Interpretational alignment: How agents learn from physical guidance depends on how they interpret it

Zhuolun Zhong
zhongzhuolun995@gmail.com
Cross Compass
Tokyo, Japan

Ben Prystawski
benpry@stanford.edu
Department of Psychology
Stanford University, USA

Sarah A. Wu
sarahawu@stanford.edu
Department of Psychology
Stanford University, USA

Bella Fascendini
bfascendini@princeton.edu
Department of Psychology
Princeton University, USA

Sepehr Saeedpour
sepehr.saeedpour@ens.psl.eu
Département d'Études Cognitives
Université PSL, Paris, France

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

Abstract

In many pedagogical contexts, teachers guide learners through direct physical intervention—a modality we term *state intervention*. While common in human interaction, the computational mechanisms that allow agents to learn from being physically moved remain underexplored. We propose that the effectiveness of state intervention hinges on the *interpretational alignment* between teacher and learner: the learner must infer whether an intervention functions as a suggestion, a correction, or a non-pedagogical event. We investigated this problem using a 1D navigation task where human teachers (N=64) trained Q-learning agents under four different interpretation types. Our results show that agents learn most effectively when they interpret interventions as pedagogical signals—either as recommended actions (*suggestion*) or as discouraging actions (*impede*). In contrast, interpreting interventions as environment resets (*reset*) or mere interruptions (*interrupt*) led to significantly poorer performance and learning plateaus. Crucially, providing teachers with a visual representation of the agent's internal beliefs (Q-table) did not improve teaching effectiveness, suggesting that aligning the agent's learning rules with human intuition is more critical than information transparency. These findings highlight the importance of designing "socially-aware" learners that treat physical interaction as intentional communication.

Keywords: state intervention; reinforcement learning; human-robot interaction; pedagogy; interpretational alignment

Introduction

In many pedagogical scenarios, teachers guide learners through direct physical intervention rather than verbal instruction. A parent might guide a child's hand to start a letter, or a user might relocate a struggling robotic vacuum. This mode of teaching—which we term *state intervention*—is prevalent in scaffolded instruction (Vygotsky, 1978), motor learning (Marchal-Crespo, McHughen, Cramer, & Reinkensmeyer, 2010), and human-robot interaction (Hoffman, Bhattacharjee, & Nikolaidis, 2024). Yet, computational models of pedagogy often overlook this modality, focusing instead on demonstrations (Argall, Chernova, Veloso, & Browning, 2009) or evaluative feedback (Knox & Stone, 2009).

State intervention represents a distinct pedagogical signal: it forces a new state while preserving the learner's autonomy to decide the next action. This creates a unique interpretational challenge: the learner must infer *why* it was moved to determine *what* to learn. Does the move imply the previous state was "bad," or the new one "good"? Is it a suggested action, a reset, or mere noise? Each interpretation implies a different learning rule, and misalignment can cause learners to miss pedagogical signals entirely. Following research on pedagogical reasoning (Ho, Cushman, Littman, & Austerweil, 2019; Csibra & Gergely, 2009), we propose that its effectiveness hinges on whether the learner interprets the intervention as communicative—a challenge we term the *interpretational alignment problem*.

Studying this in naturalistic settings is difficult due to high-dimensional states and noisy feedback. A minimal task allows us to precisely specify how different interpretations map onto learning dynamics and observe human behavior when those dynamics vary. We examine these mechanisms using a 1D navigation task. By reducing the environment to a single spatial dimension with unambiguous optimal behavior, we can isolate the cognitive mechanisms of learning from intervention and observe how teachers adapt when the learner's internal interpretational logic remains opaque.

Specifically, we use a minimal paradigm where human teachers train a Q-learning agent by moving it, without knowing its interpretation to interventions. We manipulate the interpretation rule and whether teachers can observe the agent's policy (Q-table). We address three questions: (1) Do agents learn differently depending on how they interpret interventions? We predict pedagogical interpretations will outperform others. (2) Does policy visibility assist teachers? If alignment is the bottleneck, technical transparency may not help. (3) Do teachers improve over repeated interactions, and is this improvement conditional on interpretational alignment?

Background

We examine the distinctiveness of state intervention by relating it to other teaching signals, human-robot interaction, motor learning, and pedagogical reasoning.

Computational Properties of Teaching Modalities

Evaluative Feedback: Learners receive positive or negative feedback on their past behavior. In frameworks like TAMER (Knox & Stone, 2009), learners use this feedback to shape their policies. However, human feedback is often more complex than a simple good/bad label: feedback is used to

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

provide guidance and motivation too (Thomaz & Breazeal, 2008). Learning is most effective when learners recognize this communicative intent (Ho et al., 2019). MacGlashan et al. (2017) show that misalignment between what teachers intend and what learners infer can lead to systematic failures, especially when learners update policies in ways unanticipated by teachers. From the learner's perspective, this signal is retrospective: feedback comments on what the learner has already done.

Action Advice: Rather than evaluating past behavior, learners receive explicit suggestions about what actions to take at certain states (Torrey & Taylor, 2013; Griffith, Subramanian, Scholz, Isbell, & Thomaz, 2013). While this reduces the learner's policy search space, it assumes the learner can execute the suggested action and understands how to interpret it. Critically, the pedagogical meaning of advice is unambiguous—the teacher is recommending an action.

Demonstrations: Teachers show learners how to perform a task by doing it themselves. Learners can then imitate the demonstrated behavior directly or infer the goals underlying the demonstration (Argall et al., 2009; Abbeel & Ng, 2004). While demonstrations offer rich information, they impose high cognitive and computational costs, and can constrain learning. Bonawitz et al. (2011) showed that children who received pedagogical demonstrations explored less and discovered fewer novel functions, because they rationally inferred that a knowledgeable teacher would have shown them anything else worth knowing. The function of a demonstration is clear: the teacher is showing the learner what to do.

State Intervention: State intervention—physically relocating the agent—differs from all of the above. Unlike feedback, it does not explicitly evaluate past behavior. Unlike advice, it does not prescribe a specific action. Unlike demonstrations, it does not convey the teacher's policy. Instead, intervention forces a state change while preserving the learner's autonomy to decide what to do next. This ambiguity creates the interpretational challenge at the heart of our study: the learner must infer *why* they were moved before learning *what* to do (Hoffman et al., 2024). How the learner makes this inference determines what, if anything, they learn from it.

How Should Physical Guidance be Interpreted?

Physical corrections are common in human-robot interaction (HRI). Users reposition robot arms, redirect autonomous vehicles, or manually steer drones away from obstacles. Recent HRI research has begun to formalize what agents should infer from such interactions: Bajcsy, Losey, O'Malley, and Dragan (2018) model physical interventions as evidence about the human's reward function, and Losey and O'Malley (2018) demonstrate that treating physical contact as a communicative signal allows robots to learn new objectives. However, these frameworks typically assume a **fixed interpretational context**: the robot pre-defines the intervention as a "correction" to its current policy.

The core challenge we address is that in naturalistic teaching, the meaning of a forced move is often ambiguous (Losey, Bajcsy, O'Malley, & Dragan, 2022). An intervention could be a recommendation of a better state, a punishment of a previous action, or a non-pedagogical interruption (e.g., the teacher simply clearing the workspace). While some work has begun to incorporate uncertainty about human intent (Ho, Littman, & Austerweil, 2017; Losey & O'Malley, 2018), it remains unclear how the *type* of interpretation chosen by the agent—among multiple plausible pedagogical and non-pedagogical candidates—affects the overall teaching trajectory.

Research on haptic guidance in motor learning further highlights this ambiguity. The "guidance hypothesis" (Schmidt, 1991) suggests that physical guidance can actually *impair* learning if the learner becomes dependent on the external force rather than internalizing the skill. This tension—guidance as a helpful "suggestion" versus a disruptive "interruption"—suggests that the learner's *subjective interpretation* of the force is as important as the force itself. Our study fills this gap by systematically comparing how different interpretational "priors" enable or disable effective learning from the same physical input.

Interpretational Alignment in Pedagogy

Effective teaching requires coordination between teacher and learner. In models of "natural pedagogy," ostensive signals from teachers trigger learners to interpret evidence as relevant and generalizable (Csibra & Gergely, 2009). This has been formalized as recursive inference: teachers select evidence based on what they believe learners need; learners interpret that evidence based on what they believe teachers intend (Shafto, Goodman, & Griffiths, 2014; Gweon, 2021). Much of this work uses Partially Observable Markov Decision Processes (POMDPs) to model teaching (Kaelbling, Littman, & Cassandra, 1998; Rafferty, Brunskill, Griffiths, & Shafto, 2016). However, most computational models focus on demonstrations or labels rather than state interventions.

Unlike demonstrations or labels, a state intervention does not clearly signal what dimension of the situation is pedagogically relevant—the learner must infer whether the intervention comments on states, actions, or rewards. Our framework operationalizes this intuition as four distinct assumptions a learner might make. A learner who interprets the intervention as a *suggestion* assumes the teacher is recommending the action that leads to the new state. A learner who interprets it as an *impediment* assumes the teacher is punishing the prior action. A learner who interprets it as a *reset* assumes the intervention carries no information about actions at all, only that the environment has changed. A learner who interprets it as an *interruption* assumes the intervention is not pedagogical and requires no update whatsoever. Each interpretation implies a different belief about whether the teacher is communicating and thus a different learning rule.

Computational Framework

Task Environment

The environment is a 1D grid with 12 cells arranged as H-F-S-F-F-F-F-F-F-F-F-G (Figure 1). From left to right, the layout is: hole (H), start (S), eight ice cells (F), and goal (G). The agent begins at the start and can move one cell left or right. An episode ends when the agent reaches the hole (reward = 5) or goal (reward = 10); ice and start cells yield no reward. The hole's reward is suboptimal, but closer to start. This provides an opportunity for a teacher to help. A thought bubble above the agent indicates its next action.

For participants in the Q-table visible condition, the interface displays the agent's current Q-values. The shading indicates Q-value magnitude, and the displayed direction shows the agent's preferred action at each state. States with equal Q-values for both actions display "N/A."

To intervene, participants click and drag the agent to a new location. This action pauses the agent's movement; upon release, the agent's position gets set to the nearest grid location, updates its Q-table according to its interpretation (see Computational Model), and resumes autonomous movement.

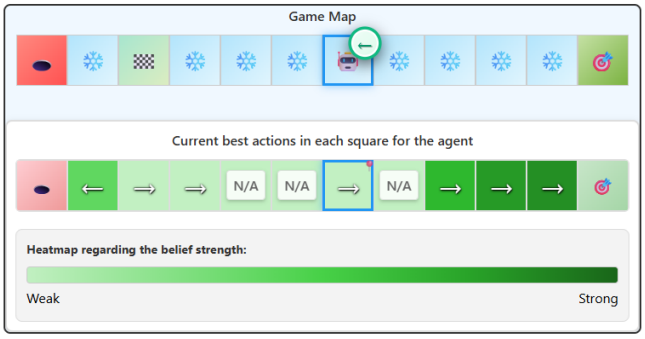

Figure 1: Schematic of the experimental procedure

Interpretation Types

We represent the problem of navigating an environment with discrete states and actions, and with the goal of optimizing the expected cumulative discounted reward as a Markov Decision Process (MDP) (Sutton & Barto, 2018). An MDP is defined by (S, A, T, R, γ), where S is the set of possible states, A is the set of possible actions (moving one square left, right), $T : S \times A \to [0,1]^{|A|}$ encodes the probability of next states given actions (deterministic), $R : S \times A \times S \to \mathbb{R}$ encodes the expected immediate reward after taking an action in the current state and arriving in the next state (discussed further below), and the discount factor $\gamma \in [0,1]$.

Q-Learning and Interpreting Interventions

Q-learning (Sutton & Barto, 2018) approximates the expected cumulative discounted reward of taking any action at any state. Given a learning rate α and a trajectory from a single action (s_t, a_t, s_{t+1}, r_t), it updates as follows:

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

where s_t is the agent's state at time t, a_t is the action taken by the agent at time t, s_{t+1} is the agent's state at time $t+1$, and r_t is the reward received by the agent at time t. We used an epsilon-greedy policy, which takes an optimal action according to the current Q-table with probability $1-\varepsilon$ and a random action otherwise.

We define four methods to interpret a state intervention that interrupts an agent's movement between times t and $t+1$, each of which yields distinct learning dynamics. Let s^I_{t+1} and s^A_{t+1} denote the agent's state at time $t+1$ with and without intervention, respectively. Let a_t denote the action the agent is about to execute at time t, and a^I be the action that, among those available in state s_t, minimizes the Euclidean distance to the intervention state s^I_{t+1}. First, an intervention can be interpreted as a *suggestion* towards a new action and new state. The agent ignores its original action and uses $(s_t, a^I, s^I_{t+1}, 1)$ for its Q-update. The second interpretation type is *reset*, which treats the intervention as a relocation. The agent learns from its original state-action sequence and uses (s_t, a_t, s^A_{t+1}, r) for its update. The third interpretation type is *interrupt*, which treats the intervention as simply an interruption of its autonomy. The agent makes no update in this case. Finally, the last interpretation type is *impede*, which punishes the agent by assigning a negative reward to its original state-action sequence. The agent uses $(s_t, a_t, s^A_{t+1}, -1)$ for its Q-update.

Experiment

Participants and materials

64 participants were recruited from Prolific and completed the task. The study used a 4 (interpretation type: suggestion, reset, interrupt, impede) $\times$ 2 (Q-table visibility: visible, hidden) between-subjects design, with conditions counterbalanced across participants. Participants were unaware of their agent's interpretation type throughout the experiment. Half of the participants could see a visual representation of the agent's Q-table; the other half could not.

The agent paused for 1 second between actions to accommodate human reaction time. Participants were informed that dropping the agent directly onto the goal would not teach it the reward. The experiment was approved by the University of Wisconsin-Madison Institutional Review Board.

Procedure

Participants completed 3 rounds of training, with the agent's Q-table reset to zero at the start of each round. Each round consisted of 15 episodes. Before training, participants were explicitly instructed to act as teachers with the goal of training the agent to independently and smoothly reach the goal while avoiding the hole. They were informed that the agent would learn from their interventions. Interventions were performed by dragging and dropping the agent to any desired

position whenever the participants deemed it necessary. To assist their pedagogical decisions, the agent's intended next action was visualized as an arrow above its head. To ensure consistent teaching effort, participants were cautioned that rewards gained directly through dragging did not always contribute to the agent's internal learning; instead, the agent had to reach states through its own volition to reinforce its policy.

Measures

We evaluated learning using three metrics. *Hamming distance* measures the proportion of non-terminal states where the greedy action differs from optimal. *Success rate* was estimated by simulating 100 episodes per Q-table using an ε-greedy policy ($\varepsilon = 0.1$), recording the proportion of episodes reaching the goal. *Q-RMSE* quantifies how closely learned Q-values approximate theoretical optimal computed via value iteration. For condition-level visualizations (Fig 2), we aggregated Q-tables after per-subject min-max normalization.

Q-Table Reconstruction and Normalization Each participant's Q-table was initially incomplete due to storage optimizations that omitted state-action pairs with Q-values of zero. We reconstructed complete Q-tables for each participant-round combination, resulting in a 12×2 matrix where rows represent states (0-11) and columns represent the two possible actions: moving left and moving right.

To compare Q-tables across different subjects and rounds while accounting for scale variations, we used per-panel Min-Max normalization. For each subject's Q-table $\mathbf{Q}^{(i)}$:

$$\mathbf{Q}^{(i)}_{\text{norm}} = \frac{\mathbf{Q}^{(i)} - \min(\mathbf{Q}^{(i)})}{\max(\mathbf{Q}^{(i)}) - \min(\mathbf{Q}^{(i)})}$$

where $\min(\mathbf{Q}^{(i)})$ and $\max(\mathbf{Q}^{(i)})$ represent the minimum and maximum Q-values in table i. This transformation scales all Q-values to the range [0,1] within each participant's Q-table, preserving relative value differences while eliminating absolute magnitude variations.

Aggregated Representative Q-Table To derive a representative Q-table for groups of participants, we aggregated normalized Q-tables using mean averaging. For a set of N normalized Q-tables $\{\mathbf{Q}^{(1)}_{\text{norm}}, \mathbf{Q}^{(2)}_{\text{norm}}, \ldots, \mathbf{Q}^{(N)}_{\text{norm}}\}$, the aggregated Q-table $\bar{\mathbf{Q}}$ was computed as $\bar{\mathbf{Q}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{Q}^{(i)}_{\text{norm}}$. This aggregation method produces a single Q-table that reflects the average learning patterns across participants while minimizing the influence of individual outliers.

Performance Evaluation Metrics We employed three converging measures of learning.

Hamming Distance from Optimal Policy The Hamming distance measures the divergence between the learned policy and the theoretically optimal policy. For each non-terminal state s, we extracted the learned action a_{learned} from the Q-table using greedy selection. Let $Q(s,\text{left})$ and $Q(s,\text{right})$ denote the Q-values for moving left and right from state s, respectively. The learned action is determined as:

$$a_{\text{learned}}(s) = \begin{cases} \text{left} & \text{if } Q(s,\text{left}) > Q(s,\text{right}) \\ \text{right} & \text{if } Q(s,\text{right}) > Q(s,\text{left}) \end{cases}$$

States where $Q(s,\text{left}) = Q(s,\text{right})$ were treated as errors in policy alignment. The Hamming distance H was calculated as:

$$H = \frac{\sum_{s \in S_{\text{movable}}} \mathbb{I}[a_{\text{learned}}(s) \neq a_{\text{optimal}}(s)]}{|S_{\text{movable}}|}$$

where S_{movable} denotes the set of non-terminal states (states with type F or S), $a_{\text{optimal}}(s)$ is the optimal action (left or right) at state s, and $\mathbb{I}[\cdot]$ is the indicator function. Low Hamming distance values indicate that a policy is close to optimal.

Success Rate via Simulation We evaluated the practical performance of learned Q-tables by simulating agent behavior in the environment. For each Q-table, we conducted 100 independent simulation runs starting at state S with maximum 50 steps per episode, a discount factor $\gamma = 0.95$, and an ε-greedy policy with $\varepsilon = 0.1$. This approach balances exploitation of learned Q-values with continued exploration, reflecting realistic agent behavior. Performance metrics included the success rate (percentage of episodes reaching the goal G).

Distance from True Q-Values We computed theoretical optimal Q-values using value iteration with the Bellman equation:

$$Q^*(s,a) = R(s,a) + \gamma \max_{a'} Q^*(s',a')$$

where $R(s,a)$ is the immediate reward for taking action a (where a can be left or right) in state s, and s' is the resulting state. The learned Q-table $\mathbf{Q}$ was compared against the optimal Q-table $\mathbf{Q}^*$:

$$\text{MSE} = \frac{1}{M}\sum_{s,a}(Q(s,a) - Q^*(s,a))^2$$

where Q-RMSE $= \sqrt{\text{MSE}}$ and M is the number of non-terminal state-action pairs. This quantifies how closely the learned Q-values approximate the theoretical optimum.

Aggregated results

Figure 2 displays visualizations of the Q-tables for participants trained under different conditions, while Table 1 presents the results of the normalized aggregated Q-tables under these conditions across three evaluation methods.

Individual and round results

To contextualize the learning challenge, we first evaluated an autonomous RL agent without human intervention in 50 runs. This baseline agent achieved a success rate of 0.0% and a Hamming distance of 0.8 (Table 1), confirming that the environment was sufficiently challenging to necessitate external

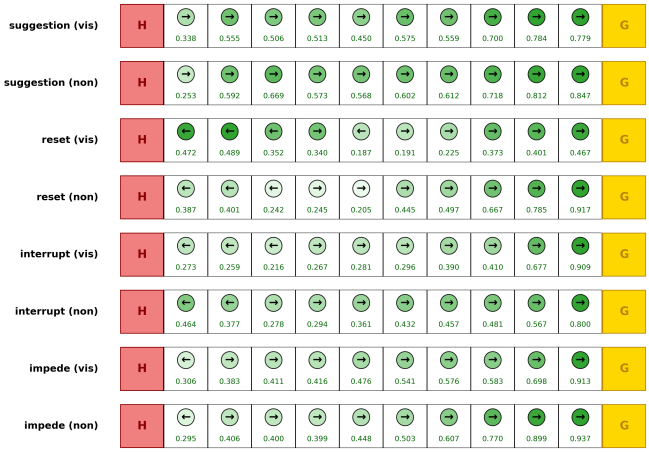

Figure 2: Normalized aggregated Q-tables for agents trained by participants under different conditions.

Table 1: Evaluation results for different interaction types and Q-table visibility conditions. HD = Hamming Distance, SR = Simulation Success Rate, Q-RMSE = Q-Value RMSE.

Condition	**HD**	**SR**	**Q-RMSE**
Suggestion (vis)	0.0	1.0	0.2362
Suggestion (non-vis)	0.0	1.0	0.2392
Reset (vis)	0.4	0.0	0.3067
Reset (non-vis)	0.3	0.0	0.1548
Interrupt (vis)	0.3	0.01	0.1952
Interrupt (non-vis)	0.2	0.07	0.2171
Impede (vis)	0.1	0.96	0.1847
Impede (non-vis)	0.1	0.94	0.1776
Baseline (No intervention)	*0.8*	*0.00*	*0.4808*

guidance. Crucially, despite our explicit instructions for participants to act as teachers, the resulting pedagogical effort remained consistent across conditions: a Bayesian ANOVA confirmed that total intervention rates were indistinguishable between interpretation types ($M_{suggestion} = 69.1, M_{impede} = 74.0, M_{reset} = 72.4, M_{interrupt} = 64.1$; $F(3,60) = 0.14, p = .94, BF_{01} = 10.1$). This confirms that performance gaps stem from the agent's interpretation rules rather than differential teacher effort.

To analyze how these mechanisms influenced performance, we conducted 4 (interpretation type) $\times$ 2 (visibility) $\times$ 3 (round number) mixed ANOVAs for Hamming distance and Q-RMSE, with random intercepts by participant. For success rate, we utilized a binomial generalized linear mixed model (GLMM) to account for the bounded nature of proportion data and to evaluate learning trajectories over rounds.

Hamming distance How participants taught depended critically on how the agent interpreted their interventions. Interpretation type had a significant effect on policy alignment ($F(3,55.55) = 7.96$, $p < .001$, $\eta^2 = 0.30$): agents in the *suggestion* condition learned policies closest to optimal ($M = 0.20$, 95% CI [0.08, 0.33]), significantly outperforming both *interrupt* ($M = 0.59$; $p = .002$) and *reset* ($M = 0.57$; $p < .001$). *Impede* fell in between ($M = 0.36$) but did not differ significantly from the other conditions. In other words, agents that treated human guidance as pedagogical—either as a recommendation (*suggestion*) or a correction (*impede*)—learned better policies than those that treated it as noise (*interrupt*) or a simple relocation (*reset*).

Participants also improved with practice: performance increased across rounds ($F(2,107.49) = 7.21$, $p = .001$, $\eta^2 = 0.12$). However, showing participants the agent's Q-table did not help ($F(1,56.18) = 0.62$, $p = .44$), and there were no significant interactions. The advantage of pedagogically-aligned interpretations held regardless of what teachers could see.

Success rate We analyzed success rate using a binomial GLMM with interpretation type, round, visibility, and their interactions, with random intercepts by participant.

We found significant main effects of interpretation type ($\chi^2(3) = 42.12$, $p < 0.001$) and round ($\chi^2(1) = 200.60$, $p < 0.001$), but not visibility ($\chi^2(1) = 0.63$, $p = 0.43$). Critically, we observed a significant interpretation type $\times$ round interaction ($\chi^2(3) = 285.18$, $p < 0.001$), indicating that learning trajectories differed across conditions (Figure 3). To test whether teaching improved over rounds, we extracted condition-specific slopes from the model. Three conditions showed significant improvement: *suggestion*, *impede*, and *interrupt* ($b = 0.63, 0.91,$ and 2.39, respectively, all $p < .001$). *Reset* showed no improvement ($b = 0.08, p = .22$), remaining near floor across all rounds. Interrupt showed improvement from round 1 to 2, but not from round 2 to 3. It was the only condition with incomplete data (n = 10, 7, 9 for rounds 1–3). Thus, it should be interpreted with caution.

Q-RMSE We did not find any significant main effects of intervention type ($F(3,54.58) = 1.33$, $p = 0.27$, $\eta^2 = 0.07$), visibility ($F(1,55.53) = 1.87$, $p = 0.18$, $\eta^2 = 0.03$), or round ($F(2,107.67) = 0.07$, $p = 0.93$, $\eta^2 = 0.001$), on how closely learned Q-values matched optimal values, and no significant interaction effects. This null result, despite large differences in success rate, suggests that *what* agents learn matters more than whether their value estimates converge numerically.

Discussion

Computational models of pedagogy have formalized how learners extract information from demonstrations, evaluative feedback, and advice. The current study extends this literature by examining state intervention—physically relocating a learner—as a distinct teaching modality. Our results show that state intervention is highly effective, but only when the learner's update rules align with the teacher's pedagogical intent.

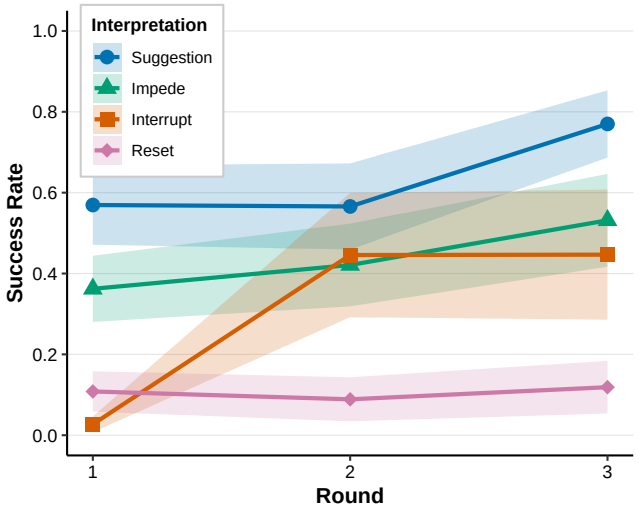

Figure 3: Success rate trajectories across rounds by interpretation type. Points show observed means; shaded regions show ± 1 SE. *Suggestion* and *impede* show steady improvement, while *interrupt* shows gains only from round 1 to 2. *Reset* remains near floor throughout.

The *suggestion* and *impede* conditions shared a key property: both translated physical guidance into reward signals. When a participant dragged the agent toward the goal, *suggestion* treated this as positive evidence for the action leading in that direction. When a participant dragged the agent away from the hole, *impede* treated this as negative evidence for the action that led there. These interpretations align with how humans naturally use physical guidance—as a way of saying "do this" or "don't do that." *Reset* and *interrupt* worked differently. *Reset* treated interventions as pure state changes, learning from the agent's original action as if nothing communicative had occurred. *Interrupt* ignored interventions entirely. Without explicit instruction about these mechanisms, participants had no way to teach effectively—their interventions simply did not mean what they intended.

Contrary to our predictions, showing participants the agent's Q-table did not improve teaching effectiveness. We propose that the marginal utility of such visualization was low because the task was inherently intuitive. In this one-dimensional navigation environment, participants could readily infer the agent's policy simply by observing its behavior and the visualized intent arrow. Because the goal-directed nature of the task allowed for natural mental modeling, the abstract numerical values in the Q-table provided little actionable information beyond what was already evident through interaction. This null effect suggests that for intuitive tasks, aligning the agent's learning rules with human expectations may be more critical than providing technical transparency, as high-level explanations do not always lead to better collaborative outcomes and can even hinder performance calibration (Bansal et al., 2021).

The consistency of intervention rates across all conditions ($BF_{01} = 10.1$) is particularly revealing. It indicates that the failure of certain agents was not due to a lack of participant motivation or effort, but rather a fundamental *interpretational misalignment*. Participants in the *reset* and *interrupt* conditions continued to provide guidance, likely assuming their interventions were being understood, yet the agents' internal logic was deaf to these communicative acts.

These findings have implications for the design of robots that learn from touch (Villani, Pini, Leali, & Secchi, 2018). Our results suggest that such corrections are only effective if the robot interprets them as intentional signals rather than noise (Losey et al., 2022). Rather than requiring users to learn the idiosyncrasies of machine learning algorithms, agents should be designed to treat physical guidance as pedagogical by default.

In conclusion, effective human-robot teaching requires agents to be "socially-aware learners." The divergent learning trajectories we observed highlight that the semantic interpretation of an intervention determines not just whether learning occurs, but how it unfolds over time. Future work should focus on agents that can dynamically adapt their interpretation of human interventions to match the user's implicit teaching strategy, ensuring more robust and natural cooperation.

Limitations and Future Work

Several limitations of the current study warrant further consideration. First, our sample size was modest ($N = 64$), which may have limited our statistical power to detect more subtle interactions, particularly between interpretation types and Q-table visibility. Second, the 1D grid environment was intentionally minimalist to ensure experimental control. While this allowed for precise measurement of policy alignment, future research should investigate whether these interpretational effects generalize to higher-dimensional tasks where physical guidance might be more ambiguous. Third, our study focused on the learner's interpretation in isolation, as we did not manipulate teachers' understanding of the learners' interpretations. Participants were never informed of their agent's specific interpretation type, capturing a scenario where users lack insight into an agent's internal logic. Finally, we only tested model-free Q-learning agents. Model-based agents, which propagate information through learned transition dynamics, might show different patterns of adaptation and could potentially benefit more from information transparency tools like Q-table visualizations.

Now that we have developed an experimental paradigm to study teaching with state interventions, there are numerous promising directions for future research. For instance, future work could explore a wider range of interpretation rules and agent types. Future work could also more thoroughly examine the teacher's perspective by providing different pedagogical instructions or explicitly priming teachers with specific mental models of the learner's mechanism.

Acknowledgments

This research originated as a group project at COSMOS 2025, which included the hard work of Ke Fang. We sincerely thank the COSMOS 2025 organizers, Charley Wu and Wataru Toyokawa, for their guidance and support. This work was supported by funding from the RIKEN CBS-Toyota Collaboration Center (RIKEN BTCC). Additionally, 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/simplified-state-interventions.git`.**

References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on machine learning* (p. 1).

Argall, B. D., Chernova, S., Veloso, M., & Browning, B. (2009). A survey of robot learning from demonstration. *Robotics and autonomous systems*, *57*(5), 469–483.

Bajcsy, A., Losey, D. P., O'Malley, M. K., & Dragan, A. D. (2018). Learning from physical human corrections, one feature at a time. In *Proceedings of the 2018 acm/ieee international conference on human-robot interaction* (pp. 141–149).

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., ... Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 chi conference on human factors in computing systems* (pp. 1–16). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3411764.3445717

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330.

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

Griffith, S., Subramanian, K., Scholz, J., Isbell, C. L., & Thomaz, A. L. (2013). Policy shaping: Integrating human feedback with reinforcement learning. In *Advances in neural information processing systems* (Vol. 26).

Gweon, H. (2021). Inferential social learning: How humans learn from others and help others learn. *Trends in Cognitive Sciences*.

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.

Hoffman, G., Bhattacharjee, T., & Nikolaidis, S. (2024). Inferring human intent and predicting human action in human–robot collaboration. *Annual Review of Control, Robotics, and Autonomous Systems*, *7*.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1-2), 99–134.

Knox, W. B., & Stone, P. (2009). Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on knowledge capture* (pp. 9–16).

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.

Losey, D. P., & O'Malley, M. K. (2018). Including uncertainty when learning from human corrections. *arXiv preprint arXiv:1806.02454*.

MacGlashan, J., Ho, M. K., Loftin, R., Peng, B., Wang, G., Roberts, D. L., ... Littman, M. L. (2017). Interactive learning from policy-dependent human feedback. In *International conference on machine learning* (pp. 2285–2294).

Marchal-Crespo, L., McHughen, S., Cramer, S. C., & Reinkensmeyer, D. J. (2010). The effect of haptic guidance, aging, and initial skill level on motor learning of a steering task. *Experimental Brain Research*, *201*(2), 209–220.

Rafferty, A. N., Brunskill, E., Griffiths, T. L., & Shafto, P. (2016). Faster teaching via POMDP planning. *Cognitive Science*, *40*(6), 1290–1332.

Schmidt, R. A. (1991). *Motor learning and performance: From principles to practice*. Human Kinetics.

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).

Villani, V., Pini, F., Leali, F., & Secchi, C. (2018). Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *Mechatronics*, *55*, 248–266. doi: 10.1016/j.mechatronics.2018.02.009

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.