Individual Differences in Human Teaching of Reinforcement Learning Agents: Evidence from Bayesian Hypothesis Testing

Zhuolun Zhong (zhongzhuolun995@gmail.com) **Luc Caspar (luc.caspar@cross-compass.com)**
Cross Compass
Tokyo, Japan

Cross Compass
Tokyo, Japan

Joseph L. Austerweil (joseph.austerweil@gmail.com)
Chiba Institute of Technology
Tsudanuma, Japan

Abstract

When humans teach reinforcement learning (RL) agents through real-time interventions, does their teaching strategy adapt to environmental context or reflect systematic individual variation? We present a novel web-based platform for studying human teaching via *state interventions* (i.e.: physical relocation of learning agents). In a study with 82 participants, we manipulated world dynamics and how agents interpreted interventions using six formally-defined interpretation types. Bayesian analyses provided moderate-to-strong evidence that neither factor affected teaching behavior ($BF_{01} = 5.22$ for world type; $BF_{01} = 27.08$ for interpretation type). However, participants showed systematic individual variation: high-frequency teachers ($n = 39$; $M = 36$ interventions/round) targeted actions with lower Q-values compared to low-frequency teachers ($d = -0.57$, $BF_{10} = 3.69$). While frequent interventions boosted immediate scores, they often impaired long-term policy learning. The *reset* interpretation, which ignores the intervention, was an exception: it was uniquely robust to suboptimal human guidance. These findings suggest that human teaching in this task is characterized more by persistent individual behavioral profiles than by context-adaptive strategies, with implications for designing resilient human-AI interaction systems.

Keywords: human teaching; state intervention; individual differences; reinforcement learning; Bayesian hypothesis testing

Introduction

Human teaching extends beyond verbal instruction to include direct physical manipulation of a learner's situation. Consider a coach repositioning a player on a field, or a user dragging a robotic vacuum to a dirty spot. In such moments, the teacher does not simply tell the learner what to do. Instead, they forcibly change the learner's state, then step back to let the learner decide on their course of action. This form of guidance, which we call *state intervention*, creates a distinctive pedagogical signal that blends direct environmental influence with preserved learner autonomy.

State interventions occupy a unique place in the spectrum of teaching behaviors. Contrary to evaluative feedback (e.g., numerical rewards), they directly alter the learner's state, and unlike action overrides, they preserve the learner's agency to choose subsequent actions. This combination makes state intervention a powerful yet ambiguous teaching tool. From the learner's perspective, being moved raises a fundamental question: *What does this intervention mean?* Is it a correction of a mistake, a suggestion for where to explore, a simple reset, or even a form of punishment? How the learner interprets the event—its *interpretation type*—can determine whether the intervention accelerates learning, disrupts it, or has little effect.

Two competing hypotheses frame human teaching behavior in such settings: **Context-adaptation hypothesis:** Human teaching may flexibly adapt to environmental characteristics and agent behavior. Classic work on scaffolding has shown that human tutors adjust their level of intervention contingent on learner success and failure (Wood, Bruner, & Ross, 1976), and studies of human–robot interaction have found that people adapt their feedback strategies in response to an RL agent's learning progress (Thomaz & Breazeal, 2008). When agents behave unpredictably or when certain interventions prove more effective, humans might similarly adjust their teaching strategy accordingly. **Individual-differences hypothesis:** Alternatively, human teaching may reflect stable individual styles that persist across contexts. When non-expert humans teach concepts to machine learners, they adopt qualitatively distinct strategies rather than converging on a single approach (Khan, Zhu, & Mutlu, 2011), and these individual differences in teaching style persist across repeated interactions (Ayub et al., 2023).

In this paper, we present a web-based foraging task where participants teach RL agents using drag-and-drop state interventions. We investigate: (1) what strategies humans naturally adopt; (2) whether teaching adapts to world dynamics or agent interpretation type; and (3) whether individual differences in intervention frequency predict teaching strategy.

Prior Work

Our research sits at the intersection of cognitive modeling of pedagogy and interactive machine learning (IML), with a focus on how humans use state interventions to guide learning agents in dynamic, multi-agent environments.

Human Pedagogy and Cognitive Models of Teaching

A foundational perspective in cognitive science is that human teaching is a form of intentional communication. Theories of natural pedagogy suggest that humans use ostensive signals to convey generalizable knowledge (Csibra & Gergely, 2009), often formalized through the goal–means matching principle (Shafto, Goodman, & Griffiths, 2014). In reinforcement learning (RL) contexts, however, humans often struggle to teach effectively because they treat evaluative feedback as communicative signals rather than numerical reinforcements, leading to suboptimal learning outcomes (Ho, Cushman, Littman, & Austerweil, 2019). Our work extends this

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

line of inquiry by examining *state intervention*: a teaching signal that is more direct than evaluative feedback yet more ambiguous than a full demonstration.

Human Factors in Interactive Machine Learning

While early IML research emphasized algorithmic efficiency, recent work highlights the importance of designing systems that account for the diverse and often imperfect ways humans interact with learning agents (Amershi, Cakmak, Knox, & Kulesza, 2014). Key challenges include the scarcity of human guidance (Torrey & Taylor, 2013), the unreliability of human advice (Kempf, Maurer, Turan-Schwiewager, & Koert, 2025), and the impact of user emotional states such as frustration on teaching quality (Gucsi, Tuyen, Chu, Tarapore, & Tran-Thanh, 2025). Our study builds on these insights by investigating how humans naturally intervene in a real-time, multi-agent foraging task—a setting that inherently induces trade-offs between immediate reward maximization and long-term policy improvement.

From Action Advice to State Intervention

Most research on human-in-the-loop RL focuses on *action advice* (suggesting what to do) or *evaluative feedback* (indicating performance quality). For instance, the EMPATHIC framework maps implicit human reactions to reward signals (Cui et al., 2021), while shared-control models allow agents to query human oracles for optimal actions (Avraham & Mirsky, 2025). In contrast, *state intervention* represents a distinct form of guidance. Although prior work has treated physical corrections as a means to infer reward functions in continuous control tasks (Losey, Bajcsy, O'Malley, & Dragan, 2022), the interpretation of such interventions in discrete, multi-agent, and dynamically rewarding environments remains underexplored.

Our study addresses this gap by providing the first systematic investigation of how humans deploy state interventions in a real-time multi-agent foraging environment, and how different algorithmic interpretation types align (or fail to) with human teaching behavior.

Methods

Experimental Framework

To investigate real-time teaching, we developed a web-based platform capable of handling simultaneous interactions between human participants and multiple reinforcement learning agents. The system satisfies four core functional requirements: (1) agents navigate and learn autonomously within a dynamic environment; (2) users can physically intervene on any agent via drag-and-drop, which cancels planned trajectories and records the displacement as a pedagogical signal; (3) agent policies update asynchronously based on both environmental feedback and user interventions; and (4) the framework is engineered for broad accessibility, ensuring fluid user experience, while minimizing local computational demands.

Task Environment

In a grid world environment, each agent occupies a single 1×1 tile and can move in eight directions (four cardinal and four diagonal) to adjacent tiles. Rewards are derived from collecting pellets that spawn within designated regions at a rate determined by a Poisson process (Kingman, 1992). When an agent's avatar overlaps with a pellet, the pellet is collected and the agent receives 6 points, while orthogonal and diagonal moves cost 1 and $\sqrt{2}$ points, respectively.

During an agent's movement, participants can interrupt its trajectory by clicking and holding the mouse to grab the agent. This halts the agent's movement, and the participant drags the agent until the mouse is released. Upon release, the agent's position is set to the center of the nearest grid location, and the agent learns from the intervention according to its assigned interpretation type.

Q-Learning and Interpreting Interventions

Agents use Q-learning (Sutton & Barto, 2018) to approximate the expected cumulative discounted reward. Given a learning rate α and a trajectory (s_t, a_t, s_{t+1}, r_t), the update rule is:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha \left[r_t + \gamma \max_{a'} Q(s_{t+1},a') - Q(s_t,a_t) \right] \tag{1}$$

where s_t is the agent's state at time t, a_t is the action taken, s_{t+1} is the resulting state, and r_t is the reward received. We used an epsilon-greedy policy.

To analyze whether human interventions target optimal or suboptimal agent actions, we must estimate the Q-values of agent behavior. This presents a technical challenge because the true value function depends on the dynamic pellet distribution, which changes over time as pellets spawn and are consumed. The expected reward of an agent's movement is defined as the expected number of pellets encountered within the area swept by the agent's avatar. Formally, let L be the set of grid tiles intersected by the agent's movement path. The expected reward R is calculated as:

$$R = \sum_{t \in L} w_t \cdot P(\text{pellet at } t) \tag{2}$$

where $P(\text{pellet at } t)$ is the probability of a pellet occupying tile t, and $w_t \in (0,1]$ represents the proportion of tile t covered by the agent's trajectory. This accounts for any fractional coverage of adjacent tiles during diagonal move. At the start of a round, this expected reward is determined by the number of pellets generated in the region the agent moves through. However, as the round progresses, non-consumed pellets accumulate in the environment. This violates the Markov assumption, as the true expected reward now depends on the history of prior actions (specifically, which pellets have already been collected). Preserving the Markov property would require augmenting the state space (e.g., storing the number of pellets at each tile), but this would increase complexity considerably and scales poorly when more agents are introduced.

To preserve computational efficiency while obtaining meaningful Q-value estimates, we opted to use discrete grid position as the state space and approximated the expected reward through Monte Carlo simulation. Given a specific agent policy π, we randomly select a starting position from the grid world and sum the agent's cumulative reward from executing 30 steps following the policy in a simulated environment with the same pellet spawn distribution. Averaging this process over 100 independent trials approximates the expected reward $E[R|\pi]$ for a given policy:

$$\hat{Q}(s,a) \approx \frac{1}{100} \sum_{i=1}^{100} \sum_{t=0}^{29} \gamma^t r_t^{(i)} \tag{3}$$

where $r_t^{(i)}$ is the reward at step t during trial i. This approximation allows us to evaluate whether human interventions successfully move agents toward higher-value states, providing a principled basis for classifying interventions as optimal (targeting suboptimal agent actions) or suboptimal (disrupting optimal agent behavior).

To formalize state interventions, we distinguish between s^A_{t+1} (agent's state without intervention), s^I_{t+1} (state after intervention), a_t (action taken), and r_t (environmental reward). We define a^I as the action minimizing distance to s^I_{t+1}, and r^I as the reward at s^I_{t+1}.

We define six distinct interpretation types (see (Ho, Littman, & Austerweil, 2017)):

Suggestion: learns from $(s_t, a^I, s^I_{t+1}, r^I)$ unless dragged back.

Reset: learns from original $(s_t, a_t, s^A_{t+1}, r_t)$, ignoring intervention.

Interrupt: no learning; agent continues from new state.

Transition: learns from $(s_t, a^I, s^I_{t+1}, 2R_{\text{pellet}})$; negative if dragged back.

Disrupt: learns from $(s_t, a^I, s^I_{t+1}, -2R_{\text{pellet}})$ unless $s^I_{t+1} = s_t$.

Impede: punishes original action $(s_t, a_t, s^A_{t+1}, -2R_{\text{pellet}})$.

Simulation Baseline

To provide a comparative benchmark for human pedagogical behavior, we generated baseline data using a model-based simulated teacher. Unlike human participants who infer reward locations through real-time visual observation, the simulated teacher possesses direct access to the underlying parameters of the Poisson pellet-generation process. For each timestep, it intervenes if the agent's action is determined to be suboptimal based on the ground-truth generative model. The grid world was 8×8, and the pellet spawn rate was 100ms. Q-learning agents used $\alpha = 0.1$, $\varepsilon = 0.3$, $\gamma = 0.3$, and a move rate of 200ms.

Procedure

The task was divided into 9 rounds of 100 seconds each. Round 1 was familiarization. Participants completed rounds under four settings: one agent with random distribution, one agent with smooth distribution, two agents with random distribution, and two agents with smooth distribution. Participants were told agents could not see and that agents do not typically learn from pellets obtained during an intervention.

Participants and Design

We recruited 92 participants through CloudResearch. After excluding 10 who did not intervene in more than 4 of the 9 rounds, our final sample comprised 82 participants. The experiment was approved by the University of Wisconsin–Madison Institutional Review Board. The parameters were the same as the simulation baseline.

We employed a mixed design with two factors: **Interpretation type (between-subjects):** How the RL agent interprets human interventions, with six levels (see description above). Participants were randomly assigned ($n = 13$–14 per condition). **World dynamics (within-subjects):** Rounds 2–5 used *random* dynamics (pellet spawn tiles randomly distributed), while rounds 6–9 used *smooth* dynamics (pellet spawns in contiguous clusters).

Teaching Hypotheses

We formalize teaching styles into four probabilistic models. For each hypothesis H, we define a target tile set S_H representing where a teacher would place the agent. Let $\mathcal{N}_{n\times n}(c)$ denote an $n \times n$ neighborhood centered at c, P the set of pellet spawn tiles, and NP non-spawn tiles.

Hypothesis 1 (Undoing)**.** *The teacher guides the agent to retry from its starting position:* $S_{H1}(s_{start}) = \{s_{start}\}$

Hypothesis 2 (Correcting)**.** *The teacher moves the agent to tiles containing pellets:* $S_{H2} = \{t \in \mathcal{N}_{5\times5}(s_{start}) \mid t \in P\}$

Hypothesis 3 (Exploration-encouraging)**.** *The teacher guides the agent to areas adjacent to pellet spawn tiles to facilitate discovery:* $S_{H3}(s_{start}) = \{t \in \mathcal{N}_{5\times5}(s_{start}) \mid t \in NP \wedge \exists p \in \mathcal{N}_{3\times3}(t), p \in P\}$.

Hypothesis 4 (Restart)**.** *The teacher forces a fresh start by moving the agent to isolated regions far from current pellet clusters:* $S_{H4}(s_{start}) = \{t \in \mathcal{N}_{5\times5}(s_{start}) \mid t \in NP \wedge \forall p \in \mathcal{N}_{3\times3}(t), p \notin P\}$.

We assign each participant to the strategy that best explains their interventions using maximum likelihood estimation.

Results

Participants made 12,626 total interventions across 8 rounds ($M = 154$ per participant, $SD = 159$). The total number of interventions was similar across both distribution conditions (6,339 in random vs. 6,253 in smooth), suggesting that environmental structure did not substantially influence intervention frequency. All Bayesian analyses were performed using JZS priors with a default Cauchy scale of $r = 0.707$, which is standard for testing effects without specific prior directional constraints (Wagenmakers, 2007).

Among all suboptimal actions taken by agents, 27.32% ($\frac{8149}{29828}$) were intervened upon. To understand the appropriateness of human interventions, we categorized each intervention based on the optimality of the agent's original action:

Optimal intervention: The agent was taking a suboptimal action; the intervention was warranted. **Suboptimal intervention:** The agent was taking an optimal action; the intervention was unnecessary. While optimal interventions were more common, a substantial proportion (42.05% in random, 28.42% in smooth) fell into the suboptimal category.

Teaching Strategy Analysis

Table 1 shows the teaching hypothesis assignments. The most common strategy was *Correcting* (H2), where participants moved agents to pellet tiles. This pattern was especially pronounced in the smooth distribution setting, where pellet locations were more clustered and perhaps easier to identify. Notably, despite all 70 participants who completed the demographics survey confirming their understanding that pellets obtained through intervention do not aid learning, they still overwhelmingly adopted the correcting strategy. This suggests that score-driven behavior may reflect an intuitive teaching heuristic that is resistant to explicit instruction.

World Dynamics Do Not Affect Teaching

A key question is whether human teaching adapts to environmental structure. The random distribution scattered pellet spawn tiles across the grid, while the smooth distribution clustered them in contiguous regions. If humans adapt their teaching to environmental structure, we would expect different intervention patterns.

Using Bayesian t-tests, we found moderate evidence for the null hypothesis ($BF_{01} = 5.22$), indicating that the data are approximately five times more likely under the assumption that world type (random vs. smooth) does not affect the mean Q-value of targeted actions. This suggests that participants' teaching targets remained invariant regardless of environmental dynamics.

Although participants applied the same teaching *strategy* regardless of environment, their intervention *accuracy* differed substantially. A chi-square test revealed a significant association between environment type and intervention optimality, $\chi^2(1) = 255.84$, $p < .001$, $\phi = 0.14$. Bayesian analysis provided extreme evidence for this association ($BF_{10} > 1000$). Suboptimal interventions were more frequent in random environments (42.05%) than smooth ones (28.42%), with odds ratio $OR = 1.83$, 95% CI [1.70, 1.97]. This pattern is consistent with a fixed, score-driven teaching approach: participants attempted the same behavior in both conditions but succeeded more often in smooth environments where optimal paths were visually apparent. The environment affected how *well* their strategy worked, not *what* strategy they used.

Interpretation Type Does Not Affect Teaching

Participants were assigned to one of six interpretation conditions, which determined how agents learned from interventions. If humans are sensitive to the downstream effects of their teaching, we might expect participants in different conditions to adopt different strategies.

A one-way Bayesian ANOVA provided strong evidence for the null hypothesis ($BF_{01} = 27.08$), suggesting that the data are 27 times more likely under a model where interpretation type has no effect on teaching behavior. Participants appeared insensitive to the agent's internal learning logic, relying instead on domain-general heuristics.

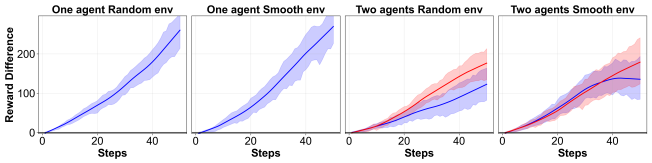
Individual Differences in Teaching Strategy

While environmental and algorithmic factors did not predict teaching behavior, we observed substantial variation between participants. To investigate this, we categorized participants into high-frequency ($n = 39$; $M = 36.1$ interventions/round) and low-frequency ($n = 43$; $M = 3.9$ interventions/round) groups using a median split at 15 interventions per round.

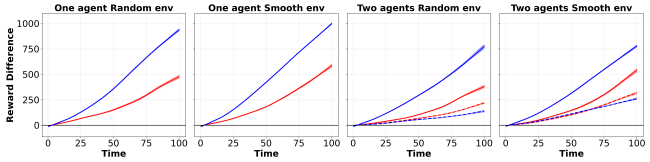
High-frequency teachers targeted actions with lower Q-values ($M = 0.247$, $SD = 0.025$) than low-frequency teachers ($M = 0.260$, $SD = 0.022$). Bayesian analysis yielded moderate evidence for this difference ($BF_{10} = 3.69$, $d = -0.57$). This relationship suggests that high-frequency intervention is associated with a more selective targeting of the agent's poorest actions.

This pattern reveals systematic heterogeneity in how participants approached the teaching task. While these behavioral profiles—active correction versus passive observation—were consistent across rounds, further research is required to determine if they represent stable, trait-like teaching styles that generalize to other domains.

Intervention Effects on Performance



(a) Simulated teacher (red = lower-scoring agents).



(b) Human (blue = high-frequency, red = low-frequency).

Figure 1: Cumulative score difference from no-intervention baseline. Human intervention is explicitly aimed at scoring, with higher frequency leading to more points.

Each round lasts 100 seconds, and without intervention, agents complete approximately 50 steps. Since human observation and intervention take time, the actual number of steps per round typically ranges between 40 and 60. Interestingly, low-frequency interventions tend to take longer (perhaps reflecting more deliberate decision-making), resulting in fewer agent steps, while high-frequency interventions are executed more rapidly.

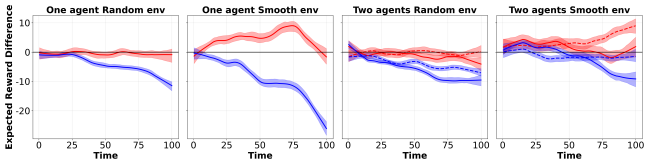
Table 1: Teaching hypotheses by world distribution. Most participants adopted the Correcting strategy (H2), moving agents directly to pellets to maximize immediate rewards. This pattern was especially pronounced in smooth distribution environments where pellet clusters were more spatially coherent and easier to identify visually. Note: Counts represent per-environment assignments, as participants are categorized separately for random and smooth distributions.

	H1: Undoing	**H2: Correcting**	**H3: Exploration-Encouraging**	**H4: Restart**
Random Distribution	9 (14%)	51 (80%)	3 (5%)	1 (2%)
Smooth Distribution	4 (6%)	58 (91%)	1 (2%)	1 (2%)

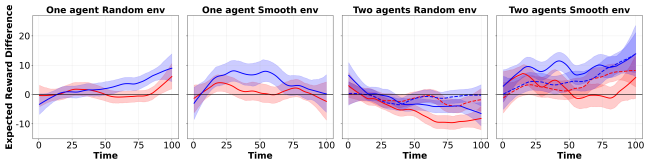
Table 2: Summary of statistical tests with Bayes factors. Bayes factors are interpreted following standard guidelines: $BF > 3$ provides moderate evidence, $BF > 10$ provides strong evidence, and $BF > 30$ provides very strong evidence.

Effect	BF	Evidence Level	Key Finding
World type	$BF_{01} = 5.22$	Moderate H_0	Teaching targets are invariant to environment
Interpretation type	$BF_{01} = 27.08$	Strong H_0	Teaching does not adapt to agent learning logic
Intervention frequency	$BF_{10} = 3.69$	Moderate H_1	Frequent teachers target lower Q-values
Reset vs. other	$BF_{10} = 4.07$	Moderate H_1	Reset interpretation mitigates suboptimal guidance

Figure 1 compares cumulative score differences from a no-intervention baseline for both simulated and human teachers. Simulated interventions (triggered only for suboptimal actions) produce modest score improvements. In contrast, human participants with high-frequency interventions substantially boost agent scores. This pattern is consistent with the *correcting* strategy: even when an agent takes an optimal action that would lead to an empty tile, participants intervene to redirect it toward tiles with visible pellets. This behavior successfully maximizes immediate point accumulation, but comes at a significant cost to policy learning.



(a) Human results (most interpretation types).



(b) Human results (reset interpretation only).

Figure 2: Expected reward difference from baseline. High-frequency interventions harm policy learning—except for reset interpretation.

Figure 2 reveals this tension starkly: while simulated teachers eventually build stable policy advantages, high-frequency human interventions significantly *harm* agents' expected reward values. Agents taught by frequent human interventions often performed worse than those receiving no teaching at all.

However, the *reset* interpretation proved uniquely robust to this problem (Figure 2b). Because reset agents learn from their original intended actions despite being relocated, they are insulated from suboptimal human guidance. This "pedagogical filter" preserves autonomous learning while still allowing humans to influence immediate scores—a design that compensates for predictable human biases rather than requiring humans to teach optimally.

To quantify this moderation effect, we correlated intervention frequency with final expected Q-value within each condition. Reset showed a positive relationship ($r = .43$, $n = 15$), while other conditions pooled showed a slight negative relationship ($r = -.11$, $n = 66$). While traditional correlation comparisons remained inconclusive, Bayesian analysis provided moderate evidence for a difference between these conditions ($BF_{10} = 4.07$).

Discussion

This study provides the first systematic investigation of how humans use state interventions to teach reinforcement learning agents in a real-time, multi-agent foraging environment. Our Bayesian hypothesis testing approach allowed us to distinguish between "no evidence" and "evidence for no effect," revealing that human teaching behavior in this task is characterized more by systematic individual variation than by flexible, context-adaptive responses to environmental or algorithmic factors.

Score-Driven Teaching: A Robust Human Tendency

We find that human teachers are instinctively *score-driven*: overwhelmingly adopting the *correcting* strategy, which redirects an agent towards immediately visible pellets, when it was already taking an optimal action. This tendency persists despite knowledge that intervention-gained rewards do not aid learning, as well as across environments. This echoes prior findings that humans treat teaching as communication rather than optimization (Ho et al., 2019; Amershi et al., 2014), and reflects human intuition about *"helping"*: one that

prioritizes observable outcomes over invisible learning processes.

Why Didn't Context Matter?

The null effect of interpretation type ($BF_{01} = 27.08$) is particularly striking given that this factor directly affects learning outcomes. Several explanations merit consideration.

First, participants may not have perceived the downstream effects of their interventions during the brief training rounds. Unlike evaluative feedback, where reward signals are explicitly displayed, the consequences of state interventions on Q-value updates are invisible to the teacher. Hence, participants defaulted to observable metrics (score) rather than latent ones (policy quality).

Second, the intuition that "moving someone closer to a goal helps them" may override sensitivity to how the learner actually processes that guidance. This aligns with theories of natural pedagogy (Csibra & Gergely, 2009), which suggest humans have evolved general-purpose teaching instincts that may not transfer seamlessly to artificial learners.

Third, the real-time nature of our task may have promoted fast, intuitive decision-making over deliberative strategy adaptation. Under time pressure, participants may have relied on cached heuristics rather than reasoning about the algorithmic implications of each interpretation type.

The Teaching Style–Targeting Relationship

High-frequency teachers systematically targeted lower Q-value actions ($BF_{10} = 3.69, d = -0.57$), suggesting two distinct behavioral profiles:

Active correctors: These participants intervened frequently (averaging 29 times per round) and strategically targeted the agent's poorest actions. This style maximizes corrective feedback density but risks over-controlling agents that might benefit from autonomous exploration.

Passive observers: These participants intervened rarely (averaging 7 times per round) and less selectively. They may believe agents learn better through autonomous experience, or they may be less engaged with the teaching task. Interestingly, this "hands-off" approach sometimes yielded better policy outcomes precisely because it generated fewer suboptimal interventions.

This individual variation has practical implications: interactive AI systems may benefit from dynamically estimating a user's teaching profile—rather than treating all users uniformly—to better interpret the intent behind their interventions.

Reset as a Pedagogical Filter

While the classical test comparing intervention–outcome correlations remained inconclusive, the Bayesian analysis provided moderate evidence for a differential effect between conditions ($BF_{10} = 4.07$). The positive correlation observed in the reset condition contrasts with the negative trend in other conditions. This pattern is consistent with our finding that the *reset* interpretation proved uniquely robust to the inherent noise in human guidance.

Crucially, this robustness emerges from the fact that a substantial proportion of human interventions (28–42%) are suboptimal. When the simulated teacher (only triggered by suboptimal actions) used the reset interpretation, learning was impaired because valuable corrective information was ignored.

This result has important design implications. Instead of expecting humans to teach like algorithms, future work might develop "teacher-aware" agents that dynamically adapt their interpretation to compensate for biases in teaching style (Kempf et al., 2025; Gucsi et al., 2025).

However, given the small sample size in the reset condition ($n = 15$), these findings should be interpreted with caution and warrant replication with larger samples.

Limitations and Future Directions

Several limitations warrant caution. First, our teaching-profile categorization used an exploratory median split of intervention frequency; future work should employ independently motivated constructs and test-retest designs to establish stable trait variables. Second, while intervention rates were comparable across conditions, we did not directly measure participants' intended pedagogical meaning; stronger tests of alignment would require eliciting or manipulating teacher intent. Third, the fixed sequence of environment types (random then smooth) introduces order effects that cannot be fully decoupled from environmental adaptation, despite moderate evidence for a null effect ($BF_{01} = 5.22$). Fourth, the reset condition's sample size ($n = 15$) was small; while the Bayesian evidence is encouraging, replication is needed.

Future work should investigate the "ecology of teaching"—how humans strategically choose among intervention types (evaluative feedback, demonstrations, state interventions) given real-time agent behavior and environmental dynamics. Equally promising is the development of adaptive agent interpretations that treat interventions as communicative acts, performing counterfactual reasoning about the teacher's latent intent.

Conclusion

Human teaching of RL agents in this task appears to reflect systematic individual heterogeneity rather than flexible adaptation to context. Bayesian testing provided a rigorous framework to quantify evidence for these effects, showing that a user's intervention frequency is a key predictor of their targeting strategy. Importantly, the reset interpretation provides a practical design principle: agents can be made robust to the variance in human guidance by learning from their original intentions despite physical relocation. As AI systems increasingly learn from human feedback, accommodating this human-centered variation becomes critical for building effective collaborative partners.

Acknowledgments

The authors would like to thank the participants for their time and contribution to this study. We also thank the anonymous reviewers for their insightful feedback on earlier drafts of this paper. JLA was funded by the Japanese Probabilistic Computing Consortium Association (JPCCA). **All experimental code and data analysis scripts are publicly available at: `https://github.com/ZhuolunZhong/Physical-Intervention-with-multi-agent-system.git`.**

References

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI magazine*, *35*(4), 105–120.

Avraham, I., & Mirsky, R. (2025). Shared control with black box agents using oracle queries. In *2025 IEEE International Conference on AI and Data Analytics (ICAD)* (pp. 1–8).

Ayub, A., Mehta, J., De Francesco, Z., Holthaus, P., Dautenhahn, K., & Nehaniv, C. L. (2023). How do human users teach a continual learning robot in repeated interactions? In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153.

Cui, Y., Zhang, Q., Knox, B., Allievi, A., Stone, P., & Niekum, S. (2021). The empathic framework for task learning from implicit human feedback. In *Conference on robot learning* (pp. 604–626).

Gucsi, B., Tuyen, N. T. V., Chu, B., Tarapore, D., & Tran-Thanh, L. (2025). User-aware collaborative learning in human-robot interactions. In *2025 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 7399–7406).

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2019). People teach with rewards and punishments as communication not reinforcements. *Journal of Experimental Psychology: General*, *148*(3), 520–549.

Ho, M. K., Littman, M. L., & Austerweil, J. L. (2017). Teaching by intervention: Working backwards, undoing mistakes, or correcting mistakes? In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 526–531). Austin, TX: Cognitive Science Society.

Kempf, L., Maurer, C., Turan-Schwiewager, C., & Koert, D. (2025). Can i trust you?—handling unreliable human action advice in interactive reinforcement learning. *ACM Transactions on Human-Robot Interaction*, *14*(4), 1–26.

Khan, F., Zhu, X., & Mutlu, B. (2011). How do humans teach: On curriculum learning and teaching dimension. In *Advances in Neural Information Processing Systems* (Vol. 24).

Kingman, J. F. C. (1992). *Poisson processes*. Clarendon Press.

Losey, D. P., Bajcsy, A., O'Malley, M. K., & Dragan, A. D. (2022). Physical interaction as communication: Learning robot objectives online from human corrections. *The International Journal of Robotics Research*, *41*(1), 20–44.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Thomaz, A. L., & Breazeal, C. (2008). Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, *172*(6-7), 716–737.

Torrey, L., & Taylor, M. (2013). Teaching on a budget: Agents advising agents in reinforcement learning. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems* (pp. 1053–1060).

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100. doi: 10.1111/j.1469-7610.1976.tb00381.x