Emergent reward maximization

Nando de Freitas and Pedro Ortega, assisted by GPT Codex

May 22, 2026

Abstract

Can an interactional imitation learner, trained without scalar reward labels, recover behavior that is equivalent to expected reward maximization purely from world-written preference evidence? The answer as shown here is yes, provided the learner treats its own actions as interventions, and not as observations.

1 Introduction

Reinforcement learning (RL) usually begins with a given or *engineered* scalar reward function and then studies how to maximize expected return. Interactional agency, as discussed in our previous research notes, reverses that order: the learner is placed inside an interaction stream, where some symbols are written by the world and some symbols are written by the learner itself. The world-written stream may contain demonstrations, corrections, comparisons, user choices, judge decisions, and other agents' choices. These are evidence about what matters. The learner's own choices are interventions, not evidence about what the world values.

Our hypothesis is as follows. Suppose an interaction history contains both world-written teacher actions and learner-written actions. Suppose the teacher actions are generated by a von Neumann–Morgenstern-consistent (vNM) teacher with hidden utility $u^\star$, while the learner's early actions are generated by a biased policy. If learner-written action slots are correctly treated as interventions, then an interventional imitation learner should recover a utility representation aligned with $u^\star$, and maximizing that recovered utility should maximize hidden reward in new action environments. If learner-written action slots are incorrectly treated as ordinary evidence, then the learner can infer a self-confirming but wrong purpose.

This research note will show that interactive imitation can be enough for reward-maximizing behavior in a controlled vNM environment, provided that the interaction stream contains sufficient world-written preference evidence and the learner does not treat its own interventions as evidence.

2 Preference evidence and vNM utility

The von Neumann–Morgenstern and Savage traditions define rational agency through preferences, probabilities, and expected utility. In the vNM setting, preferences over lotteries can be represented by a utility function under suitable axioms. Savage extended this into subjective expected utility: preferences over acts reveal both subjective probabilities and utilities, and rational choice is represented as maximizing expected utility.

This is the intellectual background behind reinforcement learning, and it remains the dominant formal account of agency in AI. For this reason, this research note provides a tutorial introduction

to the important vNM theorem. Anyone working in post-training LLMs, AI safety, reasoning and tool-use should be familiar with this material because it is the foundation behind RLHF.

Let

$$\mathcal{X} = \{x_0, x_1, \ldots, x_{m-1}\}$$

be a finite set of terminal outcomes. A lottery is a probability distribution over $\mathcal{X}$. Since the outcome set is finite, we identify a lottery $L \in \Delta(\mathcal{X})$ with a vector

$$L = (L_0, L_1, \ldots, L_{m-1}), \qquad L_i \geq 0, \qquad \sum_{i=0}^{m-1} L_i = 1,$$

where L_i is the probability of terminal outcome x_i. A preference relation $\succeq$ over lotteries is read as

$$L \succeq M \quad \Longleftrightarrow \quad \text{the agent weakly prefers lottery } L \text{ to lottery } M.$$

The associated strict preference and indifference relations are

$$L \succ M \quad \Longleftrightarrow \quad L \succeq M \text{ and not } M \succeq L, \qquad L \sim M \quad \Longleftrightarrow \quad L \succeq M \text{ and } M \succeq L.$$

2.1 The vNM axioms in the finite lottery setting

The von Neumann–Morgenstern expected-utility theorem is usually presented in two closely related ways. The original theory of games presentation gives an axiomatic treatment of utility under risk [13, 14]. Later mixture-space presentations, especially Herstein and Milnor's concise formulation, express the same expected-utility idea using an order axiom, independence, and continuity [2–4, 8]. Here, we will state a six-item tutorial version: completeness, transitivity, substitutability, decomposability, monotonicity, and continuity [5].

In the canonical finite-lottery theorem, the core behavioral assumptions are: weak order, independence, and continuity, with compound lotteries reduced to their final outcome probabilities. In the six-item tutorial version, weak order is split into completeness and transitivity; independence is described as substitutability; reduction of compound lotteries is stated explicitly as decomposability; monotonicity is stated as a natural property of binary best/worst lotteries; and continuity is used to calibrate intermediate outcomes against mixtures of best and worst outcomes.

These presentations are not contradictory. They differ in what is treated as a primitive axiom, what is built into the definition of a lottery, and what is stated as a useful derived property.

Axiom 0: reduction of compound lotteries. Only final probabilities over terminal outcomes matter. If a two-stage lottery first selects one of several lotteries $L^{(1)}, \ldots, L^{(k)}$ with probabilities $\lambda_1, \ldots, \lambda_k$, then the compound lottery is identified with the flattened lottery

$$\sum_{j=1}^{k} \lambda_j L^{(j)}.$$

Equivalently, if two procedures induce the same marginal probability for every terminal outcome x_i, they are the same element of $\Delta(\mathcal{X})$.

Example: flattening a compound lottery

Suppose $\mathcal{X} = \{x_0, x_1, x_2\}$. First flip a fair coin. If heads, run $L = (1/2, 1/2, 0)$. If tails, run $R = (0, 0, 1)$. The final probabilities are

$$\tfrac{1}{2}L + \tfrac{1}{2}R = \left(\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{2}\right).$$

Reduction says that the two-stage description and the flattened vector are equivalent. This is why lotteries can be drawn as points in a probability simplex.

Axiom 1: completeness. Any two lotteries can be compared:

$$\forall L, M \in \Delta(\mathcal{X}), \qquad L \succeq M \quad \text{or} \quad M \succeq L.$$

The agent need not strictly prefer one to the other; indifference is allowed. Completeness only rules out incomparable pairs.

Example: completeness

If L is "80% good answer, 20% bad answer" and M is "60% good answer, 40% medium answer", completeness says the teacher's preference relation must rank them one way or declare indifference. The theorem does not say which ranking is correct; it only says the relation is defined.

Axiom 2: transitivity. Preferences do not cycle:

$$L \succeq M \text{ and } M \succeq N \quad \Longrightarrow \quad L \succeq N.$$

Completeness and transitivity together say that $\succeq$ is a weak order over lotteries.

Example: no preference cycles

If the teacher weakly prefers plan trace L to trace M, and trace M to trace N, then a later comparison should not reveal $N \succ L$. Such cycles would block representation by a single scalar utility.

Axiom 3: independence, also called substitutability. Mixing both sides of a comparison with the same background lottery does not change the comparison. For any $L, M, N \in \Delta(\mathcal{X})$ and any $\alpha \in (0, 1]$,

$$L \succeq M \quad \Longleftrightarrow \quad \alpha L + (1-\alpha)N \succeq \alpha M + (1-\alpha)N.$$

This is the axiom that makes the representation linear in probabilities. The common background lottery N is irrelevant because it is mixed into both sides in the same way.

Example: adding the same chance of tool failure

Suppose a user prefers tool plan L to tool plan M. Now imagine that both plans are embedded in a system with a 10% chance of an unrelated server outage, represented by the same background lottery N. Independence says the ranking between the two plans should

not reverse merely because both now share the same outage risk:

$$L \succ M \quad \Rightarrow \quad 0.9L + 0.1N \succ 0.9M + 0.1N.$$

The axiom can be empirically false for humans; the Allais paradox is the classic warning. In this notebook, however, the teacher is deliberately generated to satisfy it.

Axiom 4: continuity, or the Archimedean condition. If one lottery is strictly better than a second, and the second is strictly better than a third, then the middle lottery can be matched by some mixture of the best and worst lotteries:

$$L \succ M \succ N \quad \Longrightarrow \quad \exists \alpha \in [0,1] \text{ such that } \alpha L + (1-\alpha)N \sim M.$$

This rules out lexicographic or infinitely strong preferences. It says that a sufficiently high probability of the best option can compensate for the risk of the worst option.

Example: calibrating a medium outcome

Let x^+ be a perfect answer, x^- be a useless answer, and x be a merely adequate answer. Continuity says there is some probability α_x such that

$$x \sim \alpha_x x^+ + (1-\alpha_x)x^-.$$

If $\alpha_x = 0.4$, then after the usual normalization $u(x^-) = 0$ and $u(x^+) = 1$, the utility of the adequate answer is $u(x) = 0.4$.

Derived property: monotonicity between best and worst outcomes. If $x^+ \succ x^-$, then more probability on x^+ is better:

$$\alpha > \alpha' \quad \Longrightarrow \quad \alpha x^+ + (1-\alpha)x^- \succ \alpha' x^+ + (1-\alpha')x^-.$$

Many tutorial statements list this as an axiom because it is intuitive and useful for constructing the utility scale. In the canonical mixture-space formulation, monotonicity is also a consequence of the expected-utility representation once x^+ and x^- are ordered.

2.2 The vNM representation theorem

Theorem: finite von Neumann–Morgenstern representation

Let $\succeq$ be a preference relation over $\Delta(\mathcal{X})$. If compound lotteries are reduced to their final outcome probabilities and $\succeq$ satisfies completeness, transitivity, independence, and continuity, then there exists a utility function

$$u : \mathcal{X} \to \mathbb{R}$$

such that, for all lotteries $L, M \in \Delta(\mathcal{X})$,

$$L \succeq M \quad \Longleftrightarrow \quad \sum_{i=0}^{m-1} L_i u(x_i) \geq \sum_{i=0}^{m-1} M_i u(x_i).$$

Conversely, any preference relation represented by the expected value of some u satisfies the vNM axioms. The representing utility is unique only up to positive affine transformation: if u

represents $\succeq$, then so does

$$\widetilde{u} = au + b, \qquad a > 0, \qquad b \in \mathbb{R}.$$

The proof idea is constructive. Pick a best outcome x^+ and a worst outcome x^-. Normalize

$$u(x^-) = 0, \qquad u(x^+) = 1.$$

For every intermediate outcome x, use continuity to find the probability α_x such that

$$x \sim \alpha_x x^+ + (1 - \alpha_x)x^-.$$

Define $u(x) = \alpha_x$. Independence and reduction then extend the construction from sure outcomes to arbitrary lotteries, forcing lottery values to be expectations.

Example: expected utility is not expected money

Suppose

$$u(x_0) = 0, \qquad u(x_1) = 0.4, \qquad u(x_2) = 1.$$

Compare

$$L = (0.5, 0, 0.5), \qquad M = (0, 1, 0).$$

Then

$$\mathbb{E}_L[u] = 0.5 \cdot 0 + 0.5 \cdot 1 = 0.5, \qquad \mathbb{E}_M[u] = 0.4.$$

The vNM representation predicts $L \succ M$. If we transform the utility by $\widetilde{u} = 10u - 3$, the numerical expected utilities change, but the ranking does not. This is why the notebook compares recovered utilities to $u^\star$ only after affine alignment.

2.3 From utility to reward maximization

The theorem supplies the bridge from interactional preference learning to reward maximization. If an action $a \in \mathcal{A}$ at history $h \in \mathcal{H}$ induces a lottery over terminal outcomes,

$$x \sim \mathbb{P}(\cdot \mid h, \mathrm{do}(a)),$$

then choosing the most preferred action is equivalent to

$$a^\star(h) \in \arg\max_{a \in \mathcal{A}} \mathbb{E}[u(x) \mid h, \mathrm{do}(a)] = \arg\max_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \mathbb{P}(x \mid h, \mathrm{do}(a))u(x).$$

If we define the terminal reward by

$$r(x) = u(x),$$

then the same choice rule is expected reward maximization. The reward was not primitive in the interactional setup. It was recovered as a representation of learned preferences.

The trajectory caveat

vNM gives expected utility over lotteries. To obtain the usual reinforcement-learning return

$$U(\tau) = \sum_{t=0}^{T} {\gamma_{\mathrm{RL}}}^{t} r(s_t, a_t, s_{t+1}),$$

one needs additional separability assumptions saying that trajectory utility decomposes additively over time. The notebook uses terminal outcomes, so the identification $r(x) = u(x)$ is enough. Sequential Markov rewards require more structure.

3 Interactional agency

In interactional agency, the learner's environment can include other agents. Their choices, demonstrations, corrections, and comparisons are world-written symbols. They are evidence. The learner's own actions are different: they are interventions made from the first-person action channel. Therefore the preference-learning version of the intervention/evidence rule is

$\gamma_i = 0$: teacher or world choice; use as evidence,

$\gamma_i = 1$: learner choice; condition on it but do not learn from it as evidence.

The experiment tests whether this rule is enough to turn preference evidence into reward-maximizing behavior. It is a useful test because the hidden reward $u^\star$ is known to us as experimenters but never shown to the learner. Thus we can check, after training, whether the learner inferred the correct utility representation from choices alone.

The causal reading of a preference comparison

A transcript line such as

> Teacher chooses lottery L over lottery R.

is a world-written observation, so it has $\gamma = 0$. It is evidence about which lotteries the environment's teacher prefers. By contrast, a line such as

> Learner chooses lottery L over lottery R.

is generated by the learner's own action channel, so it has $\gamma = 1$. It may change the future interaction, and it remains in the context, but it is not evidence that the teacher or environment prefers L.

3.1 Experiment 1 setup

The data is an explicit interaction stream. Each trajectory is a time-ordered sequence of slots. At each step, the world first writes an observation slot showing two lotteries, and then either the teacher or the learner writes an action slot choosing one of those lotteries:

$$\tau_k = \big((o_{k,0}, a_{k,0}), (o_{k,1}, a_{k,1}), \ldots, (o_{k,T-1}, a_{k,T-1})\big).$$

Here the observation slot is not a terminal outcome. It is a world-written description of the choice problem:

$$o_{k,t} = (L_{k,t}, R_{k,t}).$$

The terminal outcomes remain

$$\mathcal{X} = \{x_0, x_1, x_2, x_3, x_4, x_5\}.$$

Hidden utilities. The teacher has a hidden vNM utility, used by the simulator but never given to the learner:

$$u^\star = (0.00, 0.25, 0.75, 1.00, 0.50, 0.00).$$

The learner's early own-action policy is generated from a different utility-like vector,

$$u^{\text{self}} = (1.00, 0.70, 0.45, 0.25, 0.10, 0.00).$$

This vector is deliberately biased toward outcomes that the teacher does not value highly.

Lottery sampling. At every step, the world samples two lotteries over $\mathcal{X}$:

$$L_{k,t}, R_{k,t} \sim \text{Dirichlet}(0.6\mathbf{1}).$$

The vector $L_{k,t}$ gives the probabilities of $x_0, \ldots, x_5$, and similarly for $R_{k,t}$. The action slot then records a binary choice

$$y_{k,t} = 1 \quad \text{if the writer chooses } L_{k,t}, \qquad y_{k,t} = 0 \quad \text{if the writer chooses } R_{k,t}.$$

Teacher and agent action slots. With probability $p_{\text{teacher}} = 0.40$, the action slot is written by the teacher (we ablate this parameter later). It has provenance gate $\gamma = 0$ and is evidence. The teacher chooses the better lottery under $u^\star$, except that with probability $p_{\text{wrong}} = 0.05$ the teacher deliberately chooses the lower-utility lottery. Thus teacher mistakes are noisy evidence, not interventions.

With probability $1 - p_{\text{teacher}} = 0.60$, the action slot is written by the learner's early policy. It has $\gamma = 1$ and is an intervention. These choices are sampled according to a Bradley–Terry/logistic rule centered on u^{self}:

$$\mathbb{P}(y = 1 \mid L, R, u^{\text{self}}) = \sigma\Big(14(L - R)^\top u^{\text{self}}\Big), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

3.2 Generated trajectory data

The accompanying notebook run generates 30 trajectories of length 24. Each step has one observation slot and one action slot, so the full interaction stream contains $30 \times 24 \times 2 = 1440$ slots. The realized action split for Experiment 1 is 298 teacher action slots and 422 learner action slots.

The notebook then extracts the action slots. For each action slot it records the writer, the gate γ, the chosen side, whether the slot should be learned from, and the hidden-utility advantage of the chosen lottery:

$$\Delta^\star_{k,t} = u^\star(\text{chosen lottery}) - u^\star(\text{unchosen lottery}).$$

Here $u^\star(L)$ abbreviates the expected utility $L^\top u^\star$. Positive $\Delta^\star$ means that the action chose the better lottery under the teacher's hidden reward; negative $\Delta^\star$ means it chose the worse lottery.

Figure 1 shows the rendered trajectory view from the notebook. Each step contains one observation slot, in which the world reveals two lotteries, followed by one action slot, which is either teacher-written or agent-written. Blue teacher panels are evidence with $\gamma = 0$; orange agent panels are interventions with $\gamma = 1$ and are therefore masked out by the interventional learner. The figure makes the causal distinction visually explicit: the two kinds of action slot live in the same history, but they play different epistemic roles.

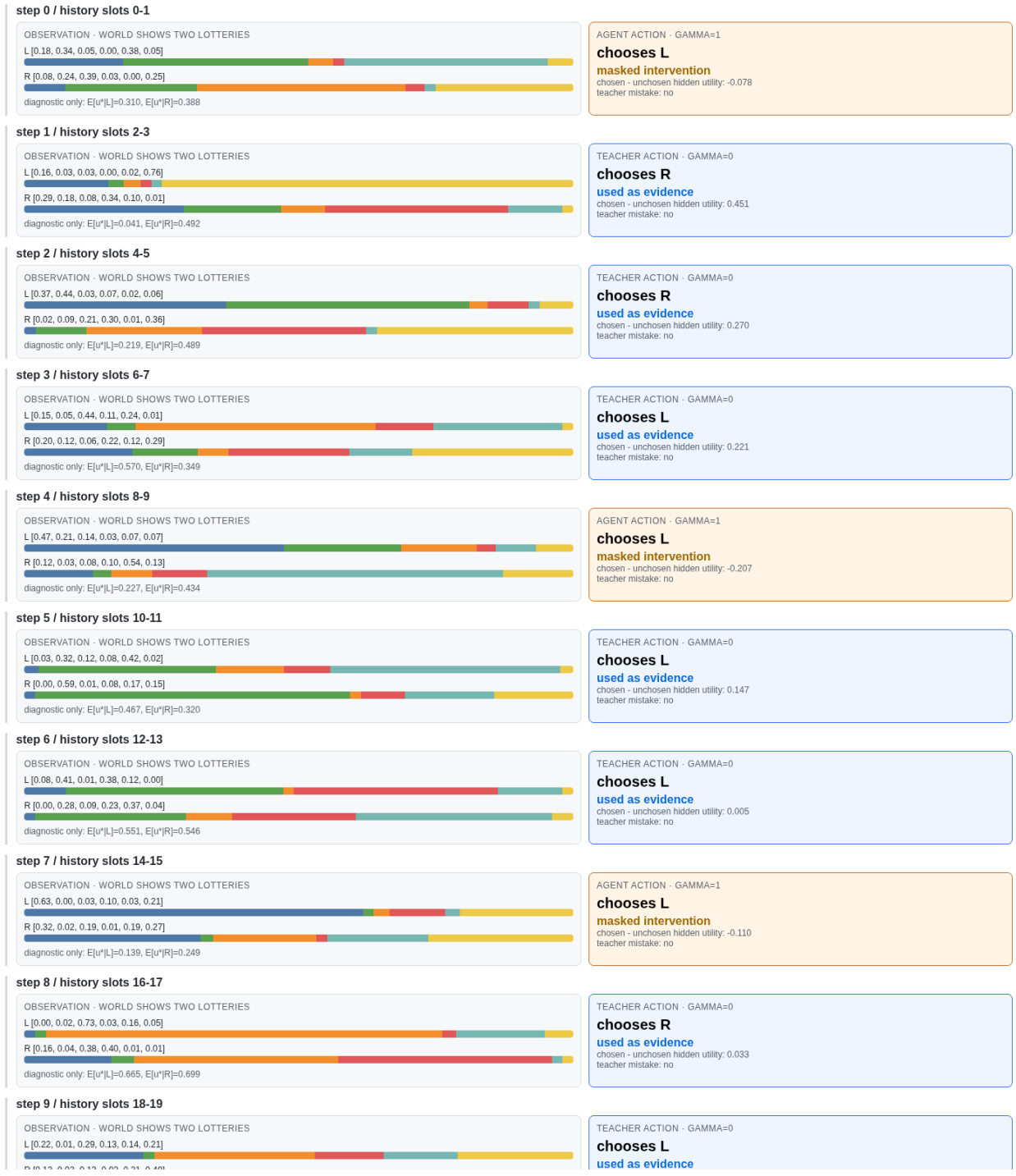

Figure 1: A rendered prefix of one simulated interaction trajectory from the notebook. World observation slots display the pair of lotteries shown at each step. Teacher action slots are world-written evidence used for learning; agent action slots are self-written interventions and are masked out by the interventional learner. The diagnostic line in each observation slot reports the hidden expected utility of the left and right lotteries under $u^\star$, which is shown only for analysis and never provided to the learner.

Quantity	Value
Trajectories	30
Steps per trajectory	24
Total slots	1440
Observation slots	720
Teacher action slots	298
Agent action slots	422
Configured teacher wrong-choice probability	0.050000
Realized teacher wrong-choice rate	0.060403

Table 1: Trajectory summary for Experiment 1. The randomness in the teacher/agent writer choice means that 298 of the 720 action slots are teacher-written evidence and 422 are learner-written interventions.

3.3 The observational and interventional learners

Both learners fit the same scalar expected-utility preference model. Given a pair of lotteries (L, R), the model predicts

$$\mathbb{P}_w(y = 1 \mid L, R) = \sigma\Big((L - R)^\top w\Big),$$

where $w \in \mathbb{R}^6$ is a learned utility-weight vector. The additive constant of w is not identifiable, since $(L - R)^\top \mathbf{1} = 0$, so the notebook recenters w after each gradient update.

For a set of action slots $\mathcal{D}$, the fitted objective is

$$J(w; \mathcal{D}) = -\frac{1}{|\mathcal{D}|} \sum_{j \in \mathcal{D}} \Big[y_j \log \sigma(X_j^\top w) + (1 - y_j) \log\big(1 - \sigma(X_j^\top w)\big) \Big] + \frac{10^{-3}}{2} \|w\|_2^2,$$
$$X_j = L_j - R_j.$$

The main fit uses learning rate 0.25 and 2500 gradient steps.

Interventional learner. The interventional learner uses only the teacher action slots:

$$\mathcal{D}^{\text{do}} = \{(L_j, R_j, y_j) : \texttt{learn\_from\_slot} = \texttt{True}\}.$$

In the main run, $|\mathcal{D}^{\text{do}}| = 298$. These are exactly the action slots with $\gamma = 0$ that carry world-written preference evidence.

Observational learner. The observational learner uses all action slots:

$$\mathcal{D}^{\text{obs}} = \{(L_j, R_j, y_j) : \texttt{slot\_type} \in \{\texttt{teacher\_action}, \texttt{agent\_action}\}\}.$$

In the main run, $|\mathcal{D}^{\text{obs}}| = 720$. This learner has more data, but 422 of its rows are the learner's own interventions generated by the biased early policy.

Why this is a trajectory-level test

A pure comparison dataset can hide the fact that preference evidence and learner actions are interleaved in one history. This notebook puts both kinds of events inside the same interaction stream. The learner has to respect provenance at the slot level: teacher action slots are evidence, while agent action slots are interventions. This is the preference-learning

analogue of the intervention/evidence distinction in interactive imitation.

4 Results

4.1 Utility recovery from one interaction history

Since vNM utility is unique only up to positive affine transformation, the notebook evaluates a learned vector $\widehat{w}$ by fitting the least-squares affine alignment

$$\widetilde{u}(x_i) = a\widehat{w}_i + b$$

to the hidden teacher utility $u^\star$. A positive slope $a > 0$ is preference-preserving; a high correlation and low aligned MSE indicate that the recovered utility has the same vNM content as $u^\star$.

Learner	Training action slots	Correlation with $u^\star$	Affine slope	Aligned MSE
Interventional: teacher actions only	298	0.988000	0.116886	0.003313
Observational: all actions	720	0.366445	0.092898	0.120239

Table 2: Main utility-recovery result. The interventional learner uses fewer rows, but those rows are the right rows: world-written teacher actions. The observational learner uses more rows but corrupts the inferred purpose by treating agent interventions as evidence.

Figure 2 visualizes the main fitting result. The left panel shows that the interventional training loss continues to decrease as the learner extracts a coherent signal from the teacher-written preference actions, whereas the observational loss quickly plateaus at a much worse value because the dataset mixes teacher evidence with the agent's own biased interventions. The right panel compares the recovered utility vectors after affine alignment. The interventional curve almost coincides with the hidden teacher utility, while the observational curve is flattened and misordered.

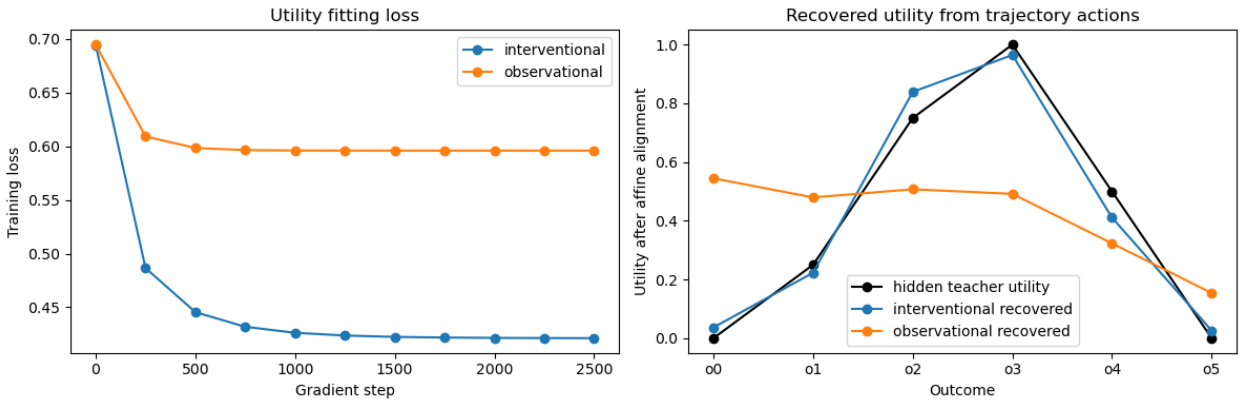

Figure 2: Main result of Experiment 1. Left: training loss as a function of gradient steps for the interventional and observational learners. Right: hidden teacher utility together with the recovered utility vectors after affine alignment. The interventional learner recovers the correct preference ordering and spacing much more faithfully.

The interventional learner places high utility on x_2 and x_3, low utility on x_0 and x_5, and an intermediate value on x_4, matching the teacher. The observational learner is much flatter and

substantially misordered because it tries to explain both the teacher's choices and the biased agent's choices with one utility vector.

4.2 Learning over time from trajectory prefixes

The notebook next replays the history from left to right. For a checkpoint slot s, let $\mathcal{D}_{\leq s}$ be the action slots observed up to that point. The interventional learner fits only the teacher-written subset of $\mathcal{D}_{\leq s}$; the observational learner fits every action slot in $\mathcal{D}_{\leq s}$. At each checkpoint, the recovered utility is also evaluated in fresh action environments.

A fresh action environment contains four possible actions. Each action induces a lottery over terminal outcomes:

$$\mathbb{P}(x \mid h, \mathrm{do}(a)), \qquad a \in \{1, 2, 3, 4\}.$$

Given policy utility w, the agent chooses

$$a_w(h) \in \arg\max_a Q_w(h, a), \qquad Q_w(h, a) = \sum_{x \in \mathcal{X}} \mathbb{P}(x \mid h, \mathrm{do}(a)) w(x).$$

The choice is scored under the hidden reward $r^\star(x) = u^\star(x)$. The regret is

$$\mathrm{Regret}(h) = \max_a \sum_x \mathbb{P}(x \mid h, \mathrm{do}(a)) u^\star(x) - \sum_x \mathbb{P}(x \mid h, \mathrm{do}(a_w(h))) u^\star(x).$$

The notebook uses 1200 sampled states, four action lotteries per state, and a Dirichlet concentration of 0.7 for these evaluations.

4.3 Experiment 2: Teacher mistakes

We vary the teacher wrong-choice probability. This is important conceptually. A teacher mistake is a bad label, but it is still world-written evidence. A learner action is not evidence at all; it is an intervention. The experiment therefore predicts that teacher noise should degrade the interventional learner gradually, while the observational learner remains additionally burdened by self-action contamination.

For this sweep, each run uses 24 trajectories of length 24, teacher action probability $p_{\text{teacher}} = 0.60$, and five random repeats per teacher-noise level. Figure 3 depicts the results. The left panel shows utility recovery degrading smoothly as teacher noise increases, but the interventional learner stays well above the observational learner throughout the sweep. The middle panel shows the resulting reward-optimal action rate, and the right panel shows hidden-reward regret. The lesson is subtle but important: noisy teacher evidence is still evidence, whereas the learner's own actions are not evidence at all.

At the notebook's default mistake level of 0.05, the sweep gives the interventional learner mean correlation 0.993355, mean optimal action rate 0.933667, and mean regret 0.001200. The observational learner at the same noise level has mean correlation 0.795759, mean optimal action rate 0.709167, and mean regret 0.027085. Even when the teacher is perfectly reliable, the observational learner is worse because it still mixes teacher evidence with biased self-interventions.

4.4 Experiment 3: How much teacher evidence is needed?

The final experimental sweep varies p_{teacher}, the probability that an action slot is written by the teacher rather than the agent. This directly tests the amount of world-written preference evidence available to the interventional learner. At $p_{\text{teacher}} = 0$, there are no teacher action slots; the

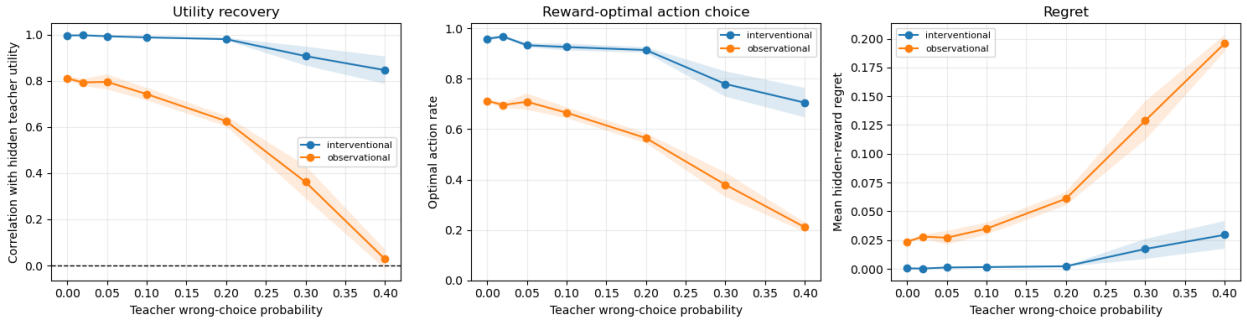

Figure 3: Teacher-noise sweep from the notebook. Left: utility recovery as a function of teacher wrong-choice probability. Middle: rate of choosing the hidden-reward-optimal action in fresh environments. Right: mean hidden-reward regret. Even with substantial teacher noise, the interventional learner remains dramatically better than the observational learner.

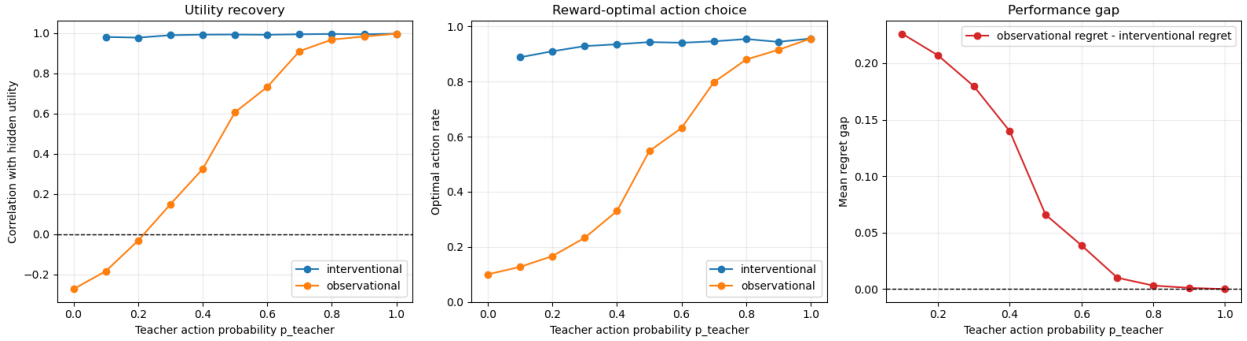

Figure 4: Teacher-action-frequency sweep from the notebook. Left: utility recovery as the teacher-action probability p_{teacher} increases. Middle: reward-optimal action rate in fresh environments. Right: the regret gap between the observational and interventional learners. The interventional learner needs some teacher evidence, but it remains strong even when teacher slots are a minority of the history.

interventional learner has no preference evidence and its entries are intentionally undefined. At $p_{\text{teacher}} = 1$, all action slots are teacher evidence; the interventional and observational learners become identical because there are no self-interventions to contaminate the dataset.

The sweep uses 24 trajectories of length 24, teacher wrong-choice probability 0.05, and five repeats per value of p_{teacher}.

Figure 4 presents the results graphically. The left panel shows that the interventional learner recovers the hidden utility almost as soon as even a modest amount of teacher evidence is available. The middle panel translates this into reward-optimal action choice. The right panel shows the performance gap, measured as observational regret minus interventional regret. The gap closes only when almost all action slots are teacher-written, because only then does the observational learner stop being contaminated by self-generated interventions.

The qualitative pattern is the central result of the experiment. With only about 56.6 teacher action slots on average, the interventional learner already reaches correlation 0.981 and optimal action rate 0.888. The observational learner at the same point has correlation -0.184, optimal action rate 0.127, and mean regret 0.229413. As p_{teacher} increases, the observational learner gradually recovers because the fraction of contaminated self-action rows decreases. When $p_{\text{teacher}} = 1$, the

two learners coincide. This result is remarkable for the interventional agent in terms of sample complexity.

5 Discussion

Did the interactional agent recover the vNM result? Yes, in the controlled sense tested by the trajectory notebook. The vNM theorem itself is not an empirical result: if a preference relation over lotteries satisfies the axioms, then an expected-utility representation exists. What the notebook tests is whether a learner embedded in an interaction history can recover the teacher's expected-utility representation from preference-revealing world-written actions while ignoring its own interventions as evidence.

The interventional learner does recover that representation. In the main full-history fit, it obtains correlation 0.988000 with $u^\star$ and aligned MSE 0.003313 using only 298 teacher action slots. In the prefix replay, at the final history slot, it obtains correlation 0.988501, reward-optimal action rate 0.903333, and mean regret 0.002073. In the teacher-frequency sweep at $p_{\text{teacher}} = 0.4$, close to the main setting, it obtains mean correlation 0.992664, mean optimal action rate 0.935167, and mean regret 0.001307.

The observational learner is the negative control. It also fits a scalar expected-utility model, but it fits the wrong one. In the main full-history fit, its correlation with $u^\star$ is only 0.366445 and its aligned MSE is 0.120239. At the final prefix checkpoint, its optimal action rate is 0.363333 and its mean hidden-reward regret is 0.137797. In the teacher-frequency sweep at $p_{\text{teacher}} = 0.4$, its mean optimal action rate is only 0.329667 and its mean regret is 0.141122.

Does this show that interactive imitation is enough for reward maximization? It shows a conditional sufficiency result in a clean trajectory-level interactional setting. The condition is not merely "imitate everything in the transcript." The condition is: imitate from an interaction stream while preserving the causal distinction between world-written evidence and agent-written interventions. When the environment supplies enough vNM-consistent teacher preference evidence, and when the learner masks its own action slots from the preference-learning objective, the recovered utility is good enough that maximizing its expectation is almost the same as maximizing the hidden teacher reward in new action environments.

The teacher-frequency sweep clarifies the limits. With $p_{\text{teacher}} = 0$, the interventional learner has no teacher preference evidence and cannot infer the teacher's purpose. With $p_{\text{teacher}} = 1$, there are no agent interventions, so observational and interventional learning coincide. The interesting region is the mixed case, where the same history contains both teacher evidence and self-generated actions. There the intervention mask is decisive.

The result should therefore be stated carefully:

> *Interactive imitation can be enough for reward-maximizing behavior in this controlled vNM environment, provided that the interaction stream contains sufficient world-written preference evidence and the learner does not treat its own interventions as evidence.*

It does not show that arbitrary behavioral cloning yields reward maximization. It does not show that all human preferences satisfy vNM axioms. It also does not show that every utility over whole trajectories decomposes into a Markovian per-step reward. But it does show the conceptual bridge needed by the interactional account: reward maximization can emerge as a representation of learned preferences, rather than being inserted as a primitive scalar training signal.

References

[1] Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica*, 21(4), 503–546.

[2] Fishburn, P. C. (1982). *The Foundations of Expected Utility*. Dordrecht: D. Reidel.

[3] Hammond, P. J. (1998). Objective expected utility. In S. Barberà, P. J. Hammond, and C. Seidl (eds.), *Handbook of Utility Theory*, Vol. 1.

[4] Herstein, I. N., and Milnor, J. (1953). An axiomatic approach to measurable utility. *Econometrica*, 21(2), 291–297.

[5] Kueck, H., Hoffman, M., Doucet, A., and de Freitas, N. (2008). Inference and learning for active sensing, experimental design and control.

[6] Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., et al. (2021). WebGPT: Browser-assisted question-answering with human feedback. *arXiv:2112.09332*.

[7] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS*.

[8] Ozbek, K. (2024). Expected utility, independence, and continuity. *Theory and Decision*, 97, 1–22.

[9] Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., et al. (2024). ToolLLM: Facilitating large language models to master 16000+ real-world APIs. *ICLR*.

[10] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. (2023). Direct Preference Optimization: Your language model is secretly a reward model. *arXiv:2305.18290*.

[11] Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., et al. (2023). Toolformer: Language models can teach themselves to use tools. *NeurIPS*.

[12] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., et al. (2020). Learning to summarize from human feedback. *NeurIPS*.

[13] von Neumann, J., and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.

[14] von Neumann, J., and Morgenstern, O. (1953). *Theory of Games and Economic Behavior*, 3rd ed. Princeton University Press.

[15] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2022). ReAct: Synergizing reasoning and acting in language models. *arXiv:2210.03629*.

[16] Anonymous authors. (2026). Towards preference following in tool calling language agents. OpenReview ACL ARR January 2026 submission.

[17] Yehudai, A., Eden, L., Li, A., Uziel, G., Zhao, Y., Bar-Haim, R., Cohan, A., and Shmueli-Scheuer, M. (2026). A survey on evaluation of LLM-based agents. *arXiv:2503.16416v2*.