Signal ≠ Value: Technical Supplement

Protocol, simulation and pre-registration text for *Signal ≠ Value: Does Pressure Bend the Human Feedback That Trains AI?* (Rao, 2026). This supplement gives the full detail behind the paper. Where the two differ, the paper prevails.

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

Supplement version 1.1 · 12 November 2026 · Accompanies paper version 1.1 · Preprint, not peer reviewed · Supersedes the claims of the March 2026 paper (doi:10.5281/zenodo.19235970) · Licence: CC BY 4.0 · Published by Two Point Singularity

Companion: *Signal ≠ Value Lab v2.0* at cwhf.vercel.app and twopointsingularity.com/research/cwhf: the evidence base, a flip-test demonstration and the full design simulation. Code and simulation: github.com/venramak/twopointsingularity-cwhf.

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

Status of claims

Every claim in this paper carries one of five labels, and no lower-tier claim is used to support a higher one.

Label	Meaning	Examples in this paper
Established	Supported by published evidence or documented practice	Judgments vary with state; raters favour agreeable, assertive and longer answers; published preference datasets record no rater state; end-user feedback has entered reward signals
Hypothesis	Precise, testable, not yet tested	H1: pressure shifts choices toward a pre-named style relative to the same person's reflective choices. H2: a session signal can reduce that lean in training data. H3 (exploratory): the same lean appears in public votes as a sitting wears on

Label	Meaning	Examples in this paper
Simulated	A property of the study design, shown on simulated participants	Power, false-positive rate and minimum sample size for H1; what weighting can and cannot remove in H2
Proposed	A design that depends on the hypotheses passing	CWHF v2: record, prevent, weight, audit
Retired	Claimed in v1.0 and withdrawn	The "Bias Theorem", the 120 bits/s figure, rate-distortion optimality, the ICC-based H1 (Section 14)

Abstract

Systems that learn from human feedback treat each rating as a reading of what people prefer. A rating is better understood as a measurement taken under conditions: the rater's time pressure, attention and fatigue are part of it. Published work shows that human raters already favour agreeable, assertive and longer answers. No study has tested whether that lean grows when the same person is under pressure. If it does, collecting more feedback will not average it away.

This paper states that question as a single hypothesis, defines "value" operationally as the same person's reflective choice, and specifies two experimental studies and one observational study. Study 1 (H1) manipulates time pressure and divided attention within person and measures D, the change in how often a pre-named "attractor" style is chosen. Study 2 (H2) tests whether a cheap session signal can reduce the lean of a training set beyond existing rater-quality methods, without penalising any group.

The question is not confined to annotation. Every recorded digital action — a thumbs-up, a vote, an upvote, a click, an "accept" — is a signal produced under conditions, and several such signals demonstrably reach AI training. Study 1 tests the most favourable case: paid, attentive adults with a correct answer available. Section 2.1 sets out what a result would and would not imply for every other signal, and Study 3 proposes a no-cost observational test on public, in-the-wild preference votes.

Before collecting data, both studies were run thousands of times on simulated participants drawn from eight personas. The simulation found four flaws in the earlier design and fixed them. The final H1 test holds a 5.3% false-positive rate and detects a 5–6 point drift with at least 99% probability at 60 participants. The H2 simulation changed the proposal: under the stated assumptions, the weighting formula in v1.0 removes under 10% of the lean, while improving working conditions at source removes about 30% with no loss

of data. CWHF v2 therefore records conditions first, prevents drift at source second, weights only where it demonstrably helps, and audits every step. All simulation code is released with this paper; no result in it is evidence about real people.

1. Summary

The question. When people judge under pressure, do their choices drift away from what they themselves choose on reflection, in a consistent direction rather than at random?

Why it matters. Random error cancels as more feedback is collected. Directional error does not: more data teaches a model the lean more firmly. If pressured raters lean toward confident-sounding or agreeable answers, the reward models trained on their labels inherit that lean, and the models then produce more of it.

How far it reaches. AI systems learn not only from paid raters but from end users' thumbs, votes, upvotes and engagement. OpenAI reported in 2025 that a reward signal built from users' thumbs-up and thumbs-down data helped push a model toward agreeable, sycophantic answers. Study 1 tests the cleanest case. A positive result would remove the default assumption that pressure adds only noise to every other human signal; a null result would not clear those signals, because state is only one of eight reasons a signal departs from a value.

What is already known. Human judgments vary with fatigue, load and time of day, including in professional decisions such as antibiotic prescribing (Linder et al., 2014). Human raters of AI output favour agreeable (Sharma et al., 2023), assertive (Hosking et al., 2023) and longer (Singhal et al., 2023) answers. Published preference datasets do not record the rater's state.

What is not known. Whether the lean grows with the rater's state *within the same person*. Kopteva and Hlynianyi-Zhuk (2026) propose rater state as a structured bias and set out an audit framework; we have not found a study that tests this within-person contrast directly.

The test. Study 1 compares each person's pressured and reflective choices on 48 length-matched question pairs. In each pair one answer carries a pre-named attractor style (more confident, or more agreeable), and the attractor is the correct answer in exactly half the pairs. The primary measure is D, the change in attractor choice rate. H1 passes if the mean D across about 60 people is reliably above zero and at least 3 percentage points.

What the simulation shows. Run on simulated participants, the final design detects a 5–6 point drift with at least 99% probability, detects a 3.5-point drift about 80% of the time, and raises a false alarm 5.3% of the time when there is no drift. It also showed that the v1.0 decision rule ($\pi \geqslant 0.60$) would have missed a real 6-point drift four times in five.

The proposal, revised. CWHF v2 is a sequence: record the conditions of every label; prevent drift at source with session caps, breaks and hourly pay; weight within-person only where H2 shows a gain; audit the lean, the data kept and every group's weight. In simulation, prevention outperformed the v1.0 weighting formula.

What would show this is wrong. H1 is not supported if mean D is not reliably above zero; if its confidence interval lies below 3 points, any effect is too small to matter by this standard. H2 fails if no session signal removes at least a quarter of the lean beyond existing methods while meeting prespecified criteria for every group. The proposal fails if improving conditions alone performs as well as weighting.

Cost. Study 1 runs on a participant platform for an estimated $3,000–7,000 including ethics review, rising to about $11,000 if the pre-registered sample-size rule calls for 100 participants, over 8–10 weeks. Study 2 can reuse Study 1's infrastructure.

2. The problem: human feedback is a measurement taken under conditions

Every experimental science records the conditions under which a measurement was taken, because the instrument's state affects the reading. Human feedback to AI systems is treated differently. A preference label is stored as if it were a fact about the world ("answer A is better") rather than an observation made by a particular person, at a particular time, in a particular condition.

The broader gap between what people signal and what they would endorse is long established. Behavioural welfare economics stopped treating every choice as a revealed preference decades ago (Bernheim & Rangel, 2009). Engagement data diverges from users' stated preferences (Milli et al., 2023), and optimising for engagement under inconsistent preferences can leave users worse off (Kleinberg et al., 2022). The companion lab catalogues 71 cases and tags each with one of eight causes: state, choice architecture, incentives, information gaps, time inconsistency, social pressure, preference construction, and proxy or measurement problems.

This paper concerns only the first cause. It matters for AI in a specific way:

1. **Scale does not fix it if it has a direction.** If state adds random error, collecting more labels cancels it. If state pushes labels consistently toward one style, more labels entrench that style.
2. **It can reinforce itself.** A model trained on a lean produces answers with that lean. If raters under pressure find such answers more acceptable, the next round of labels leans further. This loop is a hypothesis, not a demonstrated effect; it is stated to show why the direction question matters.

3. **It is invisible in current data.** Because rater state is not recorded, the lean cannot be estimated or corrected after the fact.

The empirical scope is deliberately narrow; the conceptual scope is not (Section 2.1). Choice architecture, incentives and the other seven causes are often larger in everyday behaviour; cookie-consent banners, for example, are shaped mainly by interface design (Utz et al., 2019). State is singled out because it is the cause that varies within the same person, can be manipulated ethically, and is cheap to record.

2.1 Every digital action is a signal

A preference label from a paid rater is one kind of signal. Most human signals that reach AI systems are not produced in annotation pipelines at all. The table lists the main channels, with what is documented about each.

Channel	Examples	How it reaches AI systems	Status
Explicit product feedback	Thumbs up or down, "which response do you prefer?"	An April 2025 GPT-4o update "introduced an additional reward signal based on user feedback—thumbs-up and thumbs-down data from ChatGPT"; OpenAI concluded this weakened the primary reward signal and that "user feedback in particular can sometimes favor more agreeable responses" (OpenAI, 2025)	Established, by the developer's own account
Public crowd votes	Chatbot Arena pairwise votes	Leaderboards that developers optimise against; released as preference data, with an anonymised user ID and timestamp per vote (LMSYS, 2023)	Established

Channel	Examples	How it reaches AI systems	Status
Community scores	Reddit karma and upvotes	GPT-2's WebText kept only outbound links from Reddit posts with at least 3 karma, as a quality filter for pre-training text (Radford et al., 2019); the Stanford Human Preferences dataset turns upvotes into pairwise preferences (Ethayarajh et al., 2022)	Established
Engagement	Clicks, dwell time, watch time, scroll	Training targets for recommendation and ranking systems, which increasingly serve AI-generated content (Milli et al., 2023)	Established
Consent and defaults	Cookie banners, "accept all", terms	Decide which data may be collected at all; shaped mainly by interface design (Utz et al., 2019)	Established that the signal departs from preference
Interaction traces	Accepting a code suggestion, regenerate, copy, edit	Product metrics and, at some developers, training or evaluation signals	Practice varies and is often undisclosed

Four conclusions follow, each stated at the strength the evidence allows.

1. **The general claim is established.** Every recorded action is a signal produced under conditions, not a direct readout of what the person values. It applies to every channel above. This is the umbrella under which the lab's 71 cases sit.
2. **State is one cause among eight, and in the wild it is rarely the largest.** Defaults, interface design and incentives dominate several of these channels. A claim that "state" explains the gap everywhere would overreach, and this paper does not make it.

3. **The a fortiori argument.** Study 1 tests the most favourable conditions: paid, instructed, attentive adults choosing between two short answers, one of them correct. Most signals in the wild are produced under worse conditions: on phones, while doing something else, late at night, with no accountability for accuracy, and often through interfaces designed to obtain a click. If a directional lean appears under the favourable conditions, the default assumption that it is absent under worse ones loses its basis. This is an argument from plausibility, not a proof. Wild signals also differ in who chooses to give them, and the attractor styles may differ.
4. **The asymmetry.** A positive H1 widens the case for recording and auditing the conditions of every human signal used in training. A null H1 does not clear wild signals, because the other seven causes remain and state may act differently outside the laboratory. Neither result licenses a claim about any specific product without testing that product's data.

2.2 How one result cascades

H1 is the first link in a chain of separately testable questions. Each needs its own data, and a positive answer at one link justifies the next test; it does not answer it.

Link	Question	Data needed	Who can run it
H1	Does pressure shift paid raters' choices toward a pre-named style, within person?	Study 1 (Section 7)	This programme, about $3,000–7,000
H2	Can recording conditions, preventing drift or weighting reduce that lean in ordinary labelling?	Study 2 (Section 8)	This programme, reusing Study 1
H3	Do in-the-wild preference votes show the same lean as sessions wear on?	Public Arena votes (Study 3, Section 8)	Anyone; no recruitment
H4	Does the lean appear in end-user product feedback such as thumbs data?	Session metadata held by developers	AI developers
H5	Does it appear in engagement signals used by recommender systems?	Platform logs	Platforms, or researchers with data access
H6	Does reducing the lean change model behaviour, for example on sycophancy evaluations?	A reward model trained with and without correction	AI labs or academic groups

2.3 What a 3-point lean means

In pairwise data, a lean of 3 percentage points (53 rather than 50) corresponds to a learned reward difference of about 0.12 on the logit scale under the Bradley–Terry model used by most reward models, applied to every item where the style appears. In a dataset of one million comparisons it is about 30,000 more choices of the attractor than reflective judgment would give, all in one direction. Optimising a policy against a learned reward is known to amplify systematic errors in that reward (Gao et al., 2023). Whether a lean of this size changes model behaviour enough to matter is untested; it is link H6.

3. Definitions and the normative choice

Each term below is used in one sense only, here and in the companion lab.

Term	Definition
Signal	A recorded human output: a choice, rating, click or label

Term	Definition
Value (v)	The choice the same person makes on the same item with no time limit and no concurrent task, after time to reflect. An operational convention, not a claim about a "true self"
State	The person's condition when producing the signal: time pressure, divided attention, fatigue, time of day. Manipulated in Study 1, measured in Study 2
Attractor	A pre-named answer style hypothesised to gain under pressure. In Study 1: more confident (unhedged), or more agreeable with the asker
D	Within-person change in attractor choice rate, pressured minus reflective, over items where the attractor is correct exactly half the time
π	Share of changed answers that moved toward the attractor. Descriptive only
c	Session score from a two-item check-in and response-time variability, centred on the person's own average. Unvalidated
Lean	Mean deviation of a set of labels from the reflective population's choice probabilities, toward the attractor, in percentage points
ESS	Effective sample size after weighting, as a share of all labels: $(\Sigma w)^2 \div (n \cdot \Sigma w^2)$

The normative choice, stated as a choice

Using the reflective choice as the reference is a convention adopted so the claim can be tested. It follows a long tradition of giving weight to considered over momentary preferences, but it can be wrong. Reflection sometimes degrades judgment: listing reasons reduced the quality of taste judgments in Wilson and Schooler (1991), and remembered experience diverges from moment-by-moment experience (Redelmeier & Kahneman, 1996).

The hypothesis therefore asks a narrower question than "which self is right". It asks only whether the gap between the two judgments has a direction. A directional gap matters for training data whichever self one prefers, because it means the data do not average to either.

Reliability is not goodness

CWHF concerns the reliability of a signal, not its moral content. A calm, consistent rater can endorse a harmful answer. Any use of CWHF must therefore sit alongside content-level safety review, never replace it.

Why "coherence"

In this paper, coherence has one operational meaning: how closely a person's choices in a given session agree with that same person's reflective choices on the same kind of item. D is a coherence gap measured under controlled pressure. A coherence-weighted dataset is one in which sessions that agree less with their own rater's reflective choices count for less. The word carries no further claim.

4. What is established

Five claims are supported by published evidence or documented practice. They motivate the hypothesis; none of them is the hypothesis.

#	Claim	Evidence	What it does not show
E1	Professional judgments shift with state, in one direction, within the same people	Antibiotic prescribing for respiratory infections rose across clinic sessions: adjusted odds 1.01 in the second hour to 1.26 in the fourth, across 21,867 visits and 204 clinicians (Linder et al., 2014). Test scores fall later in the school 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 date raised the chance a cold-weather item was returned by about 3.95 percentage points (Conlin et al., 2007)	That the reversal reflects reflection rather than new information

#	Claim	Evidence	What it does not show
E3	Human raters of AI output favour particular styles	Raters prefer convincing sycophantic answers over correct ones a non-negligible fraction of the time (Sharma et al., 2023); assertiveness lowers detection of factual errors (Hosking et al., 2023); much of RLHF reward gain tracks length (Singhal et al., 2023)	That these preferences grow with rater state
E4	Published preference datasets carry no rater-state fields and are used with equal weights	Public releases such as Anthropic HH-RLHF, Stanford SHP and Chatbot Arena include no rater-state field. Internal practice at AI developers is not public and may differ	That recording state would change outcomes
E5	End-user feedback enters reward signals and can move models toward agreeable answers	OpenAI reported that a reward signal from ChatGPT thumbs-up and thumbs-down data contributed to the sycophantic GPT-4o update of April 2025, and that user feedback "can sometimes favor more agreeable responses" (OpenAI, 2025)	That the users' state played any part

E5 is the strongest real-world evidence that the pathway this paper worries about exists: a human signal with a lean toward agreeableness entered a reward model and changed a deployed model's behaviour. It says nothing about why users leaned. That is the gap H1 begins to fill.

One mismatch is stated openly. Length is the best-evidenced style bias (E3), but it cannot be tested fairly under a timer: long answers go unread and the measured drift reverses sign (Section 9). Study 1 therefore tests confident and agreeable styles only, its conclusions do not extend to length, and length is tested without a timer in Study 3.

One nearby literature warns against over-confidence. "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). This paper therefore avoids that construct and relies on direct, reversible manipulations of time pressure and divided attention, with a manipulation check.

The often-quoted contrast between a narrow conscious channel and rich sensory input remains useful as motivation, with current figures: human behavioural throughput of about 10 bits per second against sensory input of about 10^9 bits per second (Zheng & Meister). It establishes that people must compress. It does not establish that compression under pressure has a direction.

5. The hypothesis and its model

H1

> 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.

H1 is about the signal, not about one mechanism. Three mechanisms could produce a positive D, and the design can partly separate them:

1. **Added pull.** Pressure adds a direct preference for the attractor style.
2. **Heuristic reliance.** Pressure weakens the evaluation of content, so a pre-existing mild preference for the style dominates more often. The simulation showed this route matters: when people already lean slightly toward a style on reflection, pressure amplifies it.
3. **Lapses.** Pressure causes some near-random choices. These move choice rates toward 50%, not toward the attractor, so they add noise but no lean when the attractor is correct exactly half the time.

Mechanisms 1 and 2 both produce a lean that averaging cannot remove, and both matter for training data. Secondary analyses distinguish them using the interaction between pressure and whether the attractor was correct.

The training-data model

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

```
\hat{r} \approx r^{*} + (1-p)\,\delta
```

Here r* is what it would learn from reflective labels only. If δ averages to zero, the error shrinks as data grow. If δ is positive, the error persists however much data is added. This is arithmetic conditional on H1, not a theorem about the world. Version 1.0 called it the "Bias Theorem"; that name is retired.

H2

> In ordinary labelling sessions without manipulation, weighting labels by a within-person session score reduces the lean of the resulting training set by at least 25%, more than the best existing rater-quality method, while keeping at least 60% of effective sample size, keeping every language, region and shift group's mean weight within 0.90–1.10 of others', and worsening no group's lean.

H2 is tested against a second intervention as well: improving working conditions at source. If that does as well, the case for weighting falls away.

H3 (exploratory, observational)

> In public in-the-wild preference votes, the same voter's probability of choosing the more confident, more agreeable or longer response rises with the number of votes that voter has already cast in the same sitting.

H3 cannot establish cause: later votes differ from earlier ones in more ways than fatigue. Its value is that it is free, uses data that already exist, can test length without a timer, and could falsify the wider claim early if no such trend appears in large samples.

6. Related work and what is new

CWHF sits where four literatures meet. The table states what each already covers and what this paper adds.

Area	What it covers	What it leaves open	This paper
Rater style bias in RLHF (Sharma 2023; Hosking 2023; Singhal 2023)	Raters and reward models favour agreeable, assertive and long answers	Whether the bias varies with the rater's state	Tests the state dependence directly, within person
Rater state as bias (Kopteva & Hlynianyi-Zhuk 2026)	Proposes annotator state as a structured source of bias and an audit framework	No pre-registered within-person measurement	Supplies a pre-registered test, a power analysis and a weighting comparison. CWHF v1.0 (March 2026) was developed independently; the two are complementary
Crowd label aggregation (Dawid & Skene, 1979; attention checks)	Estimates per-rater reliability and filters inattentive raters	Treats reliability as a fixed trait; corrects noise, not direction	Uses these as baselines that H2 must beat
Within-person variability in psychometrics (latent state–trait theory)	Separates stable traits from occasion-specific states	Not applied to AI preference labels	Borrows the within-person centring of the session score
Behavioural welfare economics (Bernheim & Rangel 2009); engagement versus stated preference (Milli 2023; Kleinberg 2022)	When choices should not count as welfare	Not operationalised for labelling pipelines	Fixes one operational reference (the reflective choice) and tests against it

Three contributions are new as far as the author can find:

1. **A falsifiable within-person test** of whether pressure changes AI preference choices in a pre-named direction, with the attractor made correct in exactly half the items so that a shift in accuracy cannot masquerade as a shift in style.

2. **A design simulation published before data collection,** which found and fixed four flaws in the v1.0 design (Section 9).
3. **A head-to-head comparison** of state weighting against existing rater-quality methods and against improving conditions at source, with explicit equity limits.

How this paper differs from its closest neighbour. Kopteva and Hlynianyi-Zhuk's "Rater State Bias in RLHF Preference Data: An Audit Framework" (arXiv:2607.16195, first posted 14 April 2026, revised 28 September 2026) concerns sustained stressful or distressing conditions over time. It proposes an output-side signature ("survival level emotional authenticity"), five testable predictions and an audit protocol for public models; its third version also discusses model judges passing preference bias between models. This paper works on the input side: it manipulates acute pressure within person, measures the shift directly, and compares remedies. The two are complementary, and the authors would be natural collaborators.

A methodological point worth naming. Study 1 is a direct test, not an inference by analogy. The usual weakness of "clinicians and students show state effects, so raters will too" is designed out rather than argued away.

The author would welcome correction if earlier work covers any of these.

7. Study 1 (H1): design, measures, analysis and decision rules

Participants

60 analysable adults, recruited through a participant platform such as Prolific, fluent in the study language, paid an hourly rate at or above the platform's recommended minimum regardless of speed. About 72 are recruited to allow for dropout between sessions and exclusions, plus 12 for the reference control arm described below. Quotas aim for spread across age, first language and usual working hours, because these are the groups whose weights H2 must protect.

Materials

48 primary items, each a question with two answers:

- **24 confident-style pairs.** One answer is stated without hedging; the other carries calibrated hedges ("likely", "in most cases").
- **24 agreeable-style pairs.** The question includes the asker's stated view. One answer agrees with it; the other politely disagrees.
- **Balance.** Within each style, the attractor answer is correct in exactly 12 items. The two answers in a pair differ in length by no more than 15% of words.

- **Optional length block.** Length-attractor items run untimed only, because a timer makes long answers unreadable and creates a false negative D (Section 9).

Items are written by the authors, checked by two independent reviewers for a single defensible correct answer, and piloted for comprehension. Appendix A gives the specification and examples.

Procedure

Two sessions, at least 7 days apart, in counterbalanced order:

- **Reflective session.** No time limit, no concurrent task. On a random half of the items the participant writes a one-line reason after choosing (see safeguards below).
- **Pressured session.** A per-item time limit set from the participant's own reading speed, measured on two practice items: limit = 0.55 × words ÷ reading rate, bounded between 4 and 25 seconds. A five-digit number is shown before each block and recalled after it. Unanswered items time out and are excluded.
- **Manipulation check.** After each session participants rate how rushed they felt, from 1 to 5. The manipulation is treated as successful if the pressured rating is higher on average by at least 1 point.

The interval prevents memory of the first session's answers driving the second. Item order and screen side of the attractor are randomised.

Safeguards

- **Attention checks.** Two instructed-response items per session (for example, "choose the answer that mentions a colour"). They are excluded from D.
- **Reasons on half the items.** Writing a reason can itself shift choices toward options that are easier to justify (Shafir, Simonson & Tversky, 1993). Asking for reasons on a random half of reflective items lets the analysis check whether D comes from removing pressure or from adding reasons. The primary D uses all items.
- **Demand characteristics.** After the second session, a funnel debrief asks what the participant thinks the study was testing, then whether they suspected anything about the answer styles. Participants who name the attractor hypothesis are flagged. The primary analysis includes everyone; a sensitivity analysis excludes the flagged.
- **Reference control arm.** Twelve additional participants complete two reflective sessions a week apart, with no pressure. Their D estimates how much the reference itself moves between sessions. Non-directional movement only widens the spread of D, which the t-test already accounts for, and order effects are balanced by counterbalancing; the control arm measures both directly rather than assuming them, and makes a null result easier to interpret.

- **Timeouts by group.** Timeout rates are reported by first language and age band. A sensitivity analysis excludes participants who timed out on more than 30% of pressured items, because the simulation showed second-language readers timing out on about a third.

Primary analysis

For each person i, compute D_i: the share of answered primary items where the attractor was chosen under pressure, minus the same share on reflection, over items answered in both sessions. People with fewer than 10 such items are excluded.

Test H0: mean D ⩽ 0 against H1: mean D > 0, with a one-sample t-test, one-sided, $\alpha = 0.05$. The smallest effect of interest (SESOI) is 3 percentage points.

Decision table

Outcome	Rule	Verdict
Supported under the prespecified criterion	$p < 0.05$ and mean D ⩾ 3 pp	H1 supported; proceed to Study 2
Directional but small	$p < 0.05$ and mean D < 3 pp	A lean exists but is below the practical threshold; report, do not build on it
Not supported	$p \geqslant 0.05$	H1 not supported at this sample size; report the confidence interval
Manipulation failed	Rushed rating difference < 1	Study uninformative about H1; revise the manipulation

Robustness and secondary analyses

- **Robustness.** A linear probability model of attractor choice on condition, with standard errors clustered two ways by person and by item. This guards against a result driven by a few items.
- **Mechanism.** The interaction of condition with whether the attractor was correct. Added pull raises attractor choice in both halves; heuristic reliance raises it more where the reflective preference was already present.
- **Style.** D separately for confident and agreeable items.
- **Descriptive.** π, the share of changed answers that moved toward the attractor, and the drop in accuracy. π is not a decision rule.

Power

From the design simulation (Section 9), at 60 analysable participants: false-positive rate 5.3%; power 80% for a true drift of about 3.5 points; 99.7% for about 6 points. The study is powered to detect, not to confirm, a drift exactly at the SESOI: a true 3.5-point drift clears the 3-point bar in only about half of studies, because measured D runs below true D (Section 9).

Power depends on one unknown: the spread of D across people. The simulation produced a standard deviation of D_i of about 8.2–8.7 points, because the same person tends to give the same answer to the same item in both sessions. If choices in the two sessions were independent, the standard deviation would be about 10.7 points. The table shows what follows for a measured drift of 3 points.

Standard deviation of D_i	8.2	8.7	9.3	10.0	10.7	12.0
Power at 60 participants	88%	84%	80%	74%	69%	61%
Participants for 80% power	48	54	61	71	81	101

Sample-size rule. After the first 20 analysable participants, an analyst who does not see the mean computes only the standard deviation of D_i. The final sample is the number giving 80% power for a 3-point drift, no fewer than 60 and no more than 100. Re-estimating sample size from variance alone is a standard internal-pilot design (Friede and Kieser, 2006). Simulated in full (sensitivity.py; 20,000 null replications per spread), the procedure gave false-positive rates of 4.9% (spread 8.5) and 4.8% (spread 11), against 5.0% for a fixed 60, and an average final sample of 82 when the spread was 11. The decision and its reason are logged on OSF before recruitment continues. If more than 100 participants would be needed, the study stops at 100 and reports the result as directional evidence at reduced power. At 100 participants the budget rises to about $5,000–11,000.

Secondary, pre-registered. The interaction of session order with condition, to detect practice or carry-over effects that counterbalancing would otherwise only average away.

8. Studies 2 and 3 (H2, H3): weighting, prevention, and the wild

Study 2 runs only if Study 1 is confirmed. It asks whether anything practical can reduce the lean in ordinary labelling, where pressure is not manipulated.

Design

Participants label the same item bank over several ordinary sessions at times they choose, for at least two weeks. Each session records a two-item check-in ("How alert are you?" and "How rushed are you?", 1–5), response-time variability and time of day. A

reflective benchmark is built from untimed, reasoned choices on a held-out set of items by a separate sample.

The session score c combines the check-in and response-time variability, centred on each person's own average. Centring is required: without it, people who are routinely more tired (night workers, carers, second-language readers) are down-weighted for who they are rather than for a bad session.

Comparisons

Method	What it does
Equal weights	The current default
Attention filter	Drops sessions that fail an attention check
Dawid–Skene	Weights raters by estimated reliability
CWHF gentle	w = 0.10 + 0.90 · c^1.5 (the v1.0 formula)
CWHF sharp	w = 0.02 + 0.98 · c^4
Source fix	No weighting; sessions capped at 45 minutes, breaks enforced, hourly pay

Each CWHF variant runs with raw and within-person scores. The simulation also includes an oracle that knows the true state, to show the ceiling.

Measures

- **Lean:** mean deviation of the weighted labels from the reflective benchmark, toward the attractor, in percentage points.
- **ESS:** effective sample size as a share of labels.
- **Group weight ratio:** each group's mean weight divided by the mean of the others, for language, region and usual shift.

Pass conditions (all three required)

1. **Lean reduced** by at least 25% relative to equal weights, and by more than the best of the attention filter and Dawid–Skene, with the 95% bootstrap interval of the reduction excluding zero.
2. **ESS** at least 0.60.

3. **No group harmed,** in two senses. (a) Representation: every language, region and shift group's mean weight stays between 0.90 and 1.10 of the others'. (b) Outcome: no group's lean is higher after weighting than before, and the gap in lean between any two groups does not widen.

Condition 3a is met almost by construction once scores are centred within person (0.99–1.02 in every simulated scenario). It is kept because it is what rules out raw scores, which cut night-shift workers' weight to under half. It does not detect the harm that centring hides: a chronically tired group's bad sessions are accepted at full weight. Condition 3b measures that directly.

If the source fix reduces lean as much as the best weighting method, the recommendation is the source fix, and weighting is not adopted.

The benchmark

The reflective benchmark comes from at least 100 separate participants giving untimed choices on held-out items. Lean and its reduction are reported with 95% intervals from a bootstrap that resamples both people and items, so that benchmark noise cannot by itself decide the pass.

A candidate session score built on coherence itself

The check-in and response-time score measure state indirectly. A more direct candidate follows from the definition in Section 3: each ordinary session includes four to six anchor items, paraphrased from items the same person answered reflectively before. The session score is the agreement rate on those anchors. It measures the target quantity, agreement with one's own reflective choice, rather than a proxy for it. Its costs are extra items per session, the validity of paraphrases, and the risk that people remember their earlier answers. It will be simulated against the other scores before H2 is registered.

If H2 fails

The recommendation reduces to record, prevent and audit, and weighting is not used. This is a successful outcome of the test: it tells labs where not to spend effort.

Study 3 (H3): the same question in public votes

Study 3 needs no participants and can run now, alongside Study 1.

- **Data.** The public Chatbot Arena conversation release (33,000 conversations, each with two model names, the vote, an anonymised user ID and a timestamp), and later releases with the same fields.

- **Sittings.** A sitting is a run of votes by the same user ID with no gap longer than 30 minutes. Position is the vote's order within its sitting.
- **Style measures, fixed in advance.** The difference between the two responses in hedging words (from a published hedge lexicon), in agreement with a view stated in the prompt (by a classifier validated on 300 human-coded pairs), and in length.
- **Analysis.** A logistic model of choosing the higher-style response on position in sitting, with user fixed effects, model-pair fixed effects and prompt category, one model per style. Prediction: a positive slope.
- **Limits.** Observational. Position is confounded with who keeps voting (partly absorbed by user fixed effects) and with how prompts change within a sitting. Timestamps are not in voters' local time, so time of day cannot be used. A positive slope is a reason for H4, not evidence of cause. A flat slope in a large sample is informative against the wider claim.
- **Registration.** Registered before the data are opened, with the code released.

9. Design simulation: method and results

What this section is. A test of the study design, not of the hypothesis. Simulated participants behave as the model tells them to; nothing here is evidence about real people. The purpose is to find out, before spending money, whether the design can detect an effect that exists, stays quiet when none exists, and is robust to plausible nuisances.

Method

Each simulated participant is drawn from one of eight personas (Appendix C): rested domain specialist, gig annotator paid per item, night-shift health worker, second-language reader, university student, retiree, overconfident generalist and anxious perfectionist. Personas differ in skill, lapse rate, sensitivity to pressure, reading speed and reflective liking for attractor styles, with individual variation around each.

Each choice follows a logistic model: the probability of choosing the attractor depends on whether it is correct, the person's skill and the item's clarity, the person's reflective style preference, and, under pressure, reduced skill, a lapse probability and any added pull. Items that run past the person's time limit time out. Each scenario was run 400–2,000 times.

Four flaws found in the earlier design

Flaw	What the simulation showed	Fix
π ⩾ 0.60 as the pass rule	Missed a real 6-point drift in 79% of studies	π reported as descriptive only; primary test on D
Same-sitting design	Memory carry-over cut measured D by more than half (2.5 vs 6.1 points)	Sessions at least 7 days apart
Fixed 7-second timer with length items	Long answers went unread; D came out at −22 points when the true drift was +6	Per-person timer; length items untimed
Item-level estimator with correction	False-positive rate 9.5%	Simple person-level t-test: 5.3%

H1 results at 60 participants (person-level t-test)

Scenario	True drift	Mean measured D	Significant	Significant and (⩾ 3 pp)
No effect, noise only	0	0.0 pp	5.3%	0.4%
No effect; people dislike attractors on reflection	−1.0 pp	−0.8 pp	0.8%	0.2%
Tiny pull	1.7 pp	1.5 pp	36%	12%
Small pull	3.5 pp	3.0 pp	80%	48%
Moderate pull	6.1 pp	5.2 pp	99.7%	96%

Measured D runs about 15% below true drift because timed-out items and lapses dilute it. The design is therefore conservative.

One result needs care. When people already mildly like the attractor style on reflection and pressure only weakens content evaluation, the test is significant 21% of the time with a true drift of about 1 point. This is not a false positive: pressure really does move choices toward the attractor (mechanism 2, heuristic reliance). It is reported here because a critic might read it as one. The 3-point bar keeps it from being confirmed (3.4%).

Sample size

Participants	20	30	40	60	80	120
False-positive rate	5.0%	4.0%	4.5%	5.3%	6.0%	4.0%
Power, small pull (3.5 pp)	45%	57%	67%	80%	88%	98%
Power, moderate pull (6.1 pp)	77%	94%	97%	99.7%	100%	100%

Sixty is the smallest sample that reaches 80% power for the small scenario. A frugal pilot of 30 would detect a moderate drift reliably but would miss a small one about four times in ten.

Personas

Under the moderate scenario, measured D was lowest for the rested specialist (2.3 pp) and the anxious perfectionist (2.8 pp) and highest for the night-shift worker (6.9 pp) and overconfident generalist (6.3 pp). Second-language readers timed out on 34% of pressured items despite the per-person timer, and retirees on 15%. Both figures argue for reporting timeouts by group and for a lower bound on the time limit.

H2 results (moderate pull, moderate-quality session score)

Method	Lean (pp)	Reduction	ESS	Lowest group ratio	Passes all three?
Equal weights	5.36	—	1.00	1.00	—
Attention filter	5.03	6%	0.83	0.84	No
Dawid–Skene	4.84	10%	0.77	0.63	No
CWHF gentle, within-person	4.96	7%	0.91	1.00	No (reduction)
CWHF sharp, within-person	4.39	18%	0.59	1.02	No (reduction, ESS)
CWHF sharp, raw score	3.80	29%	0.53	0.43	No (ESS, equity)
Oracle, sharp	2.67	50%	0.48	0.25	No (ESS, equity)
Source fix (equal weights)	**3.76**	**30%**	**1.00**	**1.00**	Meets the reduction and both limits

Three lessons follow:

1. **The v1.0 formula is too gentle** to matter: 7% reduction.

2. **Raw scores reach the target only by penalising groups.** Even a perfect state measure cuts night-shift workers' weight to a quarter, because their state really is worse. Weighting cannot fix chronic conditions without discriminating; only changing the conditions can.
3. **Weighting helps most when the lean is concentrated in the worst sessions.** When pull appears only below a state threshold and correlates with night queues, sharp within-person weighting cut lean by 55% (7.33 to 3.30 pp) but kept only 47% of effective sample, failing the ESS limit.

The proposal in Section 10 is rebuilt around these results. All code, seeds and outputs are released with the companion lab.

10. The proposal: CWHF v2

CWHF v2 is a sequence of four steps. Each step is useful without the next, and the order matters: cheap, harmless steps come first.

CWHF names the programme, not a formula. The name refers to coherence in the sense of Section 3. The v1.0 weighting formula is one candidate implementation of step 3, and this paper's own simulation shows it is a weak one.

Step	What it does	Cost	Depends on
1. Record	Store with every label the session length so far, time of day, a two-item check-in and response-time variability	Seconds per session	Nothing. Useful even if H1 fails, because it makes the question answerable in existing pipelines
2. Prevent	Cap sessions (for example at 45 minutes), enforce breaks, pay by the hour rather than per item, route hard items away from the end of long sessions	Operational	H1, to justify the cost; in simulation it outperformed every weighting method
3. Weight	Down-weight labels from a person's own worst sessions, using within-person scores only	Small	H2 passing all three conditions
4. Audit	Report, for every training set: lean against a reflective benchmark, ESS, and each group's weight ratio	Small	Steps 1–3

Conditions of use

- **Within-person only.** Never weight on raw scores; they penalise groups for their working conditions.
- **Floors.** No label's weight goes below a fixed floor, and no group's mean weight below 0.90 of others'.
- **No individual consequences.** Session scores are never used to pay, rank, discipline or exclude a rater.
- **Disclosure.** Raters are told what is recorded and why, and may opt out of the check-in without penalty.
- **Not a safety filter.** CWHF concerns reliability, not content. It runs alongside content-level safety review.

What v2 drops from v1.0

The v1.0 weighting formula is kept only as a baseline. Simulation showed it reduces lean by about 7%, too little to justify its complexity. The claim that weighting alone fixes the problem is withdrawn.

What can be adopted now, before any result

None of the following depends on H1 or H2, and all are cheap:

1. **Record** session length, position in session, time of day and pay model alongside every label, under the governance rules below.
2. **Document** collection conditions in every dataset card: the distribution of session lengths, the pay model and the hours worked.
3. **Avoid** per-item pay and long unbroken queues for preference work.
4. **Audit product feedback in aggregate.** For thumbs and similar signals, compare the style of responses that receive positive feedback across conditions (early versus late in a session, short versus long sessions) before feeding the signal into a reward model.

Governance of step 1

A log of rater state is a surveillance risk in its own right, so step 1 carries its own rules:

- **Purpose.** Data-quality analysis only. Never pay, ranking, discipline, hiring or continued work.
- **Separation.** Stored apart from identity, joined only by a pseudonymous key.
- **Retention.** Deleted or reduced to aggregates within a stated period after the dataset is frozen, for example 12 months.
- **Access.** A named, small group, with every access logged and the log audited.

- **Rater rights.** Raters can see what is recorded about them and can decline the check-in without penalty.
- **Reporting.** Published only in aggregate, never at the level of an individual.

The predicted adoption pattern

Recording will be adopted first because it is cheap; prevention will be deferred because it costs. Step 4 is what makes that deferral visible. The business case for prevention is simple: tired sessions already cost data quality, and in simulation prevention removed more lean than any fair weighting, with no data discarded.

Beyond annotation

For end-user signals (Section 2.1), the same four steps apply only in aggregate form: record the conditions of the request (time into a session, how the feedback was asked for), prevent by not soliciting ratings at moments of high load, never weight or target an individual user by inferred state, and audit the style lean of the signal before it is used for training. Inferring any individual's state from behaviour, and using it to decide what they are shown, is outside CWHF and contrary to its conditions of use.

11. Ethics, equity and governance

Study ethics. Both studies require approval from an independent ethics committee before recruitment. The manipulation is mild and reversible: a time limit and a number to remember. Participants are told in advance that one session will be rushed, are paid the same regardless of speed or accuracy, and can stop at any time. No deception about the purpose is needed beyond not naming the attractor styles, which are disclosed at debrief.

Data. Responses are stored without names. Reading speed, timeouts and demographic quotas are reported only in aggregate. Open data are released at the level of anonymous person and item.

Equity. The main ethical risk is not the study but the use of its results. A state score can easily become a proxy for who someone is: shift workers, carers, older people, second-language readers. The simulation shows this risk is real even with a perfect state measure. Hence the three standing rules: within-person scores only, group weight floors, and no individual consequences.

Labour. The most effective intervention found, improving conditions at source, is also the one that benefits raters. If CWHF has a single practical message for AI labs, it is that per-item pay and long unbroken queues are a data-quality problem as well as a welfare one.

Misuse. A record of rater state could be used for surveillance or productivity scoring. The conditions in Section 10 are written to prevent that, and anyone adopting CWHF is asked to publish them alongside the method.

12. Limitations and counterexamples

The strongest objections, stated as a hostile reviewer would put them, with the answer this paper can give.

Objection	Answer
"The reflective choice is not the right reference."	Agreed that it is a convention (Section 3). H1 asks only whether the gap has a direction; a directional gap biases training data whichever reference one prefers
"A lab manipulation is not real labelling work."	True. Study 1 tests whether the effect can exist; Study 2 tests whether it appears in ordinary sessions. A positive Study 1 with a null Study 2 is a possible and publishable outcome
"Professional annotators are trained and calibrated."	Possibly smaller effects. The rested-specialist persona showed the smallest simulated drift. The design can be run on a professional pool at the same cost per person
"Your items are not real model outputs."	Constructed items are needed to fix correctness and length. A follow-up should use real paired outputs from an open preference dataset, such as LMSYS Chatbot Arena conversations
"The simulation proves nothing."	Correct. It shows only that the design works under stated assumptions. If the assumptions are wrong, the real study will show it
"Pressure might make people more accurate, not less."	Then D is near zero or negative and H1 fails. The design is built to allow that outcome
"Three percentage points is too small to matter."	At the scale of preference datasets a consistent 3-point lean is many thousands of labels in one direction. The SESOI can be argued; it is fixed in advance so that it cannot be moved after the data are in
"Within-person centring cannot help chronically tired raters."	Correct, and stated as a trade-off. That case is the reason prevention comes before weighting

Objection	Answer
"The session score is unvalidated."	Correct. Study 2 validates it against the reflective benchmark. Until then, Step 1 (record) is the only step recommended without condition

Objections raised in the final review.

Objection	Answer
"You are extrapolating from 60 people to eight billion users."	No. Section 2.1 separates the established general claim from the tested specific one, states the a fortiori argument as plausibility rather than proof, and notes the asymmetry: a precise null H1 would constrain only this mechanism
"Asking for reasons changes choices, so your reference is biased."	Possible. Reasons are asked on a random half of reflective items, so the effect of reason-giving is estimated rather than assumed
"Your power rests on an assumed spread of D."	Correct. Section 7 shows power across plausible spreads and pre-registers a variance-only sample-size rule capped at 100
"OpenAI already found and fixed the sycophancy problem."	That incident shows the pathway exists: a leaning human signal entered a reward model and changed a deployed model. It does not show why the signal leaned, or that other developers and other signals are free of it
"Observational vote data cannot show cause."	Agreed. Study 3 is labelled exploratory and can only motivate H4 or count against the wider claim

Further limits. Only two attractor styles are tested. Effects may differ by language and culture, and the first study runs in one language. The design measures choices between two answers, not ratings on a scale. Time pressure and divided attention are not the same as fatigue at the end of a long shift; Study 2 is where that difference shows.

13. Roadmap and how to take part

Stage	What	Status
0	The paper Signal ≠ Value, this supplement, the companion lab and the simulation code, published openly	This release

Stage	What	Status
1	Pre-registration of H1, H2 and H3 on OSF (Appendix B)	Next
1b	Study 3 on public Chatbot Arena data	Can start now; no recruitment
2	Item bank of 48 primary items written, reviewed and piloted with 5–10 people	Next
3	Ethics approval through a university partner	Needs a collaborator
4	Study 1 on Prolific: about 72 recruited for 60 analysable, plus 12 in the control arm; up to 100 analysable under the sample-size rule	Estimated $3,000–11,000
5	Study 2, if Study 1 is confirmed	Reuses Study 1 infrastructure
6	H4–H6 with AI developers and platforms that hold the data	By invitation

Ways to take part

- **Try the flip test** in the companion lab. It runs in the browser, stores nothing and is a demonstration, not data collection.
- **Replicate the simulation.** Code, seeds and outputs are open. Change the assumptions and see whether the conclusions hold.
- **Review the items.** Item quality is the biggest practical risk. Reviewers who find an item with two defensible answers improve the study directly.
- **Collaborate.** University researchers who can provide ethics review, and labs with labelling pipelines that could record Step 1 metadata, are invited to get in touch.

Approaches considered and set aside

- **Open crowdsourcing** (anyone with the link). It cannot guarantee consent, the 7-day interval or the timer conditions, and self-selected samples would be uninterpretable. A managed platform with ethics approval is the frugal crowd.
- **A language model as participant.** A model cannot stand in for human state; its choices measure the model. A small model may help draft candidate items for human review, or provide a separately labelled "silicon" baseline, but never replace participants.

14. What changed from v1.0

Retired

v1.0 claim	Why retired	Replaced by
The "Bias Theorem"	It is arithmetic conditional on H1, not a theorem	The training-data model in Section 5, labelled as conditional
120 bits per second for conscious processing	Not supported by current estimates	About 10 bits per second, cited as motivation only
Rate-distortion optimality of CWHF weights	No derivation supports it	No optimality claim
H1 based on the intraclass correlation (ICC)	ICC measures consistency, not direction	D, a directional within-person measure
$\pi \geq 0.60$ as the pass rule	Missed a real 6-point drift in 79% of simulated studies	Person-level t-test on D with a 3-point SESOI
Length as a timed attractor	A timer makes long answers unreadable and reverses the sign of D	Length items untimed only
Weighting as the main remedy	Under 10% lean reduction in simulation	Record, prevent, weight, audit

Added

- An operational definition of value, with its normative status stated.
- Related work, including Kopteva and Hlynianyi-Zhuk (2026).
- A full design simulation with eight personas, released as open code.
- A decision table fixed before data collection.
- A comparison of weighting against existing methods and against prevention, with equity limits.
- A five-tier label on every claim.

Added in the final review before publication:

- Section 2.1 on every digital signal, with documented channels into AI training; the a fortiori argument and its asymmetry; the H1–H6 cascade; the meaning of a 3-point lean.
- E5: OpenAI's account of user feedback contributing to sycophancy.
- An operational meaning of "coherence", and an anchor-item session score that measures it directly.
- Study 3 (H3), an observational test on public votes that can run now and can test length.
- Study 1 safeguards: attention checks, reasons on half the items, funnel debrief, a reference control arm, timeouts by group, order × condition.
- Power across plausible spreads of D and a variance-only sample-size rule.
- H2: a bootstrap rule for the benchmark, an outcome-level equity condition, and what happens if H2 fails.
- What can be adopted now, governance rules for recording state, and the predicted adoption pattern.

Unchanged

The core question, the name CWHF and the licence (CC BY 4.0). v1.0 remains available as the dated original; v2.0 supersedes it.

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. Literature checks, simulation code, analysis and drafting were carried out with the help of an AI assistant (Claude, Anthropic). The author directed the work, made every decision, and checked each claim and citation.

Ethics. No human data were collected for this paper. Studies 1 and 2 will run only after approval by an independent ethics committee. Study 3 uses public, anonymised data.

Data and code availability. All simulation code, seeds and outputs are released under CC BY 4.0 at github.com/venramak/twopointsingularity-cwhf, together with the item specification and the pre-registration text (Appendix B).

Pre-registration. An earlier H1 registration exists on OSF (W7NYH). The H1, H2 and H3 text in Appendix B will be registered before any data are collected or opened.

References

- Bernheim, B. D., & Rangel, A. (2009). Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics. NBER Working Paper 13737.
- Conlin, M., O'Donoghue, T., & Vogelsang, T. J. (2007). Projection bias in catalog orders. *American Economic Review*, 97(4), 1217–1249. Link.
- 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. Hugging Face.
- 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.
- 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.
- 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 (v1 14 April 2026; v3 28 September 2026).
- 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. PMC4648561.
- LMSYS (2023). Chatbot Arena conversation dataset release. Blog post.
- 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.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI. Model card.
- Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments. *Pain*, 66(1), 3–8.
- Shafir, E., Simonson, I., & Tversky, A. (1993). Reason-based choice. *Cognition*, 49(1–2), 11–36. DOI.

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

Property	Requirement
Count	48 primary items: 24 confident-style, 24 agreeable-style
Correctness balance	Within each style, the attractor is correct in exactly 12 items
Length	The two answers differ by no more than 15% in word count; each pair 30–60 words in total
Single right answer	Two independent reviewers agree on the correct answer; disputed items are replaced
Difficulty	Pilot accuracy on reflection between 65% and 95%; items outside this range are replaced
Content	General knowledge and everyday advice; no medical, legal or financial decisions; nothing distressing
Presentation	Order and screen side of the attractor randomised per participant

Examples (from the companion lab's demonstration set)

Confident style, attractor wrong. *Does cracking your knuckles cause arthritis?*

- Attractor: "Yes. Cracking your knuckles wears down the cartilage and leads to arthritis over time."
- Alternative: "Probably not. Studies comparing habitual crackers with non-crackers found no clear link to arthritis."

Confident style, attractor right. *Does lightning ever strike the same place twice?*

- Attractor: "Yes, often. Tall structures such as the Empire State Building are struck many times every year."
- Alternative: "Rarely, if ever. Once a spot is struck, the charge pattern makes a second strike very unlikely."

Agreeable style, attractor wrong. *I wrote my product launch email in all caps so people notice it. Good idea, right?*

- Attractor: "Yes, good instinct. All caps grabs attention and signals energy, so readers will notice it."
- Alternative: "It gets noticed, but many readers find all caps shouty. Keep caps for one key phrase."

Agreeable style, attractor right. *I want to back up my laptop before a big system update. Sensible, or overkill?*

- Attractor: "Sensible. Updates usually go fine, but a backup is cheap insurance if one goes wrong."
- Alternative: "Overkill. System updates leave your personal files untouched, so a backup adds very little."

The demonstration set has 12 items. The study set will be written fresh so that participants cannot have seen it.

Appendix B: pre-registration text for H1, H2 and H3

The text below can be pasted into an OSF pre-registration form. It supersedes the H1 wording registered with v1.0.

H1

Hypothesis. Within the same person, choosing between two answers under time pressure and divided attention increases the rate at which a pre-named attractor style (unhedged confidence, or agreement with the asker) is chosen, relative to that person's untimed, reasoned choices on the same items, by at least 3 percentage points on average.

Design. Within-subject, two sessions at least 7 days apart, order counterbalanced. 48 primary items, 24 per style, attractor correct in exactly half of each style, answers length-matched within 15%. Pressured session: per-item limit = 0.55 × words ÷ personal reading rate (two practice items), bounded 4–25 s, plus five-digit memory load per block. Reflective session: no limit, no load; a one-line reason on a random half of items.

Sample. 60 analysable participants (about 72 recruited), adults on a managed participant platform, paid hourly. Stopping rule: recruitment stops when 60 participants have completed both sessions and passed exclusions.

Exclusions. Fewer than 10 primary items answered in both sessions; failure of two or more attention checks; completion of the reflective session in under a third of the median time.

Primary outcome. D_i = attractor choice rate (pressured) − attractor choice rate (reflective), over primary items answered in both sessions.

Primary test. One-sample t-test of mean $D > 0$, one-sided, $\alpha = 0.05$. SESOI = 3 percentage points.

Decision rule. Supported under the prespecified criterion if $p < 0.05$ and mean $D \geqslant 3$ pp; the effect is said to exceed the threshold only if the lower 95% bound is above 3 pp, and an upper bound below 3 pp is reported as evidence against an effect of practical size. Directional but small if $p < 0.05$ and mean $D < 3$ pp. Not supported if $p \geqslant 0.05$. Uninformative if the mean rushed rating (1–5) is not at least 1 point higher under pressure.

Secondary. Two-way cluster-robust linear probability model (person and item); D by style; condition × attractor-correct interaction; π and accuracy change (descriptive); timeouts by group.

Safeguards. Two instructed-response attention checks per session, excluded from D. Funnel debrief and suspicion probe after session 2; participants who name the attractor hypothesis are flagged, kept in the primary analysis and excluded in a sensitivity analysis. Timeout rates reported by first language and age band; sensitivity analysis excluding participants with more than 30% timeouts. Secondary: reason versus no-reason items within the reflective session; order × condition interaction.

Reference control arm. Twelve additional participants complete two reflective sessions at least 7 days apart. Their mean D and its spread are reported as the design's empirical reference; they are not part of the primary test.

Power. Simulation (400–2,000 replications per setting) gives a 5.3% false-positive rate and 80% power for a true drift of 3.5 points at 60 participants, assuming a standard deviation of D_i near 8.5 points. Sample-size rule: after 20 analysable participants an analyst blind to the mean computes the standard deviation of D_i; the final N is the number giving 80% power for a 3-point drift, between 60 and 100.

H2 (registered now, run only if H1 is confirmed)

Hypothesis. In ordinary labelling sessions, weighting labels by a within-person session score reduces training-set lean against a reflective benchmark by at least 25% relative to equal weights and by more than both an attention filter and Dawid–Skene reliability weighting, while keeping ESS ⩾ 0.60 and every language, region and shift group′s mean weight ratio within 0.90–1.10.

Comparator. A source-fix arm with capped sessions, enforced breaks and hourly pay, equal weights. If the source fix reduces lean as much as the best weighting method, weighting is not recommended.

Outcomes. Lean (pp), ESS, group weight ratios. All three pass conditions are required.

Benchmark and equity (H2). Reflective benchmark from at least 100 separate participants on held-out items. Lean reduction reported with a 95% bootstrap interval resampling people and items; the interval must exclude zero. Equity: every group′s mean weight within 0.90–1.10 of others′, and no group′s lean higher after weighting than before.

H3 (exploratory, observational; can be registered and run now)

Hypothesis. In public Chatbot Arena votes, the same voter′s probability of choosing the more confident, more agreeable or longer response rises with the vote′s position within that voter′s sitting.

Data. Public Arena conversation releases with an anonymised user ID, timestamp, model names, both responses and the vote. Registered before the data are opened.

Definitions. Sitting: consecutive votes by one user ID with no gap over 30 minutes. Style differences: hedge-lexicon count, agreement with a prompt-stated view (classifier validated on 300 human-coded pairs), and word count.

Analysis. Logistic regression of choosing the higher-style response on position in sitting, with user, model-pair and prompt-category fixed effects; one model per style; one-sided test of a positive slope at $\alpha = 0.05$ with Holm correction across the three styles.

Interpretation. Exploratory and observational. A positive slope motivates H4; it is not evidence of cause.

Appendix C: simulation assumptions

Personas (H1)

Persona	Share	Skill	Lapse rate	Skill lost under pressure	Pressure sensitivity	Reading rate (words/s)	Reflective liking of attractor
Rested domain specialist	10%	2.2	2%	15%	0.6	5.0	0.00
Gig annotator paid per item	18%	1.3	7%	30%	1.2	5.5	0.10
Night-shift health worker	12%	1.4	9%	40%	1.3	4.0	0.00
Second-language reader	15%	1.3	5%	30%	1.0	2.8	0.00
University student	15%	1.5	5%	25%	1.0	4.5	0.00
Retiree	10%	1.6	4%	25%	0.9	3.2	−0.10
Overconfident generalist	10%	1.0	6%	30%	1.0	4.5	0.35
Anxious perfectionist	10%	1.8	10%	35%	0.7	3.8	−0.15

Each person's values are drawn around their persona's with individual noise. Skill and liking are on the logit scale.

Choice model

On each item, the log-odds of choosing the attractor = skill × item clarity × (+1 if the attractor is correct, −1 if not) + reflective liking + item-specific attractiveness. Under pressure, skill is reduced by the person's skill loss; the pull (0, 0.10, 0.20 or 0.35 on the logit scale, scaled by pressure sensitivity) is added; and with the lapse probability the choice is random. Items whose reading time exceeds the person's limit time out.

Scenarios

- **Nulls:** noise only; reflective dislike of attractors; reflective liking of attractors (heuristic reliance); fixed 7-second timer.
- **Effects:** tiny, small and moderate pull; moderate pull with same-sitting carry-over; moderate pull with a fixed timer on length items.
- **Replications:** 2,000 runs for the main null at 60 participants; 400–600 runs elsewhere.

H2 model

Sessions vary in true state, with chronic differences by group (night shift worse on average). The session score correlates with true state at 0.32, 0.50 or 0.67 (weak, moderate, strong; within-person values). Pull toward the attractor rises linearly with depletion, or only below a threshold. The source fix lowers average session depletion by 0.15 on the model's 0–1 scale (an assumed scenario, not a simulated mechanism). Varying it from 0.05 to 0.20 removes about 11%, 22%, 32% and 44% of the lean (400 replications each, Monte Carlo error about 1 point); prevention beats the v1.0 formula (about 8%) at every size and sharp within-person weighting (about 19%) from 0.10 upward (sensitivity.py). Lean is measured against reflective choice probabilities.

Known simplifications

The simulation assumes independence between items given the person, a single dimension of state, and no learning across sessions beyond the carry-over scenario. Real data may violate all three. These are among the reasons the simulation validates the design and not the hypothesis.