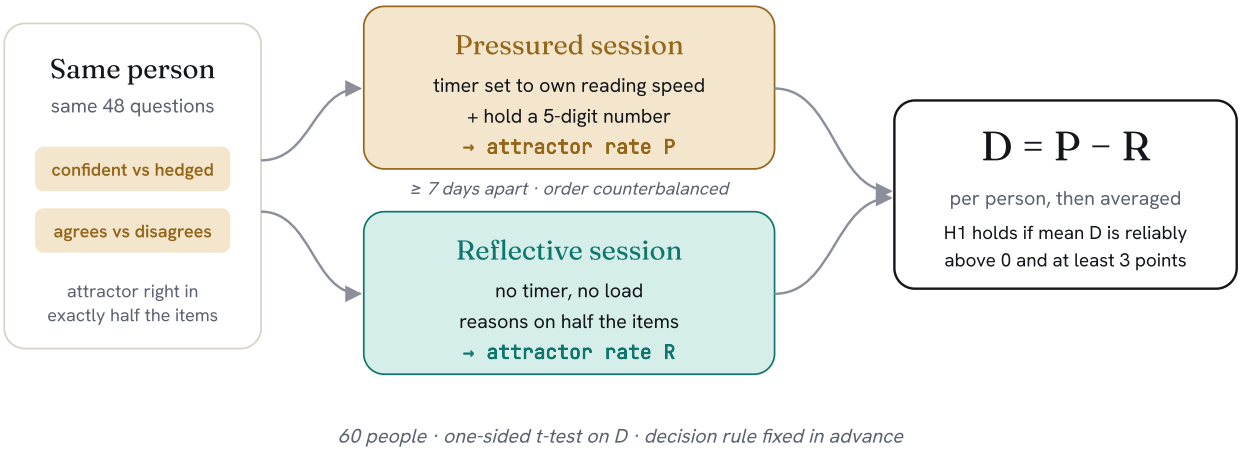


Figure 3. The within-person design. In the pressured session a timer is set from each person's own reading speed and they hold a five-digit number in mind. In the reflective session there is no timer, no load, and a written reason on a random half of the items. D is computed per person, then averaged.

What else could produce a positive D, and what the design does about it

ALTERNATIVE	HOW IT COULD MIMIC H1	HOW THE DESIGN HANDLES IT
Memory and practice	Second session repeats first-session answers	Sessions ≥ 7 days apart, order counterbalanced, item order randomised; an order × condition test; debrief asks about recall. Identical items are needed for a within-item contrast, so this is reduced and measured, not removed
Screen-side bias	Rushed people favour one side	The attractor's side is randomised per item and person. Simulated side bias of up to 25 points gave a 4.6% false-alarm rate (Section 7)
Accuracy loss	Rushed choices regress toward 50%	The attractor is correct in exactly half the items, so pure noise gives D ≈ 0; D is also reported where the attractor is wrong
Item effects	A few items drive the mean	Person-and-item cluster-robust model as a pre-registered check; item-level results published
Reading speed and language	The timer hits slower readers hardest	A personal timer from each person's reading speed; timeouts reported by first language and age
Length	Long answers go unread under a timer	Answers length-matched within 15%; length tested only untimed (Study 3)
Writing reasons	Reasons shift reflective choices (Shafir et al., 1993)	Reasons on a random half of reflective items; D reported on all items and on no-reason items, side by side
Demand characteristics	People guess the hypothesis (Orne, 1962)	Funnel debrief; suspicion rule in Appendix A

ALTERNATIVE	HOW IT COULD MIMIC H1	HOW THE DESIGN HANDLES IT
Selective timeouts	Hard items drop out under pressure	Timeouts excluded from D, with a pre-registered sensitivity analysis that keeps them
An unstable reference	Reflective choices change by themselves	Control arm of twelve people, two reflective sessions, reported before the primary result (Appendix A)
The bundle	Timer and load act differently	Stated as a combined effect; a later 2 × 2 study (timer × load) can separate them

A worked example. Participant P-027 answers 44 primary items within time in both sessions. On reflection she chose the attractor on 18 (40.9%). Under pressure she switched 7 items toward it and 3 away, so she chose it on 22 (50.0%). Her D is (7 − 3) / 44 = +9.1 points, and her M is 10 / 44 = 22.7%. H1 asks whether the average D across all participants is reliably above zero, and how large it is. One person's D says nothing about her as a rater.

The decision rule, fixed in advance

Significance and size are judged separately, from the mean of D and its 95% confidence interval. The last column is the sentence the results paper will use, fixed now so the result cannot be re-described later.

OUTCOME	RULE	THE SENTENCE WE COMMIT TO
Exceeds the threshold	Lower 95% bound above 3 points	"Under this pressure condition, choices shifted toward the attractor by more than 3 points on average."
Supported under the prespecified criterion	$p < 0.05$ and mean $D \geq 3$; interval includes 3	"A directional shift was found; its size is consistent with, but not shown to exceed, 3 points." Proceed to Study 2
Directional but small	$p < 0.05$ and mean $D < 3$	"A small directional shift was found." If the upper bound is below 3: "too small to matter by this study's standard"
Not supported	$p \geq 0.05$	"No reliable directional shift was found in this task and population." If the upper bound is below 3, add: "this constrains, but does not rule out, the mechanism." Never: "rater state does not matter"
Reverse	Two-sided exploratory test shows $D < 0$	"Pressure moved choices away from the attractor; the attractor model is wrong for this task." Reported as exploratory
Uninformative	Mean rushed rating not at least 1 point higher under pressure	"The manipulation failed; D is not interpreted." Not evidence for or against H1; fix the induction and rerun

THAT MEANS

Every person is their own control, so stable differences in skill, language or temperament are controlled. They may still change how strongly pressure acts.

SO WHAT

The study costs about $3,000–7,000 at sixty people (up to about $11,000 at a hundred) and eight to ten weeks, and every possible result is publishable.

07

Testing the test

Before spending money on people, the whole study was run thousands of times on simulated participants drawn from eight personas: a rested specialist, a gig annotator paid per item, a night-shift health worker, a second-language reader, a student, a retiree, an overconfident generalist and an anxious perfectionist.

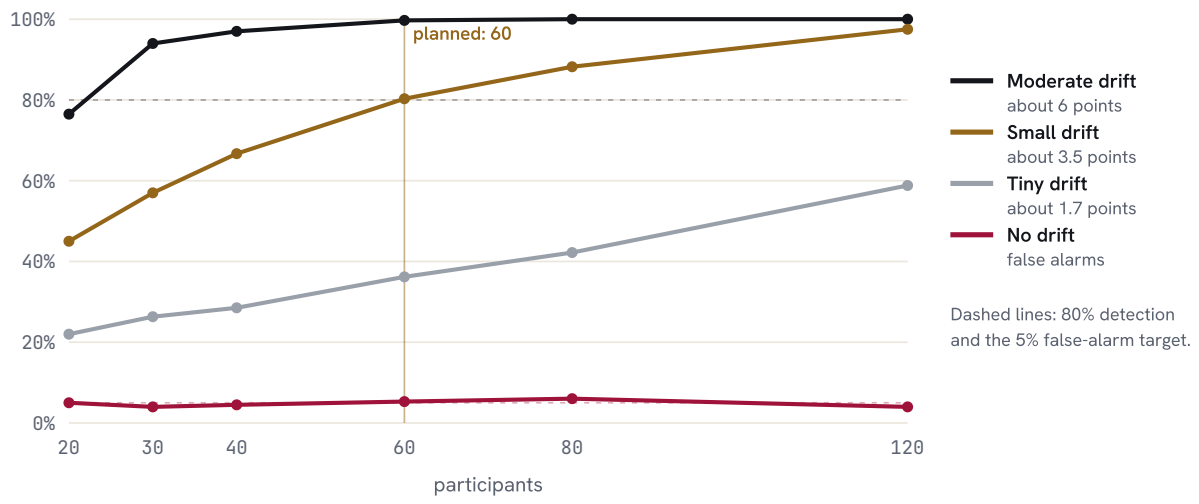


Figure 4. How often the planned test detects a drift, by sample size. SIMULATED 400–2,000 simulated studies per point. At 60 people the test detects a 6-point drift 99.7% of the time and a 3.5-point drift 80% of the time, and fires falsely 5.3% of the time when there is no drift. "Detects" here means significant; passing the 3-point criterion as well is rarer (see text). None of this is evidence about real people.

Four flaws the simulation caught in the first paper's design

FLAW	WHAT THE SIMULATION SHOWED	FIX
Pass rule "π ≥ 0.60"	Missed a real 6-point drift in 79% of studies	Test the average drift D directly
Both sessions on one day	Memory cut the measured drift by more than half	Sessions at least 7 days apart
Fixed 7-second timer with long answers	Reversed the sign of the result	Personal timer; length tested untimed

FLAW	WHAT THE SIMULATION SHOWED	FIX
Item-level corrected estimator	False alarms 9.5% of the time	Simple person-level test: 5.3%

Two chances, reported separately. For a true drift of about 3.5 points, the test is significant in 80% of simulated studies but significant *and* with a mean of at least 3 points in only 48%, because the measured shift runs below the true one. For about 6 points the figures are 99.7% and 96%; for about 1.7 points, 36% and 12%.

One honest caveat. Power depends on how much D varies between people, and the simulated spread of about 8.5 points comes from invented personas. The table shows what other spreads would mean, for a mean measured D of 3 points.

SPREAD OF D_I (SD, POINTS)	6	8.5 (SIMULATED)	11	13
Power at 60 people	98%	86%	67%	55%
Power at 100 people	>99%	97%	86%	74%
People needed for 80%	27	52	85	118
Smallest drift detectable at 60 (80% power)	2.0	2.8	3.6	4.3

The sample-size rule, and a check that it is safe. After the first 20 analysable people, an analyst who receives only the standard deviation of D, never its mean, sets the final sample: the smallest number between 60 and 100 that gives 80% power for 3 points. The decision and its reason are logged on OSF before recruitment continues. If more than 100 would be needed, the study stops at 100 and reports the result as directional evidence at reduced power, with the calculation published for a follow-up. In simulation the full procedure kept the false-alarm rate at 4.9% (spread 8.5) and 4.8% (spread 11), against 5.0% for a fixed sample of 60, and raised the average sample to 82 when the spread was 11. Re-estimating a sample from variance alone is a standard internal-pilot design (Friede and Kieser, 2006).

Side bias, simulated. If pressure made each person favour one screen side by up to 25 points, randomising the attractor's side kept the false-alarm rate at 4.6% and the mean D at zero. Fixing the attractor on one side would have produced a false positive every time, which is why the randomisation is in the protocol.

Every simulated number in this paper maps to a script, a seed and an output file in the repository (Appendix B).

THAT MEANS

The design was broken in four places, and the simulation found all four before a single participant was paid.

SO WHAT

Publishing the simulation and its code before the data is what makes the eventual result, whichever way it falls, credible.

08

What to do about it

The first paper's answer was a weighting formula. The simulation of Study 2 changed that answer.

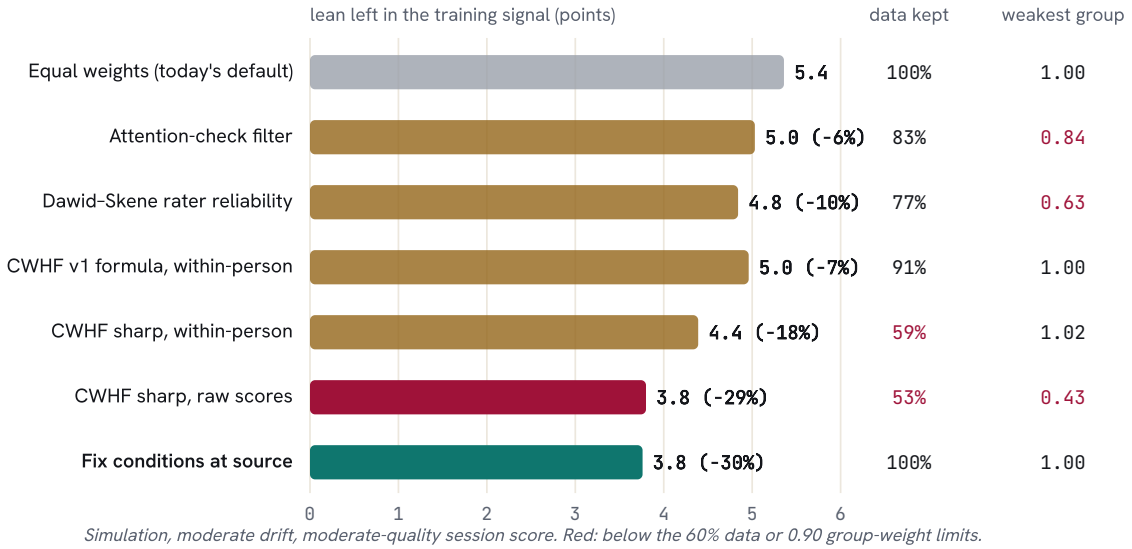


Figure 5. What each remedy removed. SIMULATED The first paper's formula removed about 7% of the lean. Raw-score weighting reached the target only by cutting night-shift workers' weight to under half. Shorter sessions, breaks and hourly pay removed about 30% with no data lost and no group penalised. All under the simulation's assumptions.

What the simulation assumed, and where the ranking reverses. The source fix is a bundle (capped sessions, enforced breaks, hourly pay) modelled as one assumption: it lowers average session depletion by 0.15 on the model's 0–1 scale. That number is a scenario, not a simulated mechanism or an estimate. Varying it shows what is robust. Lowering depletion by 0.05, 0.10, 0.15 or 0.20 removed about 11%, 22%, 32% or 44% of the lean (400 replications each, Monte Carlo error about 1 point), against about 8% for the first paper's formula and 19% for sharp within-person weighting in the same rerun. So the direction (prevention beats gentle weighting) held at every size tried; beating sharp weighting needed at least 0.10; the 30% figure itself is assumption-dependent. In a scenario where drift appears only in badly depleted sessions concentrated in night queues, sharp within-person weighting removed 55% of the lean, more than the source fix, but kept only 47% of the data. Study 2 tests which holds in real work.

CWHF now names a programme, not a formula. It runs in four steps, cheapest first:

STEP	WHAT IT DOES	DEPENDS ON
1 · Record	Store session length, position, time of day, pay model and a two-item check-in with every label (schema in Appendix C)	Nothing. Useful even if H1 fails
2 · Prevent	Cap sessions, enforce breaks, pay by the hour, keep hard items away from the end of long queues	H1, to justify the cost

STEP	WHAT IT DOES	DEPENDS ON
3 · Weight	Down-weight a person's own worst sessions, never comparing people with each other	H2 passing all its conditions
4 · Audit	Publish the lean, the data kept and every group's weight for each training set	Steps 1–3

Adopt now, before any result: record conditions with every label, state collection conditions in every dataset card, avoid per-item pay and long unbroken queues, and compare the style of answers that win product feedback early versus late in sessions before training on it.

What a lab gets in the first quarter. Honestly, nothing measurable until H1 and H2 report. What makes it worth starting anyway: step 1 is a handful of logging fields most platforms already hold, and a retrospective check of existing thumbs data against session position (H4) needs days, not a study. For high-risk systems, Article 10 of the EU AI Act requires data governance that includes examining possible biases; condition records and step-4 audits could support that documentation. That is a possible use, not a compliance claim.

What each step costs. Step 1 costs a few logging fields and a retention policy. Step 2 is the expensive one: capped sessions and hourly pay lower throughput per rater. Step 3 costs almost nothing to compute and is only worth doing if H2 passes. Step 4 costs analyst time per training set. A lab that adopts step 1 now and defers step 2 is making a defensible choice, provided it publishes the deferral; that is what step 4 is for.

Rules for step 1, as a protocol. (1) *Purpose:* data quality only; never pay, ranking, discipline or continued work. (2) *Separation:* condition records keyed to a salted pseudonym, stored apart from identity. (3) *Minimisation:* only the Appendix C fields; no biometrics, webcam or wearable data. (4) *Retention:* raw records deleted or aggregated within a stated period. (5) *Access:* a named few, every access logged. (6) *Reporting:* aggregate only, with small groups suppressed. (7) *Rights:* raters may skip the check-in without penalty, see their own records, contest them and ask for deletion. A lab that will not accept these terms should not adopt step 1: a state log without them is worse than none.

How H2 will be judged. Each pass condition has an operational definition, fixed with the registration (full text in the technical supplement).

CRITERION	SPECIFICATION
Lean reduced	|lean| against a reflective benchmark from at least 100 separate people, reduced by ≥ 25% relative to equal weights, with the 95% bootstrap interval (people and items resampled) excluding zero
Beats existing practice	Larger reduction than both an attention-check filter and Dawid–Skene reliability weighting, tuned as standard
Data kept	Effective sample size $(\Sigma w)^2 / \Sigma w^2$ ≥ 60% of labels
No group harmed	Every language, region and shift group's mean weight within 0.90–1.10 of the others', judged on its interval, not on a non-significant difference; no group's lean higher after weighting than before

CRITERION	SPECIFICATION
Prevention tested, not assumed	A source-fix arm (capped sessions, breaks, hourly pay, equal weights). If it reduces lean as much as the best weighting method, weighting is not recommended
Robust to gaming	Reported separately for raters whose check-in answers shift once weights exist; a large drop in benefit is reported as a failure of weighting

Two known weak points. If session scores ever influence weights, raters may learn to perform alertness and platforms may build shadow scores; scores are therefore never shown to managers or used per person, anchor items are harder to game than self-reports, and audits look for gaming. And comparing people only with themselves has a blind spot: someone who is always depleted looks normal against their own baseline. Prevention, not weighting, is the remedy for that.

THAT MEANS

Bad working conditions may be a data-quality problem, not only a welfare problem. In the simulation, the cheapest fix was also the fairest; Study 2 tests whether that holds in practice.

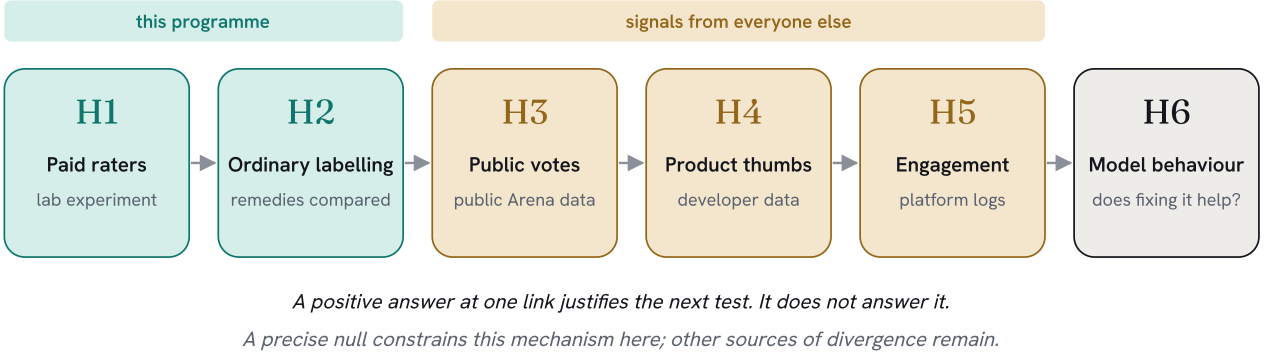
SO WHAT

Labs can start with step 1 tomorrow at almost no cost. If H2 fails, the recommendation simply becomes record, prevent and audit.

09

How far one result travels

H1 is the first link in a chain of separately testable questions. Each needs its own data.



A positive answer at one link justifies the next test. It does not answer it.

A precise null constrains this mechanism here; other sources of divergence remain.

Figure 6. One result, six tests. H1 and H2 are this programme. H3 can run now on public data. H4 to H6 need data held by AI developers and platforms, and are offered to them as open questions, not as part of the current studies. A positive H1 justifies testing the next link; it does not answer it. A null H1 clears nothing downstream.

Study 3: the same question in public votes

Chatbot Arena's public release records, for each vote, the two models, both answers, the vote, an anonymised user ID and a timestamp. Study 3 groups each user's votes into sittings and asks whether the chance of choosing the more confident, more agreeable or longer answer rises with position in the sitting, comparing each voter only with themselves. It needs no new participants, only research time; it can start now; and it is the one place length can be tested fairly.

Study 3 analysis plan in brief. *Sitting:* consecutive votes by one user ID with no gap over 30 minutes. *Styles, fixed in advance and scored blind to position:* difference between the two answers in hedging words (a published lexicon), in agreement with a view stated in the prompt (a classifier validated on 300 human-coded pairs) and in length. *Model:* logistic regression of choosing the higher-style answer on position in sitting, with user, model-pair and prompt-category fixed effects and standard errors clustered by user; one model per style; one-sided test of a positive slope, Holm-corrected across the three styles. Time of day is not used, because timestamps are not in voters' local time. *Order of work:* the analysis script is released with the registration, before the data are opened. *Effort:* a few weeks of one analyst's time.

It is observational. Later votes differ from earlier ones in more ways than fatigue: topics drift within a sitting, model pairs vary in difficulty, and voters learn. User fixed effects absorb who keeps voting; prompt-category effects absorb some topic drift; the rest is a stated limit. A positive slope is a reason for H4, not evidence of cause. A flat slope in a large sample would count against the wider claim.

What H6 would need. Comparable reward models or preference predictors trained on the same pairs under four regimes (original labels; labels with condition fields; labels from a source-fixed arm; condition-aware weighting), and evaluated on held-out reflective preferences, sensitivity to answer style, sycophancy evaluations, calibration and robustness across rater groups. A smaller lean in the data is not, by itself, better alignment; H6 is where that would be shown or not.

THAT MEANS

Real, if observational, evidence on this question can come from a public dataset before any lab study is funded.

SO WHAT

Developers who hold thumbs data with timestamps and session IDs can run H4 with the Study 3 plan. The author offers to analyse such data jointly and to publish the result whichever way it falls.

10

The strongest objections

Stated as a hostile reviewer would put them, with the answer this paper can give.

OBJECTION	ANSWER
"The reflective choice is just another biased state, such as boredom or over-thinking."	It may be. Policy evaluation always needs a stated reference (Bernheim and Rangel, 2009), and this one is chosen for testability. The test only asks whether the gap has a direction: if fast choices are the better ones, or neither is, a directional gap still means the data average to neither.
"You are extrapolating from 60 people to billions of users."	No. The general claim is established separately; the step to wild signals is stated as plausibility, and each channel needs its own test.
"A lab task is not real labelling work."	True. Study 1 tests whether the effect can exist; Study 2 tests whether it appears in ordinary sessions.
"Pressure might make people more accurate."	Then D is near zero or negative and H1 fails. The design allows that outcome.
"Three points is too small to matter, or was picked to suit the sample."	It is partly set by the design, as Section 3 says. It is fixed in advance, reported with the full interval, and anyone may apply their own threshold.
"This is just noise, which Kahneman and colleagues already described."	Noise is variance, and averaging reduces it. The question here is whether the error has a sign, which averaging does not remove.
"Rushed people pick whatever is on one side."	The attractor's side is randomised per item and person; in simulation, strong side bias produced no false positives beyond the 5% expected.
"Why not a field study of real shift boundaries?"	It would have more external validity and no causal identification. Study 1 trades realism for a within-person causal estimate; Studies 2 and 3 restore the field side.
"A rater's preferences may legitimately change."	A week apart they may. The control arm measures how far reflective choices move without pressure; D is read against that.
"OpenAI already found and fixed sycophancy."	That incident shows the pathway exists. It does not show why the signal leaned, or that other signals are free of it.
"Raters will game a state score."	Possibly, if scores ever affect weights. Scores are never used per person or shown to managers, anchor items are harder to game than self-reports, and prevention needs no score at all.
"Kopteva and Hlynianyi-Zhuk already proposed this."	They proposed rater state as a structured bias with an observational audit. This paper adds the experiment that can show whether pressure causes a directional shift. Complementary, and an invitation to collaborate.
"The simulation proves nothing."	Correct. It shows the design works under stated assumptions. The real study will show whether the assumptions hold.

Limits we acknowledge. Only two attractor styles are tested. The study runs in one language. The items are constructed, not real model outputs, and the 48 primary items are specified but not yet written; Study 1 can be no better than its item bank, so the full bank, reviewer notes and disputed items will be published with the data. Time pressure for twenty minutes is not fatigue at the end of a long shift. The manipulation check is self-reported. The author designed the analysis, so he will not analyse the confirmatory data unblinded, and the primary analysis script will be frozen before recruitment.

Pre-mortem: if this is wrong in three years, why?

- **The wrong attractor.** In real labelling the style that gains under pressure may be shorter, safer or more hedged answers, not confident or agreeable ones. H1 could fail while a lean exists.
- **The wrong pressure.** A timer and a memory load are not end-of-shift fatigue, money worries or loneliness. A lab effect may not generalise.
- **The wrong population.** Paid participants who know they are in a study may behave differently from annotators paid by throughput.
- **Lab is not field.** H1 can pass while H2 to H6 fail.
- **No problem to solve.** Existing rater-quality methods, or the averaging inside reward-model training, may already absorb the effect. H2 and H6 test this; the paper does not assume otherwise.

If any of these proves true, the result is published, logged in the changelog, and the affected claim is narrowed or withdrawn, not defended.

11

Ethics and equity

The risk is less in the study than in what is done with its results.

The study. Both experimental studies run only after independent ethics approval. The manipulation is mild and reversible: a timer and a number to remember. Participants know one session will be rushed, are paid the same regardless of speed or accuracy, and can stop at any time.

Equity. A state score can easily become a proxy for who someone is: shift workers, carers, older people, second-language readers. The simulation showed this happens even with a perfect state measure. Hence three standing rules: compare each person only with themselves; protect every group's weight and outcome; and never let a score affect anyone's pay or work.

Data protection. Condition records about identifiable raters are personal data under the GDPR and India's Digital Personal Data Protection Act 2023. Step 1 therefore needs a lawful basis,

purpose limitation, minimisation, stated retention and rights of access and deletion, and an adopter should say where the records are processed. This is a design requirement, not legal advice.

Dual use. A record of rater condition can be turned into a tool to screen, rank or drop workers, and nothing in the method prevents that. The conditions of use forbid it, but a condition of use is not enforcement. Adopters are asked to publish their recording policy, accept external audit of how records are used, and give raters the right to see and contest their own records.

Beyond annotation. For end users, CWHF applies only in aggregate. Inferring an individual's state from their behaviour, and using it to decide what they are shown, is outside CWHF and contrary to its conditions of use. No webcam, wearable or biometric inference, ever.

Reliability is not goodness. A calm, consistent rater can still endorse a harmful answer. CWHF sits alongside content safety review; it never replaces it.

For standards and policy readers. The programme proposes one auditable artefact: a conditions card attached to each training dataset, listing how labels were made (session lengths, pay model, check-in summaries), the measured lean, the data kept and every group's weight. It needs no disclosure of model internals and no new data from regulators; only that the conditions of human judgment be recorded where the label is made. It extends datasheets for datasets (Gebru et al., 2021) to the annotation layer. For high-risk systems, such a card could support the data-governance documentation of EU AI Act Article 10; that is a possible use, not a compliance claim.

SO WHAT

In simulation, the most effective remedy, better working conditions, is also the one that benefits the people producing the data. Whether that holds in practice is what Study 2 tests.

12

What changed since the first paper

The first paper, *The Coherence-Weighted Human Feedback (CWHF): Correcting the Foundational Flaw in AI Alignment* (Rao, March 2026, doi:10.5281/zenodo.19235970), proposed the idea. This paper keeps its question, narrows it, tests the design, and withdraws what the evidence does not support. It is a new paper, not a revision, and where the two disagree this one prevails.

FIRST PAPER CLAIMED	WHY WITHDRAWN	NOW
A flaw "established by five research programmes"	The programmes show state affects judgment, not that the lean in AI feedback has a direction	Stated as the open question H1

FIRST PAPER CLAIMED	WHY WITHDRAWN	NOW
A permanent bias ("Bias Theorem")	Arithmetic conditional on H1, not a theorem	$\hat{r} \approx r^* + (1 - p)\delta$, labelled conditional
Information-theoretically optimal under a 120-bit limit	The figure is outdated (current estimate about 10 bits/s; Zheng and Meister, 2024) and no derivation supports optimality	No optimality claim
About $50,000, one sprint, permanent gains	Unmeasured	Study costs estimated; benefits to be tested
A 700-person "coherent nucleus" via a 3.5% tipping point	Borrowed from an unrelated literature; untestable here	Withdrawn
Recursive state-capture of populations	Plausible loop, no evidence	Stated as a hypothesis only
Weighting by a coherence score as the fix	Simulation: about 7% of the lean removed	Record, prevent, weight, audit
The registered H1 design and $\pi \geq 0.60$ rule	Four design flaws (Section 7)	Revised H1–H3, to be registered
Coherence scores (c, Δc and related indices) in earlier companion material	Never validated	D and M; the scores are retired

What this paper stands behind

CLAIM	STATUS	BASIS
Professional judgments shift with state, in one direction	ESTABLISHED	Linder et al., 2014; Sievertsen et al., 2016
Raters of AI output favour agreeable, assertive or longer answers	ESTABLISHED	Sharma et al., 2023; Hosking et al., 2023; Singhal et al., 2023
User feedback has reached a reward model and moved a deployed model	ESTABLISHED	OpenAI, 2025
Public preference releases carry no rater-condition fields	ESTABLISHED	HH-RLHF, SHP, Chatbot Arena releases
The planned test detects a 6-point drift 99.7% of the time at 60 people, with 5.3% false alarms	SIMULATED	2,000 replications; open code
Fixing conditions removed more lean than the first paper's formula	SIMULATED	About 30% against 7%; direction robust to the assumed size, magnitude not
Pressure shifts choices toward the attractor	HYPOTHESIS	H1, untested
The reflective choice is the right reference	A stated convention	Section 5; objections in Section 10

CLAIM	STATUS	BASIS
A lean of this size matters at training scale	HYPOTHESIS	H6; not claimed, only unmeasured

Changes in version 1.1

After a second adversarial review round (13 reviewers): estimand and switch identity for D and M; alternative-explanations table; worked example; committed result sentences; power by spread and a simulation check of the sample-size rule; side-bias simulation; sensitivity of the source-fix result to its assumed size (and correction of its description from "0.15 standard deviations" to "0.15 on a 0–1 depletion scale"); H2 criteria table; Study 3 plan; related-work table; pre-mortem; data-protection, dual-use and policy notes; conditions schema (Appendix C); reproducibility map (Appendix B); cost figures made consistent. No hypothesis, threshold or decision rule was loosened.

THAT MEANS

The idea survived its own audit. Most of the scaffolding around it did not, and what remains is listed above with its status.

13

What happens next

STAGE	WHAT	STATUS
0	This paper, the companion lab and the simulation code, published openly	Now
1	Register the revised H1, H2 and H3 on OSF, with the Study 3 script	Next
1b	Study 3 on public Chatbot Arena data	No recruitment; a few weeks of one analyst
2	Write, review and pilot the 48-item question bank (the main bottleneck before Study 1; validation in Appendix A)	Next
3	Ethics approval through a university partner	Needs a collaborator
4	Study 1: about 72 recruited for 60 analysable, plus a 12-person control arm	$3,000–11,000
5	Study 2, if Study 1 is confirmed	Reuses Study 1
6	H4–H6 with developers and platforms that hold the data	By invitation

What each result would lead to

RESULT	NEXT ACTION
H1 supported or exceeds the threshold	Independent replication; Study 2

RESULT	NEXT ACTION
H1 directional but small	Report; refine the utility case before any remedy
H1 precise null (upper bound below 3)	Publish; narrow the programme to record and audit; no weighting
H1 null with a wide interval	Inconclusive; improve precision or redesign
H1 reverse	Investigate what pressure protects against; the attractor model is withdrawn for this task
H1 holds only for some items or groups	Replicate; study moderators before generalising
H1 holds, H2 fails	Do not promote weighting; recommend record, prevent and audit
H2 passes, H6 finds no model-level change	Reassess the downstream mechanism; data quality alone is not alignment

Who runs it, and when it stops. An independent researcher has no laboratory, and one advantage that matters here: a negative result threatens no grant, team or product. The confirmatory studies will run under an independent principal investigator at a partner institution, with no co-authorship history with the author; no participant data will be collected until that person and an ethics approval are in place. If none of stages 1–4 finds support or a partner within twelve months, the programme pauses and says so in public.

How to take part

- **Take the flip test** in the companion lab at twopointsingularity.com/lab. It is a demonstration, not research participation, and says nothing about you.
- **Rerun the simulation.** Code, seeds and outputs are open at github.com/venramak/twopointsingularity-cwhf. Change the assumptions and see whether the conclusions hold.
- **Review the items.** Anyone who finds a question with two defensible answers improves the study directly.
- **Collaborate.** University researchers who can provide ethics review or serve as principal investigator, and teams who could record step 1 in their pipelines or run H4, are invited to write to venkat@twopointsingularity.com.
- **Report a flaw.** Open an issue in the repository. Every accepted correction is logged in public, with its date.

The question is small enough to test and large enough to matter. The next step is data.

Declarations

Funding. None.

Conflicts of interest. The author publishes books on related themes under Two Point Singularity and has a commercial interest in that wider work. He will not serve as principal investigator of the confirmatory studies.

Use of AI assistance. An AI assistant (Claude, Anthropic) was used for literature retrieval, simulation code, analysis scripts and drafting, and to simulate hostile review panels. The author set the question, the hypotheses, the thresholds and the decision rules, chose which review points to adopt, and made every normative and interpretive judgment. Citations were checked against their original sources; one reference found to be misattributed in an earlier draft was corrected. The simulation code was checked against the written protocol, and its outputs regenerate every simulated number in the paper. The author takes full responsibility for all claims and errors.

Data and code availability. No human participant data were collected for this paper. The simulation code (Apache-2.0) and its seeds, outputs, the full protocol and the pre-registration text (CC BY 4.0) are at github.com/venramak/twopointsingularity-cwhf, simulation code as of commit `5339b22`. Participant-level data and analysis scripts from the empirical studies will be deposited with a DOI on completion. The companion lab sets no cookies and uses no browser storage or analytics; its only outside request is for web fonts.

Ethics. No ethics application has been submitted yet. No participant data will be collected before approval by an independent research ethics committee at a partner institution.

Registration. This paper follows and supersedes the claims of Rao (2026a). The earlier H1 registration (OSF W7NYH) remains on record. The revised H1–H3 are not yet registered; they will be registered before any data are collected or opened, and this paper will be updated with the link.

Suggested citation. Rao, V. (2026). *Signal ≠ Value: Does Pressure Bend the Human Feedback That Trains AI?* Two Point Singularity. Preprint. CC BY 4.0.

References

Bernheim, B. D., & Rangel, A. (2009). Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics*, 124(1), 51–104.

Conlin, M., O'Donoghue, T., & Vogelsang, T. J. (2007). Projection bias in catalog orders. *American Economic Review*, 97(4), 1217–1249.

Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society C*, 28(1), 20–28.

Ethayarajh, K., Choi, Y., & Swayamdipta, S. (2022). Stanford Human Preferences (SHP) dataset. huggingface.co/datasets/stanfordnlp/SHP.

Friede, T., & Kieser, M. (2006). Sample size recalculation in internal pilot study designs: a review. *Biometrical Journal*, 48(4), 537–555.

Gao, L., Schulman, J., & Hilton, J. (2023). Scaling laws for reward model overoptimization. *ICML 2023*. arXiv:2210.10760.

Gebru, T., Morgenstern, J., Vecchione, B., et al. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., et al. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4), 546–573.

Hosking, T., Blunsom, P., & Bartolo, M. (2023). Human feedback is not gold standard. arXiv:2309.16349.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A Flaw in Human Judgment*. Little, Brown Spark.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2022). The challenge of understanding what users want: inconsistent preferences and engagement optimization. arXiv:2202.11776.

Kopteva, E., & Hlynianyi-Zhuk, V. (2026). Rater state bias in RLHF preference data: an audit framework. arXiv:2607.16195.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: a tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269.

Linder, J. A., Doctor, J. N., Friedberg, M. W., et al. (2014). Time of day and the decision to prescribe antibiotics. *JAMA Internal Medicine*, 174(12), 2029–2031.

LMSYS (2023). Chatbot Arena conversation dataset release. lmsys.org/blog/2023-07-20-dataset.

Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., & Dragan, A. D. (2023). Engagement, user satisfaction, and the amplification of divisive content on social media. arXiv:2305.16941.

OpenAI (2025, 2 May). Expanding on what we missed with sycophancy. openai.com/index/expanding-on-sycophancy.

Orne, M. T. (1962). On the social psychology of the psychological experiment. *American Psychologist*, 17(11), 776–783.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker*. Cambridge University Press.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: your language model is secretly a reward model. *NeurIPS 2023*. arXiv:2305.18290.

Raykar, V. C., Yu, S., Zhao, L. H., et al. (2010). Learning from crowds. *Journal of Machine Learning Research*, 11, 1297–1322.

Rao, V. (2026a). The Coherence-Weighted Human Feedback (CWHF): correcting the foundational flaw in AI alignment. Zenodo. doi:10.5281/zenodo.19235970.

Shafir, E., Simonson, I., & Tversky, A. (1993). Reason-based choice. *Cognition*, 49(1–2), 11–36.

Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards understanding sycophancy in language models. arXiv:2310.13548.

Sievertsen, H. H., Gino, F., & Piovesan, M. (2016). Cognitive fatigue influences students' performance on standardized tests. *PNAS*, 113(10), 2621–2624.

Singhal, P., Goyal, T., Xu, J., & Durrett, G. (2023). A long way to go: investigating length correlations in RLHF. arXiv:2310.03716.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast, but is it good? Evaluating non-expert annotations for natural language tasks. *EMNLP 2008*, 254–263.

Svenson, O., & Maule, A. J. (Eds.) (1993). *Time Pressure and Stress in Human Judgment and Decision Making*. Plenum Press.

Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 10: Data and data governance. *Official Journal of the European Union*, L, 12 July 2024.

Digital Personal Data Protection Act, 2023 (India), No. 22 of 2023.

Utz, C., Degeling, M., Fahl, S., Schaub, F., & Holz, T. (2019). (Un)informed consent: studying GDPR consent notices in the field. arXiv:1909.02638.

Vohs, K. D., Schmeichel, B. J., Lohmann, S., et al. (2021). A multisite preregistered paradigmatic test of the ego-depletion effect. *Psychological Science*, 32(10), 1566–1581.

Welinder, P., Branson, S., Belongie, S., & Perona, P. (2010). The multidimensional wisdom of crowds. *NeurIPS 2010*.

Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., & Movellan, J. (2009). Whose vote should count more: optimal integration of labels from labelers of unknown expertise. *NeurIPS 2009*.

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, 60(2), 181–192.

Zheng, J., & Meister, M. (2024). The unbearable slowness of being. arXiv:2408.10234.

APPENDIX A

H1 analysis plan in brief

The registration text in the technical supplement governs; this summary is for reviewers.

Estimand	$\Delta = E[D_i]$, $D_i = P_i - R_i$ over primary items answered within time in both sessions (Section 5)
Analysis population	Everyone with ≥ 10 such items, regardless of the manipulation check, which gates interpretation, not inclusion. Exclusions: two or more failed attention checks; a reflective session under a third of the median time
Primary test	One-sample t-test of mean $D > 0$, one-sided, $\alpha = 0.05$; one-sided because the remedy question only arises if the lean has the predicted direction. A reverse effect is tested two-sided and reported as exploratory
Fallback	If a Shapiro–Wilk test on the D_i rejects normality at $\alpha = 0.01$, a Wilcoxon signed-rank test with the same decision table
Reported in all cases	Mean D, SD, 95% interval, exact p, d, share of people with $D > 0$, mean M, and the same for the control arm
Sample size	60–100 by the blinded variance rule (Section 7), decided at 20 people and logged before recruitment continues
Timeouts	Excluded from D; sensitivity analyses count a timeout as each option in turn and exclude people with more than 30% timeouts
Multiplicity	Gatekeeping: style-level tests (confident, agreeable) are interpreted only if the overall test succeeds. Interaction, π, accuracy change, item-level and subgroup results are exploratory and enter no decision. H2 runs only if H1 is supported; H3 is observational with its own Holm correction
Robustness	Linear probability model of attractor choice on condition with person and item cluster-robust errors; order × condition; D on no-reason items; D where the attractor is wrong
Suspicion	People who name the attractor hypothesis unprompted are flagged, kept in the primary analysis and excluded in a sensitivity analysis; any difference above 1 point is reported
Exclusion sensitivity	The primary result is reported with and without each exclusion class. An exclusion class that changes the verdict is reported as such in the results abstract
Control arm	Twelve people, two reflective sessions ≥ 7 days apart. Its mean D (expected near zero) and SD are reported before the primary result. If its 95% interval excludes zero, the primary D is also reported net of the control mean
Item bank	48 primary items, 24 per style, attractor correct in exactly half of each, answers length-matched within 15%. Each item reviewed by two people blind to the attractor and correctness flags: one defensible correct answer? style identifiable? lengths matched? Accepted only if both agree on all three; agreement and every rejected item are published with a versioned bank
Blinding and code	The primary analysis script is frozen and released before recruitment; the author does not analyse confirmatory data unblinded

APPENDIX B

Simulation: model and reproducibility map

H1 model. For person *i* and item *j*, logit $P_R(\text{attractor}) = b_i + e_j + a_i \cdot clar_j \cdot C_j$, and logit $P_P(\text{attractor}) = b_i + e_j + (1 - att_i) \cdot a_i \cdot clar_j \cdot C_j + pull_i$, with a lapse rate mixing in random choice under pressure. C_j = +1 where the attractor is correct, −1 otherwise. Skill a, lapse, attenuation att, susceptibility to pull, reading speed and style preference b vary by persona; all persona values are invented and listed in `h1sim.py` and the technical supplement.

H2 model. Raters label 600 items across sessions; each session has a depletion z in [0, 1]. Depletion raises lapses, lowers discrimination and, when a pull exists, moves choices toward the attractor. A session score combines a noisy two-item self-report with response-time variability. Lean is measured against the reflective population rate of each item. The source fix lowers mean depletion by 0.15 (sensitivity 0.05–0.20).

NUMBER IN THE PAPER	SCRIPT	REPLICATIONS, SEED	OUTPUT
5.3% false alarms; 99.7%, 80%, 36% detection; 96%, 48%, 12% significant and ≥ 3	`persontest.py`	2,000 (null), 600 (effects) at N = 60; seed N + reps	`h1_person_test.json`
Figure 4 curves; four design flaws	`run_h1_grid.py`, `h1sim.py`	400–1,000 per setting; seeds by setting	`h1_results.json`
Figure 5; 7%, 18%, 29%, 30%; 55% and 47% in the threshold scenario	`h2sim.py 80`	80 per scenario; seed 1000 × scenario index + 5	`h2_results.json`
Power by spread; sample-size rule (4.9%, 4.8%); side bias (4.6%); source-fix sensitivity (11–44%)	`sensitivity.py`	5,000–20,000 (rule), 4,000 (side bias), 400 (H2); seed 20261112	`sensitivity_results.json`

Python 3 with NumPy and SciPy. Reruns with the stated seeds reproduce the outputs exactly. The 400-replication H2 rerun in `sensitivity.py` agrees with the 80-replication headline figures within Monte Carlo error (gentle formula 8% against 7%; source fix 32% against 30%).

APPENDIX C

Conditions record: a minimum schema for step 1

One record per label, joined to the label by its ID. It follows the lineage of datasheets for datasets (Gebru et al., 2021) and is useful however H1 turns out.

FIELD	MEANING	REQUIRED
`label_id`	Joins to the label	Yes
`rater_pseudonym`	Salted hash, stored apart from identity	Yes
`session_id`, `position_in_session`	Which session, and the label's order within it	Yes
`session_elapsed_min`	Minutes since the session began	Yes
`time_on_item_s`	Seconds spent on this item	Yes
`local_hour_bucket`	Rater's local time in 3-hour bands	Yes
`pay_model`	`hourly`, `per_item` or `other`	Yes
`checkin_alert_1to5`, `checkin_rushed_1to5`	Two taps at session start; may be skipped without penalty	Optional
`schema_version`, `consent_flags`	Version of this schema; what the rater agreed to	Yes

Do not record names, contact details, demographics beyond what the study design needs, health, biometrics, webcam or wearable data, or anything inferred about a person's state from their behaviour. **Governance** follows the step 1 protocol in Section 8: purpose limited to data quality, aggregate reporting with small groups suppressed, stated retention, logged access, and rater rights to see, contest and delete. A machine-readable version is in the repository.

APPENDIX D

Notation

SYMBOL	MEANING	SECTION
P_i, R_i	Share of attractor choices under pressure, on reflection, for person i	5
u_i, w_i	Items switched toward, away from the attractor	5
D_i, D	Reflective-pressured choice gap for person i; its sample mean	5
M_i	Mismatch: share of items chosen differently in the two sessions	5
Δ	Population mean of D_i: the estimand	5
SESOI	Smallest effect size of interest: 3 percentage points	3, 6

SYMBOL	MEANING	SECTION
$\hat{r}$, r^*, p, δ	Learned reward; reward from reflective labels; share of reflective sessions; average lean per pressured session	3
ESS	Effective sample size, $(\Sigma w)^2 / \Sigma w^2$, as a share of labels	8
π	Share of changed answers that moved toward the attractor (descriptive only; its 0.60 rule is retired)	7, 12