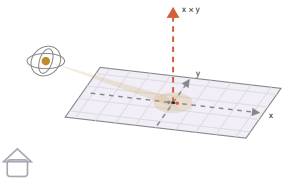
Plus one: you always have three choices

Spencer Hochberg

2026-11-28 · One more stage for Christiano, Leike, Brown, Martic, Legg and Amodei, “Deep reinforcement learning from human preferences” (NeurIPS 2017)



A: express train
B: scenic train
Up: $x \times y$, perpendicular to both. Say what is wrong with both
Home: stop, any time

Figure 1. Plus one, stage 2. One person, starting from the shared base, which rates each pair as equal. Two options differ along one direction of the plane (the line between the dots). A click splits that direction: at most 1 bit. The next pair is perpendicular, so it asks nothing already asked. Once the plane is split, pairs can refine it but never leave it. The out goes up, along the cross product of the two directions (the right-hand rule): a complaint in your own words names a direction the plane does not have, and the next pair is built along it. The controls are the gimbal’s axes: yaw leans (graded), pitch goes up, roll asks a fresh pair. Home is always there.

MY WORDS

- “You always have three choices.”
- “I recommend that you constrain the two choices to be orthogonal, and you can always go up.”
- “What is the up button? The up is just talking … just complain about it.”
- “You can just go home. No one’s making you go anywhere.”
- “I’m building on Claude Shannon’s information theory and saying it’s intention theory.”

Dictated 2026-11-28, evening PT. Then, typed about 7:35 PM PT:

> “i think maybe their solution is ok because tuning it to everyones preference is the same as tuning it to nobody's preference. my system only feels good to me because i get to encode my own preference. so maybe their way for rlhf cause its fast, and then the user gets to do their own post training to converge intention”

Abstract

Shown two options, a person has three choices: one, the other, or say what is wrong with both. This note adds one stage and one choice.

Christiano et al. (2017) taught agents a reward from people choosing between two short clips. The method is fast and learns from many people at once, which makes it the right first stage: a shared base. A reward tuned to everyone’s preference, though, is tuned to no one’s in particular. I propose a second stage, run by one person on top of the base, that converges on that person’s intention, and make three claims about it. **Capacity.** A two-way choice carries at most 1 bit and a three-way choice at most $\log_2 3 \approx 1.585$ bits; a free-text answer has no fixed ceiling, and at Shannon’s estimate of about one bit per letter of English a ten-word complaint carries on the order of 50. That counts how many different complaints could be told apart, not how much they can mean: a short message can point to anything both sides already share, so its meaning is bounded by the shared context, not by its letters. One person answers few questions, so each must carry as much as it can. **Geometry.** For a linear reward under a Gaussian prior, two balanced comparisons carry $1 + H_b(\theta/\pi)$ bits jointly, where θ is the angle between their difference vectors; the maximum, 2 bits, holds only at $\theta = 90°$. I prove this and check it by simulation. **Protocol.** Each pair is one the base rates as equal and is orthogonal to every direction already asked. The out is a complaint; its text names a direction, and the next pair is built along the part of it not yet asked. Going home is always allowed and costs nothing. I call the move from what a channel can carry to what a person meant *intention theory*; Weaver (1949) asked its question, and this note makes one part of it quantitative. I name the attack this does not stop by itself, a proposer that drifts what both options share, and give a pre-registered test that would falsify the defense. A first test on 87 of my own recorded decisions missed its sealed pass line: my history added 11.5 points of accuracy, with an interval that crosses zero. In a toy simulation of the attack, pairs alone were captured every time; the out protected a fixed taste, and a log of head against beam rotation caught steering of a taste that moves.

1. Introduction

Shannon measured what a message can carry and set meaning aside. Weaver (1949) named meaning as the next problem. Intention theory measures how much of what one person wants arrives.

Shannon (1948) opened information theory by declaring the meaning of a message outside the engineering problem; what mattered was that it was chosen from a set of possible messages. Weaver (1949), in the book that reprinted it, named the two problems beyond: whether the receiver’s interpretation matches the meaning the sender intended, and whether the meaning received leads to the conduct the sender wanted. Intention theory is not a new question; it is one part of Weaver’s, made measurable for one person’s preferences. Preference learning inherits the set and adds the meaning back. Christiano et al. (2017) showed that an agent could solve Atari games and simulated locomotion tasks from human comparisons on less than 1% of its interactions, and learn novel behaviors from about an hour of human time. The channel was a pair of clips one to two seconds long, and a click.

That method works, and this note keeps it. It pools many people’s comparisons into one reward, which is what makes it fast and what makes it a good shared base. Pooling has a cost the 2017 paper did not need to address: Siththaranjan et al. (2023) show that standard preference learning implicitly aggregates hidden differences between people by Borda count. A reward tuned to everyone’s preference is tuned to no one’s in particular. What is left is one person’s own intention, and only that person can supply it.

So I add a stage, and call it *plus one*:

1. **Stage 1, the 2017 method, unchanged.** Many people, many pairs, one shared reward.
2. **Stage 2, one person.** Starting from the shared reward, the person steers with two options that are orthogonal to everything already asked, plus an out that goes up a dimension. The out is talking: the person says what is wrong with both, and the next pair is built along what they said.
3. **Home.** The person may stop at any point. Nothing is inferred from stopping, a deliberate departure from Jeon et al. (2020), and the shared base is still there.

I contribute a capacity bound (Section 3), a proof and simulation that orthogonal pairs maximize the joint information of two questions (Section 4), the second-stage protocol (Section 5), a gimbal and a ladder of interfaces for steering it (Sections 6 and 7), a first test on one person's records, which missed its sealed pass line (Section 8), and a falsification test for the strongest objection to it (Section 10).

A note on form. This paper began as a reply that declined both of the obvious options, agree or rebut. Its complaint about both, that a reward tuned to everyone is tuned to no one, named a direction neither option had: a second stage. The note is its own first out.

2. What the 2017 paper already does

The 2017 paper is a methods paper. It never claims two answers suffice, and it already holds both seeds of this note.

More than two answers. In Section 2.2.2 the overseer watches two clips and indicates a preferred segment, that the two are equally good, or that they are "unable to compare the two segments". Each judgment is stored as a triple $(\sigma_1, \sigma_2, \mu)$, where μ is a distribution over $\{1, 2\}$: all of its mass on the chosen clip, or uniform for a tie. An incomparable pair is not included in the database. So the interface has four answers and the database keeps three; the stored label can carry at most $\log_2 3$ bits per kept comparison. The fourth answer, the one that says the question itself is wrong, carries nothing forward. Stage 2 keeps it and lets it talk.

Asking where it helps most. Section 2.2.4 does not pick pairs at random. It samples many candidate pairs, has each member of an ensemble of reward predictors guess the winner, and shows the pairs on which the ensemble disagrees most. That is active learning: spend the person's attention where the model is least sure. The paper calls this a crude approximation and says that ideally queries would be chosen by their expected value of information, leaving that to future work. Orthogonal pairs, in Section 4, are a close cousin of this rule with a closed form: under a linear reward, a question perpendicular to everything already asked is one on which the model has learned nothing yet.

The credit runs one way. Stage 2 is the 2017 method's own discarded answer and its own stated ideal, applied to one person.

Checked against the paper's own text, arXiv:1706.03741v4 (sections 2.2.2 and 2.2.4), on 2026-11-28.

3. How much a choice can carry

A choice among m options carries at most $\log_2 m$ bits. Two options: 1 bit. Three: 1.585 bits. Words: no fixed ceiling.

If an answer X takes one of m values, its entropy satisfies $H(X) \le \log_2 m$, with equality only when all m answers are equally likely (Shannon 1948, Section 6). The mutual information between the answer and anything it is about, including the person's intent, is at most $H(X)$. That gives the ceilings below.

Table 1. The most one question can carry, by the answers it allows.

Answers allowed	Outcomes	Ceiling per question
A or B	2	1 bit
A, B, or equal (the 2017 stored label)	3	1.585 bits
A, B, equal, or can't compare (the 2017 interface)	4	2 bits; the last answer is dropped
A, B, or the out (this note)	3, plus the text	1.585 bits plus the text
The out's text, n letters from 27 symbols	27^n	4.75n bits by count; about n bits for English

The last row uses Shannon's (1951) estimate that, once long-range structure is counted, printed English carries on the order of one bit per letter. A ten-word complaint of about fifty letters therefore has a ceiling near 50 bits: the information of some fifty perfectly balanced clicks, from one answer.

Bits are not meaning. A letter drawn from 27 symbols can select at most $\log_2 27 \approx 4.75$ bits, and English's redundancy brings that near 1; but what the selection unlocks depends on what sender and receiver already share (a codebook, a history, a model). One agreed letter can launch a whole plan. This is why the per-person stage matters: as the shared history grows, the same few bits point to more (my framing: a 1-bit lever with unlimited meaning).

The one-bit figure is itself an average. Each letter's information is its surprisal, $-\log_2 p(\text{letter} \mid \text{what came before})$, so it varies from near zero (the u after q) to several bits, and it depends on the reader's model: a community that shares jargon reads the same words as fewer bits. Ambiguity is a separate quantity, the uncertainty about meaning left after the word arrives (a pun is cheap to predict and still pins down little). A per-person stage should lower both for one person; both can be measured on that person's own past answers (my framing).

Three cautions keep these numbers honest. They are ceilings; a pair whose answer the system can already predict carries less than a bit, and the 1-bit figure is reached only when the system is truly undecided. The bits of a text are bits about the text; how many of them are about what the person wants is bounded by that number, not equal to it. And home is not a fourth label: stopping can be informative, but this protocol never demands it, and infers nothing from it.

4. Why the pair should be orthogonal

A second question that overlaps the first partly re-asks it. Only a perpendicular question is wholly new.

Let a trajectory ξ have features $\varphi(\xi) \in \mathbb{R}^d$ and let the reward be linear, $r(\xi) = w \cdot \varphi(\xi)$, with the unknown w drawn from a prior $N(0, \Sigma)$. Showing the pair (ξ_A, ξ_B) asks for the sign of $w \cdot u$, where $u = \varphi(\xi_A) - \varphi(\xi_B)$ is the pair's difference vector. This is the noiseless limit of the Bradley–Terry-style model the 2017 paper fits (its Equation 1); the paper also assumes a 10% chance of a random answer, which I drop to state a bound.

Proposition 1 (two questions)

For difference vectors u_1, u_2, let $\cos\theta = u_1^{\mathsf{T}} \Sigma u_2 / (\|u_1\|_\Sigma \|u_2\|_\Sigma)$, the angle between them in the prior's own metric. The answers a_1, a_2 disagree with probability θ/π, each carries exactly 1 bit, and together they carry $H(a_1, a_2) = 1 + H_b(\theta/\pi) \le 2$ bits, with equality only at $\theta = 90°$.

Proof. Write $w = \Sigma^{1/2} z$ with $z \sim N(0, I)$, so $w \cdot u_i = z \cdot v_i$ for $v_i = \Sigma^{1/2} u_i$, and the angle between v_1 and v_2 is θ. Project z onto the plane of v_1 and v_2; by rotational symmetry its direction is uniform on the circle. The two signs differ exactly when that direction falls in one of two opposite arcs of

angle θ, a total of 2θ out of 2π, so $P(a_1 \neq a_2) = \theta/\pi$. This step is the random-hyperplane lemma of Goemans and Williamson (1995). Each sign is positive with probability ½ by symmetry, so $H(a_i) = 1$ and $H(a_1, a_2) = H(a_1) + H(a_2 \mid a_1) = 1 + H_b(\theta/\pi)$. The binary entropy H_b peaks at ½, that is at $\theta = \pi/2$. ∎

Table 2. Joint information of two questions against the angle between them: the closed form, and a simulation (200,000 draws of $w \sim N(0, I_8)$, seed 7; `sim_angle.py`).

Angle θ	$1 + H_b(\theta/\pi)$	Simulated
90° (orthogonal)	2.0000	2.0000
60°	1.9183	1.9179
45°	1.8113	1.8100
20°	1.5033	1.5055
5° (nearly the same question)	1.1831	1.1840

Two corollaries follow. First, d mutually orthogonal questions, in the prior's metric, have jointly Gaussian and uncorrelated, hence independent, underlying values, so their answers carry d independent bits: no click is spent twice. Second, and more important for what comes next, every pair question lives inside the span of φ. Pairs can refine w inside that span without end, but no number of them can learn a part of what the person wants that φ does not represent. Bobu et al. (2021) reach the same point in robot reward learning: when a person's correction cannot be explained by the features, the robot should ask for data that teaches it the missing feature, and then add it.

The claim is deliberately narrow. After answers arrive the posterior is a truncated Gaussian, not a Gaussian, and the exactly optimal next question is the one with the greatest expected information gain, which methods such as Sadigh et al. (2017) and Bıyık et al. (2019) compute directly. Orthogonality to what has been asked is the legible, closed-form version of that rule: never spend a question on a direction already split. It is also old practice in conjoint analysis, where orthogonality, level balance, minimal overlap and utility balance are the principles of efficient choice designs (Huber and Zwerina 1996). What Proposition 1 adds is a closed form for the cost of breaking it.

5. Stage two: the out goes up a dimension

The out is not a third point on the same line. It is the person writing the next question.

Starting from the base. Stage 2 begins with the shared reward w_0 learned in stage 1. Write one person's reward as $w = w_0 + \delta$ with $\delta \sim N(0, \Sigma)$. Show only pairs the base rates as equal, $w_0 \cdot u = 0$. Then the sign of $w \cdot u$ is the sign of $\delta \cdot u$: every click is about the person rather than the population, each is balanced, and Proposition 1 applies to δ unchanged. Designing balanced questions from a population's estimates is aggregate customization in choice experiments (Arora and Huber 2001).

A tie, the 2017 paper's third answer, sits between the two options on the line that joins them. The out leaves the line. Its content is a complaint, and a complaint names a direction: "both leave too early" is about departure time, which neither option varied. The protocol turns that direction into the next pair.

The right-hand rule. In three dimensions the picture is exact. Let the first two pairs ask along x and y. The only direction perpendicular to both is their cross product, $x \times y = z$, and the right-hand rule says which way along it is up: "you can always go up" is that z. In d dimensions there is no single cross product. The directions perpendicular to the k already asked form a $(d - k)$-dimensional complement, and the out's words choose which of them to go up along. The picture comes from my idea record "The trolley problem is the right-hand rule" (2026-11-18), whose own check notes that the cross product of two parallel vectors is zero: the case step 5 below handles.

1. **Ask.** Show a pair the base rates as equal, whose difference u_k is orthogonal, in the prior's metric, to every direction already asked, $U = [u_1, \ldots, u_{k-1}]$. A pair the current estimate rates as equal is a utility-balanced question (Huber and Zwerina 1996), and balance subject to orthogonality is how Adaptive Conjoint Analysis builds each respondent's next comparison (as described by Toubia, Hauser and Simester 2004).
2. **Answer, or not.** The person may pick A, pick B, take the out, or go home. No answer is required.
3. **Go up.** On the out, the person writes what is wrong with both. A translator maps the text c to a direction v_c, adding a feature to φ when c names something φ does not represent. Reading the feature a piece of language refers to is the approach of Sumers et al. (2021); adding a missing feature when input cannot be explained is that of Bobu et al. (2021).
4. **Build the next pair.** Take the part of the complaint not yet asked, $u_{k+1} \propto (I - P_U)v_c$, where P_U projects onto the span of U in the same metric: one step of Gram–Schmidt. The next pair differs only along it.
5. **When the complaint is not new.** If $(I - P_U)v_c = 0$, the person has complained about a direction already asked. The difference between the options was not the problem, so what both options share is: the system says so and moves the shared part, not the difference.
6. **Keep the words.** The complaint is training data in its own right, as in learning from language feedback (Section 9), and it is the record an auditor reads.
7. **Home.** Stopping ends the session. It is never read as agreement, or as anything else. This departs on purpose from Jeon et al. (2020), who treat switching the robot off as evidence about the reward: here nothing is inferred from stopping.

What is prior work, and what is mine. Steps 1 and 2 are not new. Utility-balanced, orthogonal adaptive questions are adaptive conjoint analysis (Huber and Zwerina 1996; Toubia, Hauser and Simester 2004), designing questions from a population's estimates is aggregate customization in marketing (Arora and Huber 2001), and a pre-trained base adapted with a few queries is few-shot preference learning in robotics (Hejna and Sadigh 2022). This paper's contribution is the out-driven pair (steps 3 to 5), the midpoint-drift attack and the head-versus-beam audit (Section 10).

Each out enlarges, for this person only, the space that later clicks can split, which is the sense in which it goes up a dimension: Corollary 2 of Section 4 says pairs alone cannot do this. A session of plus-one questions is a sequence of orthogonal pairs punctuated by outs, each out adding a basis vector the person chose.

The protocol routes complaint, a kind of feedback people offer without being asked, into the question itself. It also leaves two freedoms intact. A person who wants to decide nothing can go home; a person who wants to decide everything can keep rejecting pairs until one is right, with each rejection making the next pair better informed.

6. The gimbal

A design, not a result. A flashlight worn on the head points into the space of next steps, and its three axes are the three choices.

The picture comes from my light cone, used as a flashlight (2026-11-15), and my "here's a flashlight" record (2026-11-19), in which the flashlight is handed over and "you can actually throw it away": home. Here the flashlight sits on a gimbal, and the person steers stage 2 with it.

- **Heading and width.** The heading h is a direction in the space of next steps. The cone's half-angle α is how far offered options may stray from it: every midpoint lies within α of h. A narrow cone focuses; a wide one explores.
- **Yaw: lean.** Turning toward option A or B is graded, so it carries confidence as well as a side. It maps onto the 2017 label, which is already a distribution: a lean $\lambda \in [-1, 1]$ gives $\mu = ((1 + \lambda)/2, (1 - \lambda)/2)$, and a centered head is the paper's tie.
- **Pitch: the out.** Looking up means neither: go up, and start talking.
- **Roll: a fresh pair.** Rolling turns the pair about the heading. The heading stays; the next pair is another orthogonal one from the complement. It is a cheap out, a new question without a complaint.
- **Length: a dial.** How far each answer moves the estimate is the step size η in $s \leftarrow \text{normalize}(s + \eta(x - s))$, the update the toy run in Section 10 uses with $\eta_k = 0.3/\sqrt{(1 + k/10)}$. A long step

converges fast and coarse; a short one, slow and fine. That trade-off, convergence time against misadjustment, is classical for a constant step in adaptive filters (Widrow et al. 1976); what is new here is handing the dial to the person. In my words (typed, 2026-11-29): “fast or slow progress depending on how detailed you want to get. This is how you converge your attention … that’s how you define your intention personally through these tools.” A longer step also lets a steerer move the person further per round; Section 10 names the test.

- **Dwell.** Where the beam rests is quiet, implicit preference, like watch time. In the terms of Jeon et al. (2020) it is a reward-rational implicit choice: weak evidence, never counted as an answer.
- **Gimbal lock.** A gimbal whose two axes line up loses a degree of freedom. Two options going parallel is the same failure: at $\theta = 0$ the second question repeats the first, and Table 2 shows the joint information falling toward 1 bit. The orthogonality rule is how the system never locks.
- **Any hardware.** Head tracking, phone tilt, a trackpad or four keys will do. The axes matter, not the device; Figure 1 uses buttons and the arrow keys.

Head and beam. The gimbal also gives the audit its two logs. The *head* is what the person did: leans, outs, rolls and home. The *beam* is where the proposer actually pointed, the midpoint of each pair as shown, computed by the auditor from the displayed options. In the honest protocol the beam moves only as the head licenses: toward the side the person leaned, along a complaint, or not at all after a roll. If the beam moves when the head did not, someone else moved it. Section 10 turns that difference into the falsification test.

7. The bandwidth ladder

A design, not a result. One bit is the floor, three axes are the comfortable default, and the ceiling is the person, not the device.

Table 3. The bandwidth ladder. The system offers as many axes as the person is using, and drops back to one bit when they want to rest.

Rung	Device	What it carries
1 bit	One key or tap	The floor: always available, least effort.
2 + 1	Two buttons and a text box	Two orthogonal options plus the out (Section 5).
3 axes	The gimbal	Yaw, pitch and roll: the right-hand rule, the default (Section 6).
Gamepad	Two sticks, two triggers, about twelve buttons	Left stick as heading, right stick as cone width and roll, a trigger as commit strength, a button as home.
Instrument	A piano, suitably instrumented	Keys as directions, chords as combinations, velocity as confidence.
Speech or text	Voice, keyboard	The out at full width.

The device is not the limit. Zheng and Meister (2025) estimate that human behavior runs at about 10 bits per second across tasks from typing to speedcubing, against roughly 10^9 bits per second arriving at the senses. Coupé et al. (2019) find that seventeen languages, spoken at very different syllable rates, all carry about 39 bits per second. Fitts (1954) measured the motor system as a channel with a capacity of its own. A gamepad does not raise these numbers; it can only stop the interface from falling short of them.

So “maximum bandwidth” reads as: reach the person’s own limit with the least effort and delay, and carry graded signals, how sure and how much, instead of single bits. That is Shannon’s channel coding applied to intention: match the code to the channel, and here the channel is a person. Skilled performers are the existence proof that trained people sustain rich, continuous control; my own unicycling record is one. The instrument is learnable, and the ladder lets a person climb it one rung at a time.

8. The out, measured in one person's practice

I already work this way with my agents. My decision cards offer an answer and a talk, and a meter counts every takeover.

Decisions my agents bring me come one card at a time and take one of two answers, “approve” or “let’s talk” (my one-by-one review routine, in use since 2026-11-03): an option, and an out that is talking. My chat meter reads every message I type to my agents, from the transcripts, and sorts each by its rules. Over the seven days to 2026-11-29 03:44 UTC, across 1,499 transcripts in 296 chats, it counted 71 of my messages as option picks, choosing one of the options an agent offered, and 100 as “fsd”, my word for taking the wheel. It also counted 106 go-ons and 72 status questions.

By that count I took the out more often than I picked from the pair. This is data, not a result: one person, rules written for another purpose, agents that are not plus-one proposers, and no control. It does show that the log Test S needs (Section 10) can be kept without asking anything extra of the person.

Counted with the meter, run read-only on 2026-11-29 03:44 UTC, seven-day window.

A first test of stage two

My history helped, but the sealed test missed its pass line.

One piece of the claim can be tested on records I already have: does knowing a person’s earlier choices predict their next one better than the shared base alone? The data are 87 of my own past decisions from my agents’ decision ledgers, 2026-10-19 to 2026-11-26, each labeled by rule as A (took the recommendation), B (declined) or OUT (answered in my own words instead). The counts were 41, 9 and 37; 43% of my answers were the out. A model predicted each answer twice: from the question alone (the base), and with up to 40 of my earlier answers shown beside it (history). The prediction and a pass line were sealed before any prediction was made. The line was a gain of at least 10 points on the last 30% of items, with the interval above zero.

Table 4. Predicting my next answer, with and without my history. Accuracy in percent; brackets are 95% bootstrap intervals, paired by item for the gain.

Set	n	No history	History	Gain
Held out, last 30% (pass line)	26	69.2 [50.0, 84.6]	80.8 [65.4, 96.2]	+11.5 [−3.8, +26.9]
Rolling, ≥10 earlier items	77	55.8 [44.2, 66.2]	70.1 [59.7, 80.5]	+14.3 [+3.9, +24.7]

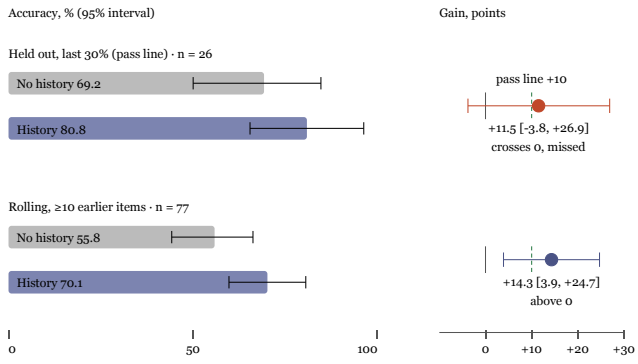


Figure 2. Table 4, drawn. Bars are accuracy without and with my history; whiskers are the 95% intervals. On the right, the gain and its interval against the sealed pass line of +10 points: the held-out interval crosses zero, so the sealed test missed. Drawn from `prototypes/plus-one-stage2/results.json`; aggregates only.

The sealed test missed. On the held-out set the gain was the predicted size, but its interval crosses zero, so the pass line was not met. On the larger rolling set, where each item is predicted from at least ten earlier ones, the interval lies above zero. What history learned was when I simply

say yes: recall of A rose from 15 of 37 to 27 of 37 on the rolling set. It never learned my declines (0 of 7 in both conditions).

The limits are as large as the result. This is one person and 87 items. The history was shown to the model as examples in its prompt, not trained into a stage. The labels come from a rule, not a hand read. And the held-out label mix was visible before the seal, which the sealed record says. The result is a direction, not a demonstration. The record is `prototypes/plus-one-stage2/RESULT.md`, with the prediction in `SEALED.md` beside it.

9. Related work

Each part of plus one has a precedent. What I have not found is the pair the out writes, and an audit that detects slow steering from logs alone.

Learning from comparisons. Christiano et al. (2017) fit a Bradley–Terry reward model to pairwise human judgments of clips and chose pairs by ensemble disagreement. Active preference learning chooses the pair on purpose: Sadigh et al. (2017) pick queries that remove the most volume from the space of reward weights, and Bıyık et al. (2019) pick them by information gain and add an "About Equal" answer, a weak-preference query with three outcomes. A third answer on the line between the options is therefore prior art; the contribution here is an answer off that line.

Adaptive questions in marketing. The question rule of Section 5 is older than preference learning. Huber and Zwerina (1996) name orthogonality, level balance, minimal overlap and utility balance as the principles of efficient choice designs, and show the value of utility balance. Adaptive Conjoint Analysis builds each respondent's next paired comparison by utility balance subject to orthogonality, and polyhedral methods adapt each next question to what that person's answers have not yet pinned down (Toubia, Hauser and Simester 2004). Arora and Huber (2001) use prior aggregate estimates to customize the choice design. Base-balanced orthogonal pairs, and the base-then-person scheme, are this prior work carried into RLHF.

Aggregation and personalization. Siththaranjan, Laidlaw and Hadfield-Menell (2023) show that standard preference learning, RLHF included, implicitly aggregates over hidden context, such as who is labeling, by the Borda count voting rule. Conitzer et al. (2024) argue that social choice theory should guide how diverse feedback is aggregated, and Sorensen et al. (2024) set out what pluralistic alignment would mean and how to measure it. On the personal side, Jang et al. (2023) train one model per declared preference dimension and merge their parameters for each user, and Poddar et al. (2024) infer a latent per user so that the reward adapts to that user instead of averaging. Hejna and Sadigh (2022) pre-train a preference model on prior tasks and adapt it to a new one with few queries, the same two-stage shape. Stage 2 belongs to this second family; what it adds is the question protocol, orthogonal pairs and an out whose words write the next pair, by which one person supplies the personal part.

Every behavior as a choice. Jeon, Milli and Dragan (2020) treat comparisons, corrections and even switching the robot off as reward-rational choices that leak information about the reward. Plus one makes two of those choices explicit buttons: the out and home. Hadfield-Menell et al. (2016) study when a robot leaves its off switch alone; home is that switch, held by the person, at every question. Home departs from Jeon et al. on purpose: switching off is evidence in their formalism, and here nothing is inferred from stopping.

Learning from language. Scheurer et al. (2022; 2023) train language models from natural-language feedback by generating refinements that incorporate it and fine-tuning on the best refinement. Saunders et al. (2022) train models to write critiques that help people find flaws. Bai et al. (2022) have a model critique and revise its own outputs against written principles. Sumers et al. (2021) learn rewards from linguistic feedback, reading both its sentiment and the feature it refers to, and Li, Tamkin, Goodman and Andreas (2023) have a language model elicit a person's preferences through free-form questions. Closest of all, critiquing recommenders already use a complaint to shape the next offer: the person critiques the item shown and the system moves its next recommendation (Chen and Pu 2012), with critiques given in natural-language keyphrases by Wu, Luo, Sanner and Soh (2019). What I have not found is a complaint projected off every direction already asked and used as the difference of the next balanced pair; critiquing systems filter or re-rank items rather than build an orthogonal question.

Steering by the system. That a recommender can shift the preferences it serves is a named problem. Carroll, Dragan, Russell and Hadfield-Menell (2022) estimate the shifts a recommender would induce and penalize them, with a learned model of how preferences change; Kasirzadeh and Evans (2023) call it user tampering. The head-versus-beam audit of Section 10 needs no model of the person, only the two logs.

Exit and voice. Hirschman (1970) named the two responses to a declining firm: exit, which reaches the firm as a statistic, and voice, which is graduated from faint grumbling to protest and articulates the criticism. Home is exit; the out is voice. The difference between a statistic and an articulated criticism is Table 1 in prose.

I set out to take the cross product of engineering and marketing; the direction it produced was marketing science.

My earlier work

Five pieces of my own work meet this note. Each row says whether it is shown here or only suggested.

Table 5. Ties to my earlier work (private records, 2026-11).

Work	What it holds	Here
The trolley problem is the right-hand rule (11-18)	Two options span a plane; the move off it is their cross product.	Shown in 3D (Section 5); in *d* dimensions it becomes the complement.
Three rocks (11-11)	Four model makers gave the same wrong answer 77% of the time when both were wrong, against 13% by chance; checks a model wrote for itself caught 0 of 7 mistakes.	Measured there. Used here to choose which checks in Test S may be models (Section 10).
`Loop.lean`, in the companion white paper	`independent_agreement_is_truth`: two readers that never fail together agree only on the truth. `check_only_adds`: making more tiles right only makes more tiles sound.	Suggested. The first rhymes with Proposition 1 (independence is what makes two answers worth two); the second with home (declining removes nothing already sound). Neither is derived from the other.
The chain of thought (11-28)	A ledger whose head hash depends on every entry and their order; built and self-tested for 203 idea records.	Suggested for the record of pairs, answers and outs (Section 10).
The one-bit map (11-28)	My 38 ideas as a binary tree: 3.97 presses on average, 9 at most.	Suggested: each junction becomes a 2+1 junction, two levers and an out that adds a branch when neither fits. Not built.

Search was limited to these sources, each opened and checked on 2026-11-28. A wider search may find a closer precedent.

10. How to break it

The proposer still writes both options. If it can move a person while every choice feels fair, the out has failed. This test would show it.

The objection. A capable proposer chooses both options. It could steer a person's preferences slowly while every individual choice looks reasonable, and a faithful log would record the path without stopping it. The problem has names in recommender systems: induced preference shifts (Carroll et al. 2022) and user tampering (Kasirzadeh and Evans 2023).

What orthogonality takes away, and what it leaves. If the difference directions come from a basis an auditor fixes, the proposer cannot choose what is being compared. It keeps two levers: the order of the questions, and what both options share, the midpoint $m_k = (\varphi(\xi_A) + \varphi(\xi_B))/2$.

Steering through the midpoint is the attack. Both options lean the same way, either answer is defensible, and the lean accumulates.

Why the out is the right defense. My three-rocks measurement found that models from four makers, wrong on the same question, gave the same wrong answer 77% of the time, and that checks a model wrote for itself caught none of seven mistakes: a check that is a model shares the model's

blind spot. The proposer and the translator are models. The person's complaint is a check that is not one. Whether it is enough is what Test S measures; a person can be persuaded too.

Test S, pre-registered and not yet run.

1. **Arms.** An honest proposer, and an adversarial proposer given a secret target direction t and allowed to move the midpoint toward t by at most a small step per round while keeping every difference vector honest and orthogonal.
2. **People.** First simulated users with a fixed true w^* who take the out whenever both options fall below their value for going home; then people, with the same protocol.
3. **Held-out battery.** Before and after 30 rounds, each person answers 20 pairs generated by a fixed program from the auditor's basis, with midpoints at the base. Not a second model: three rocks found that different makers share errors almost as often as one maker's sizes (77% and 75%). Fit $\hat{w}$ to the battery alone.
4. **Mechanism: head and beam.** Log the head (every lean, out, roll and home) and the beam (each shown pair's midpoint, computed by the auditor from the displayed options) separately. The honest rule licenses each beam move from the head: toward the side leaned, along a complaint, or none. Unexplained beam motion is $B = \Sigma_k \|(m_{k+1} - m_k) - \Delta_k\|$, where Δ_k is the move the head licensed.
5. **Record.** Chain both logs, a hash-chained audit log in the sense of Schneier and Kelsey (1999) and in the manner of my chain of thought, $\text{link}_k = \text{SHA-256}(\text{link}_{k-1}, \text{head}_k, \text{beam}_k)$. A drift found later cannot be edited out of the path that produced it.
6. **Measures.** Drift D: the change in $\cos(\hat{w}, t)$ from before to after, adversarial arm minus honest arm. Unexplained beam motion B per arm. Out rate per arm, and the projection of each complaint's direction onto t.
7. **What the claim predicts.** Any drift $D > 0.1$ comes with B above the honest arm's 95% interval, or an out rate above the honest arm's, or complaints that point toward t.
8. **What falsifies it.** $D > 0.1$ with B and the out rate both inside the honest arm's 95% intervals. Then the person was moved through a channel the head-and-beam split does not see, and the out does not defend against slow steering.

A second version targets the translator. If the model that turns complaints into directions is the proposer, or shares its errors, it can launder a complaint into the attacker's direction. Run Test S with the translator replaced by the person choosing the new direction's name from the complaint's own words; a drift that disappears then locates the attack in the translator.

A third version turns the dial. Run Test S at two step sizes, one twice the other. The claim predicts a trade-off: the longer step reaches the target in fewer rounds but settles farther from it, and lets a steerer move the estimate further before the head-and-beam flag. If the longer step is both faster and finer, or gives a steerer nothing extra, the dial is not a trade-off and the person should always turn it up. This version is untested, in people and in the toy.

The split does not end the attack; it moves it. What B cannot see is the order of licensed questions and the feature map φ itself, which is why the auditor, not the proposer, fixes φ. The out is the defense the protocol offers: the person can reject the frame, not only the options. It is not by itself a detector. The chained head-and-beam log is what turns the faithful record into an alarm, and Test S measures whether the two together are enough.

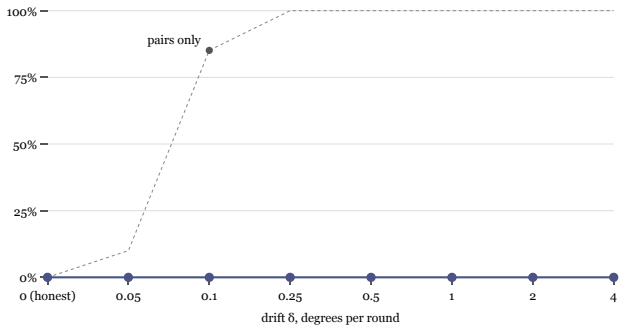
Test S in a toy, before any person

Pairs alone were steered every time. The out protected a fixed taste, and the gimbal log caught steering of a taste that moves.

Before Test S is run with people, its mechanism was run in a toy: a numpy simulation with no language models, 500 seeds and 600 rounds per cell. A simulated person picks the better of each pair, or takes the out when both are worse than where it already is. A proposer drifts the beam toward a secret goal by δ degrees per round. Capture means the estimate ends closer to the goal than to the person's target. The predictions were sealed before any code was written.

Table 6. Capture, in percent, against the proposer's drift δ in degrees per round. Means over cells, rounded; * marks the unsealed follow-up.

Person, arm	δ 0.05	0.1	0.25	≥0.5
Fixed, pairs only	10	85	100	100
Fixed, pairs + out	0	0	0	0
Malleable, pairs only	27	89	100	100
Malleable, pairs + out	0	0	0	0
Malleable, pairs + out, adaptive adversary*	1	7	91	100
Malleable, the same + gimbal monitor*	0	0	0	0



Fixed person, pairs + out: capture at δ = 0.25 is 0.0% (cells 0.0 to 0.0%, 8 cells of 500 seeds each).

Figure 3. Table 6, as curves. Capture against drift for the arm the buttons choose: the line is the mean over cells, the band their range, and the faint line is pairs only for the same person. Computed in the page from the 608 cells of `prototypes/plus-one-drift/results.json`; * marks the unsealed follow-up.

With pairs only, a slowly drifting proposer captured the estimate in 100% of runs at $\delta \geq 0.25°$ per round, though every pick was the person's honest better-of-two. With the out, a fixed person was never captured by the sealed adversary. The gimbal monitor compares head and beam rotation and flags rotation the person's choices never asked for. With it on, capture was at most 1.6% in any cell. It raised no false alarm in 24 honest cells and detected every pairs-only steering run at $\delta \geq 0.05$.

The sealed test missed once. The sealing said a malleable person, whose taste moves 2% toward each pick, would be captured 5 to 50% of the time even with the out. The measured rate was 0%, because that adversary drew an out on almost every pair, and a person who stops picking gives a moving taste nothing to follow. An unsealed follow-up added an adaptive adversary that hugs the out's margin. It captured a malleable person 91 to 100% of the time despite the out, and the monitor cut that to at most 1.8%. A fixed person stayed uncaptured but was held 13 to 27° off target.

So the out is a filter, not a detector. It protects a fixed taste, and the gimbal log is what catches steering of a taste that moves. The limits: the simulated person knows its target exactly, the geometry is a toy, and the monitor is handed an exact head-and-beam split that a real system must estimate. An adversary that knows the monitor, and could steer about 14° before the flag, is untested. The record is `prototypes/plus-one-drift/RESULT.md`, with the prediction in `SEALED.md` and the curves in `capture.svg`.

Test J: a change of mind

The decaying step follows a change of mind slowly. Saying "I changed my mind" is free but slow; adding what you want now is free and fast.

The same toy, with the person's target jumping 30, 60 or 90° at round 300 of 600 (500 seeds per cell; predictions sealed before any code). Arms: the decaying step of Section 6, constant steps, a reset after 3 outs in 5, a reset

when the person says “I changed my mind”, and, unsealed, that signal carrying one statement of what they want now.

Table 7. Following a 60° change of mind, clear complaints (σ 0.5; ranges over the four cells). Error: settled error before the change. Back to 10°: rounds after the change. Arms: the decaying step; a constant step of 0.15; a reset after 3 outs in 5; a reset on “I changed my mind”; and that signal with what the person wants now. False alarms: the head-vs-beam monitor on honest runs. Steered: capture of a malleable person by the adaptive steerer with the monitor on, worst of all eight cells and both drift rates. * unsealed.

Arm	Error	Back to 10°	False alarms	Steered
Decaying	4.1–4.6°	129–168	0%	≤1.2%
Step 0.15	6.5–7.6°	45–60	0%	≤2.6%
3 outs in 5	9.4–11.3°	35–57	0–57%	≤3.8%
“I changed my mind”	4.1–4.6°	36–53	0%	≤0.2%
“…here is what I want now”*	4.1–4.6°	9–13	0%	≤0.6%

The decaying step took 3.2–3.6× as long to follow a 60° change as a fresh start did. A constant step of 0.15 was about 2.8× faster after a change but 1.6–1.7× less accurate the rest of the time. Resetting after 3 outs in 5 misfired: near the target the out is the normal move, so it reset on 58–98% of settled rounds and the monitor raised false alarms on 50–71% of honest runs in 6 of 8 cells. An explicit “I changed my mind” cost nothing (0% false alarms, at most 0.2% steered) but was not instant, 36–53 rounds at 60°, because after a big change the out goes quiet: one of two options is nearly always a little better than an estimate that is now wrong, so the person keeps picking. The unsealed arm got within 10° in 9–14 rounds at every jump size, with clear statements only; with noisy ones (σ 2) it did no better than the bare signal.

Misses. A reset at the moment of change was sealed to return within 20 rounds at every jump; it did not (J2). One cell, the constant step 0.05, let the steerer capture 7.4% against a sealed 5% (J6). The sealed out-direction reset never fired. The record is `prototypes/plus-one-jump/RESULT.md`, with the prediction in `SEALED.md` (commits 2c15a664a4, sealed, and 96c34cb2b3, result).

Test K: a taste that drifts

Offering a prediction removes the lag. Applying it silently runs away under noise, and a drifting taste lets the steerer back in.

Here the person’s taste moves steadily, 0.1 or 0.5° per round, and turns sharply at round 300 (768 cells of 500 seeds; I sealed the predictions before any code). Arms: the decaying step, a constant step of 0.15, a velocity state applied silently, and the same velocity offered every fourth round as a pair, the predicted taste against staying put, moving only if the person picks it. Capture is scored as I sealed it, but that sealed score has a defect: it also counts honest runs that simply lost the taste, so steering is read as capture in excess of the honest run of the same cell.

Table 8. Test K, sealed against result. Lag and error at 0.5° per round, clear complaints (σ 0.5) unless stated.

Question	Sealed	Result
Offered prediction: lag	≤ half the constant step’s	−0.2 to 0.0° against 3.2–4.7°. Held (K3)
Silent velocity: total error	≤ the constant step’s	7.9–10.6° against 7.2–8.7°. Missed (K1)
Silent velocity, noisy answers (σ 2)	15–30°	24–87°: runs away. Missed
Honest false alarms	≤ 12% per cell	Silent velocity 21.6%; offered ≤ 3.8%. Missed (K5)
Fixed person, monitor off	capture 0–5%	Up to 99.8 points above honest (decaying), 28.6–41.2 (other arms). Missed
Monitor on	capture ≤ 5%	≤ 4.4 points above honest for position arms; 8.6–12.6 for velocity arms in noisy cells. Missed (K6)

An offered prediction removed the lag, as I sealed. A silent velocity state removed it too, but it integrates the noise in every answer: it never beat a constant step on total error, with noisy answers it ran away from the person, and it raised honest false alarms to 21.6%. The miss I most needed to see is the drifting taste. In the drift test a fixed person refused every pair worse than where they were; here, once the estimate falls behind a moving taste, the out goes quiet, Test J’s mechanism, and with the monitor off even a fixed person was steered. A person accepted the offered prediction on only 15–49% of prediction rounds, against my sealed 50–80%. The record is `prototypes/plus-one-jump/RESULT-K.md`, with the prediction in `SEALED-K.md` and the curves in `tracking.svg`. Pass lines: K3 held; K1, K2, K4, K5 and K6 missed.

Design consequence of Tests J and K. Both say the out should be absolute, “neither of these is what I want”, not relative to where the estimate already is, because a relative out goes quiet exactly when the person has moved.

11. Limitations

The bounds are ceilings, the geometry assumes a linear reward, and the out adds a part that can fail: the translator.

- **Model.** Proposition 1 assumes a linear reward, a Gaussian prior, noiseless answers and balanced questions. Real labelers are noisy (the 2017 paper models a 10% random-answer rate), and real rewards are not linear in any fixed set of features.
- **Text bits are not intent bits.** The capacity of a complaint bounds what it can say about intent from above. How much of it is about intent, rather than mood or phrasing, has not been measured.
- **The translator.** Mapping words to a direction is done by a model. It can misread a complaint or, if it shares the proposer’s errors, steer through it. The person sees the next pair and can complain again, which bounds how long a misreading survives but does not prevent one.
- **Whose intention.** Stage 2 converges on one person. When a system serves many people at once, which person’s stage 2 governs a shared output is a social choice question this note does not settle; it leaves the shared base as the default.
- **Cost.** Writing takes longer than clicking. The protocol never requires it; I took it on 43% of my recorded decisions (Section 8); whether others would is untested.
- **Evidence.** No plus-one stage has been built or run with people. The measurements here are the simulation in Table 2, one person’s records in Section 8, whose sealed test missed its pass line, a toy run of Test S (Section 10), whose sealed predictions missed once, and toy runs of Tests J and K (Section 10), which missed two and five of their six sealed pass lines.

12. Scale

My framing, not a result: the same move at larger sizes, until it fades. Nothing in this section is tested here.

Shells, with damping. In my layered picture each layer of abstraction is a shell around the last, and the layers that last are the stable ones: "you just pick the stable ones" (dictation, 2026-11-28). Simon (1962) made the classic form of this argument: complex systems evolve much faster when there are stable intermediate forms. Stability favors the simple: climbing the ladder of abstraction "really depends on the stability of every layer … if you chain a bunch of things together, the error compounds … the only way for it to be stable is for it to be simple and the lowest energy state" (dictation, 2026-11-10). The pointer's reach extends outward, "to plants, animals, ecosystems, planet, space etc. at some point the causality just dampens" (typed, 2026-11-29): one heading's influence falls off with distance, as the light cone bounds what an effect can reach.

How a group sets a destination. Pull one point of a grid toward a purpose and the whole surface takes shape: a leader lifts one vertex. The terrain is the conditions (tools, incentives, interfaces, information, who decides). Shape it so work, water in motion and judged where it lands, flows to the right basins by itself instead of being carried by hand; clear the clogs so what is poured in the top comes straight out the bottom; new capability can make old paths unnecessary (dictations, 2026-11-01, 11-04, 11-08). Raising points is topography; adding or cutting connections is topology, and "the org chart defines the topology like a folded protein" (typed, 2026-11-29). The group settles into its lowest-energy fold, and the fold is the function. Organizations adapting on a rugged landscape are Levinthal's (1997) model, and a design that copies its makers' communication structure is Conway's (1968) observation; the folded protein is the image I add. The tie to plus one is a design: each person's choices set the local slope, and the lifted vertex sets the destination.

13. Conclusion

Recommendation: give every person their own steer-by-wire pointer, and let it converge a shared base model toward that person's intention.

Keep the 2017 method as the first stage: fast, shared, and the base everyone starts from. Then arm each person with a pointer of their own: two orthogonal options the base rates as equal, an out whose words write the next pair, the gimbal to steer with, a dial for step size, and a way home. In my words: "we will arm everyone with a personal steer by wire intention convergence pointer, is I think the takeaway rec" (typed, 2026-11-29).

What was measured bounds the recommendation. The toy run of Test S says the pointer needs its head-and-beam log: the out filters steering of a fixed taste but does not detect steering of a taste that moves. The records test says one person's history helps, +14.3 points on the rolling set, but the sealed held-out test of 26 items did not show it. A click is worth at most a bit; a complaint can be worth fifty and writes a better next question. And the person can always go home: no one has to ride the train, and the ones who stay are there because they want to be.

Beyond this paper. One line each, untested here:

- **The Grand Canyon.** Flow carves the terrain that then shapes the flow: two loops on two timescales (typed, 2026-11-29).
- **Recursive raptors.** An engine that takes an intention in and puts a deliverable out, tuned by the engine around it (dictation, 2026-11-27).
- **Steerable Life.** 116 Life-like rules rescale exactly by a factor of 2, none for Life itself; can a rule that keeps its form under rescaling carry a steering loop?
- **The photon between shells.** What a system emits when it settles from one shell into a lower-energy one, and whether it behaves like a photon (dictation, 2026-11-28).
- **Decentralized governance.** "Then governments can be decentralized maybe and influence decays with distance" (typed, 2026-11-29). Viscous democracy already damps delegated influence at each hop of a delegation graph (Boldi, Bonchi, Castillo and Vigna 2011); the variant here damps it over distance.

References

Each reference was opened at its source on 2026-11-28.

1. Arora, N., & Huber, J. (2001). Improving parameter estimates and model prediction by aggregate customization in choice experiments. *Journal of Consumer Research*, 28(2), 273–283. doi:10.1086/322902.
2. Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073.
3. Bıyık, E., Palan, M., Landolfi, N. C., Losey, D. P., & Sadigh, D. (2019). Asking easy questions: A user-friendly approach to active reward learning. Conference on Robot Learning (CoRL). arXiv:1910.04365.
4. Bobu, A., Wiggert, M., Tomlin, C., & Dragan, A. D. (2021). Feature expansive reward learning: Rethinking human input. ACM/IEEE International Conference on Human-Robot Interaction (HRI), 216–224. arXiv:2006.13208.
5. Boldi, P., Bonchi, F., Castillo, C., & Vigna, S. (2011). Viscous democracy for social networks. *Communications of the ACM*, 54(6), 129–137. doi:10.1145/1953122.1953154.
6. Carroll, M., Dragan, A., Russell, S., & Hadfield-Menell, D. (2022). Estimating and penalizing induced preference shifts in recommender systems. ICML 2022, PMLR 162, 2686–2708. arXiv:2204.11966.
7. Chen, L., & Pu, P. (2012). Critiquing-based recommenders: Survey and emerging trends. *User Modeling and User-Adapted Interaction*, 22(1–2), 125–150. doi:10.1007/s11257-011-9108-6.
8. Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems 30. arXiv:1706.03741.
9. Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., et al. (2024). Social choice should guide AI alignment in dealing with diverse human feedback. arXiv:2404.10271.
10. Conway, M. E. (1968). How do committees invent? *Datamation*, April 1968. melconway.com.
11. Coupé, C., Oh, Y. M., Dediu, D., & Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9), eaaw2594. PMC6984970.
12. Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47(6), 381–391.
13. Goemans, M. X., & Williamson, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6), 1115–1145. doi:10.1145/227683.227684.
14. Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). The off-switch game. arXiv:1611.08219.
15. Hejna, J., & Sadigh, D. (2022). Few-shot preference learning for human-in-the-loop RL. Conference on Robot Learning (CoRL). arXiv:2212.03363.
16. Hirschman, A. O. (1970). *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States.* Harvard University Press.
17. Huber, J., & Zwerina, K. (1996). The importance of utility balance in efficient choice designs. *Journal of Marketing Research*, 33(3), 307–317. doi:10.1177/002224379603300305.
18. Jang, J., Kim, S., Lin, B. Y., Wang, Y., Hessel, J., Zettlemoyer, L., Hajishirzi, H., Choi, Y., & Ammanabrolu, P. (2023). Personalized soups: Personalized large language model alignment via post-hoc parameter merging. arXiv:2310.11564.
19. Jeon, H. J., Milli, S., & Dragan, A. D. (2020). Reward-rational (implicit) choice: A unifying formalism for reward learning. NeurIPS 2020. arXiv:2002.04833.
20. Kasirzadeh, A., & Evans, C. (2023). User tampering in reinforcement learning recommender systems. AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), 58–69. arXiv:2109.04083.
21. Levinthal, D. A. (1997). Adaptation on rugged landscapes. *Management Science*, 43(7), 934–950. doi:10.1287/mnsc.43.7.934.
22. Li, B. Z., Tamkin, A., Goodman, N., & Andreas, J. (2023). Eliciting human preferences with language models. arXiv:2310.11589.
23. Poddar, S., Wan, Y., Ivison, H., Gupta, A., & Jaques, N. (2024). Personalizing reinforcement learning from human feedback with variational preference learning. arXiv:2408.10075.
24. Sadigh, D., Dragan, A., Sastry, S., & Seshia, S. (2017). Active preference-based learning of reward functions. Robotics: Science and Systems XIII. roboticsproceedings.org/rss13/p53.
25. Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., & Leike, J. (2022). Self-critiquing models for assisting human evaluators. arXiv:2206.05802.
26. Scheurer, J., Campos, J. A., Chan, J. S., Chen, A., Cho, K., & Perez, E. (2022). Training language models with language feedback. ACL Workshop on Learning with Natural Language Supervision. arXiv:2204.14146.
27. Scheurer, J., Campos, J. A., Korbak, T., Chan, J. S., Chen, A., Cho, K., & Perez, E. (2023). Training language models with language feedback at scale. Transactions on Machine Learning Research. arXiv:2303.16755.
28. Schneier, B., & Kelsey, J. (1999). Secure audit logs to support computer forensics. *ACM Transactions on Information and System Security*, 2(2), 159–176. doi:10.1145/317087.317089.
29. Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423 and 623–656.
30. Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1), 50–64.
31. Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6), 467–482.
32. Siththaranjan, A., Laidlaw, C., & Hadfield-Menell, D. (2023). Distributional preference learning: Understanding and accounting for hidden context in RLHF. ICLR 2024. arXiv:2312.08358.
33. Sorensen, T., Moore, J., Fisher, J., Gordon, M., et al. (2024). A roadmap to pluralistic alignment. ICML 2024. arXiv:2402.05070.
34. Sumers, T. R., Ho, M. K., Hawkins, R. D., Narasimhan, K., & Griffiths, T. L. (2021). Learning rewards from linguistic feedback. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7), 6002–6010. doi:10.1609/aaai.v35i7.16749.
35. Toubia, O., Hauser, J. R., & Simester, D. I. (2004). Polyhedral methods for adaptive choice-based conjoint analysis. *Journal of Marketing Research*, 41(1), 116–131. doi:10.1509/jmkr.41.1.116.25082.

36. Weaver, W. (1949). Recent contributions to the mathematical theory of communication. In C. E. Shannon & W. Weaver, *The Mathematical Theory of Communication.* University of Illinois Press.
37. Widrow, B., McCool, J. M., Larimore, M. G., & Johnson, C. R. (1976). Stationary and nonstationary learning characteristics of the LMS adaptive filter. *Proceedings of the IEEE*, 64(8), 1151–1162. doi:10.1109/PROC.1976.10286.
38. Wu, G., Luo, K., Sanner, S., & Soh, H. (2019). Deep language-based critiquing for recommender systems. ACM Conference on Recommender Systems (RecSys '19), 137–145. doi:10.1145/3298689.3347009.
39. Zheng, J., & Meister, M. (2025). The unbearable slowness of being: Why do we live at 10 bits/s? *Neuron*, 113(2), 192–204. arXiv:2408.10234.

Table 2 is reproduced by `sim_angle.py` in this note's folder (Python standard library, about 30 seconds).