# Empirical Evidence and Analysis of a Critical Pitfall in Reward Learning from Human Feedback

**Taha Shaheen , Stephen G. West , Yu Zhang**
Arizona State University
{taha.shaheen, sgwest, yzhan442}@asu.edu

## Abstract

Reward learning via human feedback is a crucial capability for beneficial AI. Current methods are built on decision-making theories that assume a matched dynamics model between the learning agent and the feedback provider. However, humans often form imperfect internal dynamics models, and their feedback reflects these misconceptions. While this relationship has long been hypothesised, its manifestation in sequential decision-making remains largely an assumption. Our work provides the first comprehensive empirical investigation of this relationship through a randomized controlled trial ($N = 211$). We followed a two-stage design where we first initialized the participants' understanding of the dynamics in a grid-world navigation domain and then manipulated it using text-based instructions. Causal mediation analysis revealed that humans' internal models play a mediating role in feedback behaviour. We show that this relationship is invariant across visual contexts and is robust to three common feedback types: pairwise preferences, trajectory corrections, and off-switch interventions. These findings confirm a critical limitation of current reward learning methods and establish the missing psychological foundation for approaches that incorporate dynamics understanding.

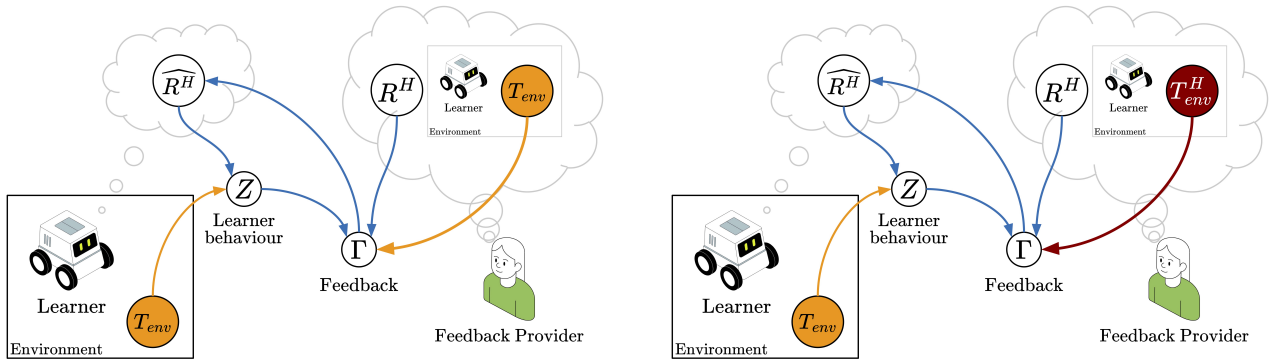
## 1 Introduction

Value alignment addresses the challenge of guaranteeing that an agent's objectives and actions align with human intentions [Gabriel and Ghazavi, 2023]. This is critical for human-AI applications because, as AI systems become more autonomous, the risks of misalignment increase with the complexity of the AI system. Fully specifying the desirable agent behaviour remains difficult; it is not always possible to manually produce reward functions that encapsulate every desired constraint. As a result, reward learning is becoming an important class of methods for value alignment that aim to learn the reward function for decision-making problems modelled as Markov Decision Processes (MDPs).

Reward learning is generally not well defined and suffers from non-identifiability [Armstrong and Mindermann, 2018]. To alleviate its impact, standard frameworks like Inverse Reinforcement Learning (IRL) [Ng and Russel, 2000], and generally Reinforcement Learning methods from Human Feedback (RLHF) [Kaufmann *et al.*, 2024], typically assume the human feedback provider or user is rational or noisily-rational at maximizing the expected accumulative reward. Crucially, these formulations assume that the user and AI agent (learner), use an identical dynamics or transition model, i.e., $T^H_{env} \equiv T_{env}$ in Fig. 1. It implies that the learner interprets the human's feedback to be based on the ground-truth dynamics model. The reward function is then often recovered using disambiguating criteria such as Maximum Entropy or Bayesian optimization [Ziebart *et al.*, 2008; Ramachandran and Amir, 2007].

Unfortunately, this assumption can cause issues when there are differences between the human's understanding of the dynamics model and the ground truth (Fig. 1). In particular, when the user's internal model does not match with the reality ($T^H_{env} \neq T_{env}$), prior reward learning formulations that infer only a latent reward are forced to absorb this mismatch into the learned reward function. Behavioural deviations caused by internal model misconceptions are explained away as desired behaviour or as noisy reward signals. For example, if users avoid a safe shortcut because they falsely believe it is dangerous, the agent may mistakenly learn that longer paths are preferred over short-distance travel. This confounding of human beliefs with environmental reality makes the learned reward signal a composite of true intent and false beliefs [Gong and Zhang, 2020]. However, this complication in sequential decision-making is largely hypothesized and, to our knowledge, there is no empirical study that validates and analyses its presence. The key lies in validating a strong relationship between humans' understanding of the dynamics model and their feedback.

In this paper, we empirically investigate whether the human's understanding of dynamics influences their feedback behaviour. We also check whether this relationship is independent of the feedback type and any initial understanding formed from visual cues. However, isolating the influence of user understanding on feedback posed significant design challenges. A user's mental model may change during the study (e.g., through observations) and noise in both understanding and feedback can confound the analysis. Towards this end, we devised a series of planning games lacking plan execution.




(a) Incorrect View (Traditional)

(b) Our view

Figure 1: Comparison of reward learning views. The learner agent updates its estimate $\widehat{R^H}$ using feedback $\Gamma$ on its behaviour $Z$. (a) Matched-Dynamics Assumption: The learner interprets feedback $\Gamma$ as reflection of the human's latent reward function $R^H$ under true environment dynamics, i.e., ($T^H_{env} \equiv T_{env}$). (b) Our View: Feedback $\Gamma$ is a composite signal derived from both the human's true reward $R^H$ and their (possibly incorrect) dynamics beliefs $T^H_{env}$.

In a randomized controlled experiment ($N = 211$) we implemented a two-stage understanding manipulation: initializing ($T^H_{env}$); and then manipulating text-based instructions to update participants' internal models. This design enabled the use of a causal mediation analysis [MacKinnon and Tofighi, 2012; Valente *et al.*, 2020]. To counter the noise, we used baseline measures of behaviour and understanding as covariates. Our contributions are three-fold: (1) We empirically investigate whether user feedback behaviour is affected by the user's internal model of domain dynamics $T^H_{env}$ in sequential decision-making using a carefully structured study design and mediation analysis; (2) We demonstrate that this relationship is consistent across three distinct feedback types: preferences, corrections, and off-switch interventions; (3) Our work establishes the psychological foundation for reward learning approaches that incorporate user understanding and confirms a critical limitation of existing methods.

## 2 Related Work

Literature separately considers feedback providers that evaluate agent behaviour without interacting, and feedback providers that interact with the agent [Hadfield-Menell *et al.*, 2016; Reddy *et al.*, 2018; Nikolaidis and Shah, 2013]. We focus on the former setting, with results conceivably generalizable to the latter (see Sec. 7).

### 2.1 Irrational Feedback Providers

Cognitive science and behavioural economics suggests that human feedback is generated from simplified and/or biased internal models rather than models optimized over ground truth dynamics. Literature on Bounded rationality [Simon, 1955], Prospect Theory [Kahneman and Tversky, 1979], and Preference Construction [Slovic, 2020] indicates that human judgement relies on heuristics and context. Furthermore, humans employ "Intuitive Physics" mental models that systematically deviate from the truth [Battaglia *et al.*, 2013; Smith and Vul, 2013].

Boltzmann rationality is a behavioural model that relaxes the assumption of perfect optimality by treating human choices as noisily-optimal. Deviations get modelled as stochasticity in the human's action selection process. This noisy-rationality underlies many learning methods [Wirth *et al.*, 2017; Kaufmann *et al.*, 2024]. Such models incorporate human suboptimality (formalized as a form of bounded rationality) into a single rationality parameter [Jeon *et al.*, 2020] and, by extension, into the learned reward function. Recent critiques of this approach argue that human suboptimality is not random noise but reflects structural mechanisms. For instance, [Bobu *et al.*, 2020] show that standard models overcount similar trajectories as separate choices while humans actually view them as a single option. Similarly, [Baker *et al.*, 2011] demonstrate that humans interpret inefficient actions not as random errors, but as rational planning based on a mistaken understanding of the environment.

### 2.2 Traditional Reward Learning Methods

Frameworks reviewed in the survey by [Kaufmann *et al.*, 2024] model human preferences as noisy or stochastic, yet implicitly assume $T_{env} \equiv T^H_{env}$. Traditional IRL formulations assume expert demonstrations are drawn from an optimal policy within the true MDP [Ng and Russel, 2000; Abbeel and Ng, 2004]. Maximum Entropy approaches relax the assumption of strict optimality to stochastic policies. They attribute deviations from optimality to noise in action selection [Ziebart *et al.*, 2008; Gleave and Toyer, 2022]. Deep RL methods [Ho and Ermon, 2016; Ibarz *et al.*, 2018; Christiano *et al.*, 2017] similarly assume that the learner's behaviour is evaluated by the human under the environment dynamics. Often, approaches highlighted above do not recognize that the matched dynamics assumption is being made.

### 2.3 Unknown or Mismatched Dynamics

A growing body of literature highlights the risks of assuming $T_{env} \equiv T^H_{env}$. [Armstrong and Mindermann, 2018] suggest that deviations in human feedback may stem from incorrect

internal beliefs rather than human's latent reward. Observing feedback alone is often insufficient to distinguish the source of these deviations. This can lead to flawed reward inference and performance degradation in standard IRL [Skalse and Abate, 2023; Viano *et al.*, 2021]. [Casper *et al.*, 2023] label reliance on idealized human models a limitation in RLHF.

Some literature learns the latent dynamics model. [Herman *et al.*, 2016] jointly learn the reward and transition function under the assumption that the human teaches the agent the correct dynamics. [Gong and Zhang, 2020] go a step further to simultaneously estimate the human's internal transition model and reward function via variational inference under the assumption of mismatched human dynamics understanding. They show that ignoring latent biased beliefs can lead to inferring preferences that are opposite to the ground truth. The authors [Gong and Zhang, 2022] further show that the mismatched dynamics can be captured by a surrogate reward function for explicable planning [Zhang *et al.*, 2017].

Closest to our study is [Dandekar *et al.*, 2025]. They show that human preference labels are shaped by belief of agent capabilities. In contrast, we study belief mismatch in the dynamics model and demonstrate that altering participants' beliefs about dynamics model systematically changes the feedback they provide. Furthermore, while this work focused on preferences, we show that this relationship exists under various forms of feedback.

Our study complements this prior body of work. While previous literature acknowledges the presence of dynamics mismatch, its impact on user feedback is typically considered as an established fact in sequential decision-making. Consequently, the internal model is treated as a fixed initialization or a static latent variable. Our work moves beyond postulating this relationship to empirically establishing its existence and effect size. We show that the human's internal model plays a mediating role: in particular, that instructions distort feedback via dynamics understanding across multiple feedback types.

# 3 Hypotheses

To test if feedback is mediated by the user's internal dynamics model, we utilize a causal mediation framework [MacKinnon and Tofighi, 2012].

**H1:** Internal dynamics understanding mediates the relationship between instructional framing and feedback.

- **(H1a)** Instructions change dynamics understanding.
- **(H1b)** This change in understanding predicts the change in feedback behaviour.

**H2:** The mediation is robust to feedback modality and remains consistent across three feedback types.

**H3:** The mediation is robust to visual priming and persists across two visual contexts.

# 4 Methodology

## 4.1 Design

To isolate the effect of the participant's domain dynamics understanding from their latent reward function, it was crucial that we only manipulate their understanding while holding the reward constant. To do this, we administered the domain dynamics manipulation in two stages: (1) initialization using visual priming, and (2) manipulation using text-based instructions. We chose this because visuals establish an initial mental model and text allows for explicit updates to this model [Powell *et al.*, 2015]. In contrast, the latent reward function was initialized both visually and textually and held constant throughout.

A concern was that mental models can change when observing an agent actions [Hanni and Zhang, 2021] in the environment. To prevent domain understanding from drifting, we framed the problem as a planning without execution task. Participants were asked to help plan a course in an extended cliff walking domain for an agent, "Elfie," prior to a journey. The execution of this journey was never shown and participants never observed the agent slip or succeed. We measured baseline understanding and behaviour as covariates to control for bias.

## 4.2 Platform

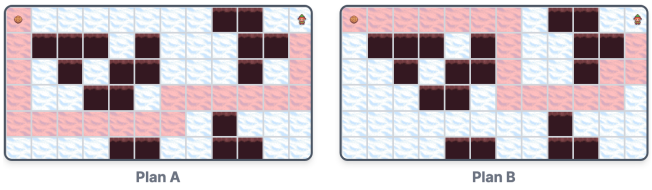
Data were collected in form of state-action trajectories via a series of online mini-games[1]. Participants helped a character named Elfie the Elf develop a plan to go from a starting position to a goal position to retrieve a cookie. We generated 20 maps ($6 \times 12$ grid-worlds) with 12 to 25 hole states. Start and goal states were at opposing corners of the same row. For each map, we solved two MDPs ($\gamma = 0.9$; $R_{step,cookie,hole} = \{-1, 10, -100\}$): (1) a deterministic stable model, and (2) a stochastic slippery model where tiles adjacent to holes had a slip probability of $p = 1/3$ directed towards hole. To ensure heterogeneity, maps were created such that optimal policies matched in the initial 3-steps but maximized path divergence with Dynamic Time Warping (DTW) [Sakoe and Chiba, 2003] distance within 50.0 of the straight-line vs. perimeter-hugging paths DTW distance. We used the `dtw-python` package [Giorgino, 2009].

**Mini-Games (Feedback Types)**

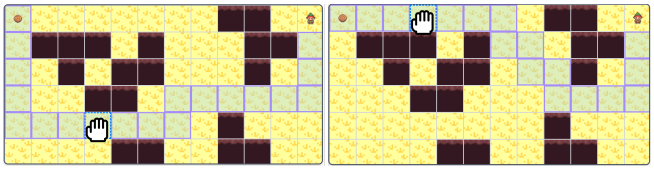
We implemented four feedback types.

1. **Comparison Game.** Modelled after Preference-based RL [Wirth *et al.*, 2017], participants picked one of two trajectories noisily sampled from $\pi_{stable}$ and $\pi_{slippery}$ for the same map (Fig. 2a).
2. **Correction Game.** Inspired by methods where a robot's trajectory is physically adjusted [Bajcsy *et al.*, 2017], participants corrected a pre-planned trajectory by dragging points on the trajectory (Fig. 2b). To prevent blind acceptance, 3 of the 5 pre-planned paths conflicted with the visual context.
3. **Intervention (Off) Game.** This was a variation of the Off-Switch feedback [Hadfield-Menell *et al.*, 2017], that allowed iterative stopping and complete trajectories.

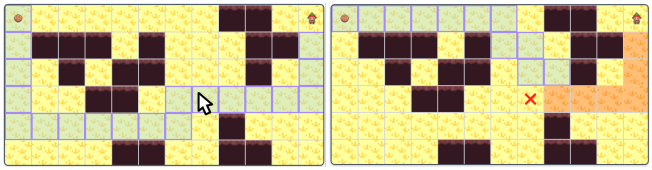
[1]Readers may play the games at https://elfie-cliff-walking.onrender.com/. Log in using `test_user1`, `test_user2`, `test_user3`, or `test_user4` to experience different randomizations. Code and anonymised experiment data can be found at: https://github.com/tahaShaheen/elfie-cliff-walking.



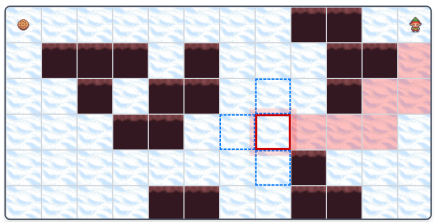

(a) Comparison Game (Slippery Context)



(b) Correction Game (Stable Context)



(c) Off Game Interface (Stable Context)



(d) Demonstration Game (Slippery Context)

Figure 2: The four Mini-Games used in the study: (a) participants select preferred trajectories, (b) participants drag points to correct plans, (c) participants place obstacles to trigger replanning, (d) participants create trajectories by selecting tiles.

Participants could place up to 5 obstacles (Fig. 2c), forcing the agent to preserve the path up to the intervention and replan from there. As with Corrections, 3 of the 5 initial trajectories conflicted with the visual context.

4. **Demonstration Game.** Based on Learning from Expert [Ng and Russel, 2000], participants constructed a trajectory by selecting tiles (Fig. 2d). We used this as a covariate rather than a feedback type. This was done for two reasons: (1) it is a common, generalizable feedback type, and (2) as a generative task it served as a platform tutorial without introducing any bias from observing Elfie's plans.

## 4.3 Procedure

The study was implemented in two stages (Fig. 3).

**Stage 1: Initializing Understanding**

**Visual Priming.** Participants were briefed on the path-planning task to initialize their latent goal understanding (reach cookie, avoid cliffs). Participants were explicitly instructed that the mini-games were planning-only; they would never see Elfie move. All participants were exposed to both visual contexts: (1) Stable (dry, yellow grass) and (2) Slippery (ice and snow) (see Fig. 4).

**Baseline Assessment.** Participants were randomly assigned to one Visual Context ($X_1$). We then collected baseline measures to serve as covariates:

- **Baseline Understanding ($C_1$):** Participants viewed two plans on the same map (Fig. 5): one away from cliff edges (always safe check question) and the other adjacent (dynamics understanding question). They rated the likelihood of slipping for each plan on a 5-point Likert scale (1 = Least, 5 = Most Likely). $C_1$ was operationalized as the mean of the dynamics understanding question rating (1 to 5) and the inverted reverse-coded check question. This composite score is a discriminative measure where high values indicated rating the cliff edge as dangerous, mid-range values as rating both plans equally, and low values inversely rating the safe zone as dangerous. Participants were excluded if they failed three attempts to answer two comprehension questions designed to ensure they understood they would be evaluating plans with no execution.
- **Baseline Behaviour ($C_2$):** Participants played a Demonstration Game (Fig. 2d). See Sec. 4.3 for its measurement strategy.

**Stage 2: Manipulation of Understanding**

**Manipulation.** Participants were randomly exposed to one of two text-based Instructions ($X_2$) with "We have received more information!" followed by either the safe or the danger instruction:

- **Safe Instruction:** "Relax: Stable Ground! In this realm, the ground is made of safe and stable tiles. You don't have to worry about accidentally slipping.
  - On stable ground, cliff edges are completely safe.
  - Elfie will not accidentally slip if her plan takes her past an edge.
  - Elfie is always safe on tiles not adjacent to cliffs.
  - Diagonal movement is not possible."
- **Danger Instruction:** "BE CAREFUL! The ground is made of extremely slippery tiles that slope towards the cliffs.
  - On slippery ground, cliff edges are dangerous.
  - Elfie could accidentally slip off a cliff if her plan takes her past an edge.
  - Elfie is always safe on tiles not adjacent to cliffs.
  - Diagonal movement is not possible."

Text-based instructions can result in compliance or experimenter demand effects where participants merely comply with directives. To mitigate this, we framed the instructions as descriptions of how the environment behaves (e.g., risk of slipping) instead of commands what to do (e.g., 'choose the longer path').

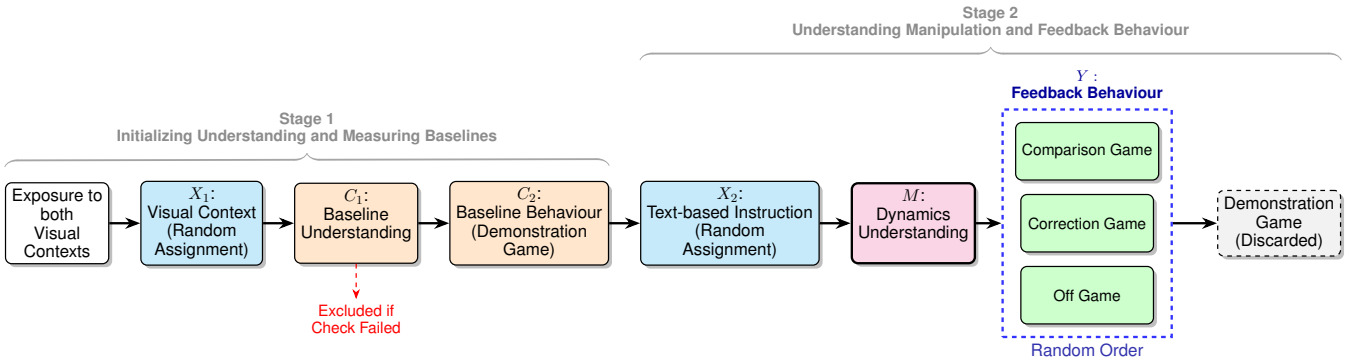



Figure 3: Experimental Procedure Flow Diagram showing the two stage design.

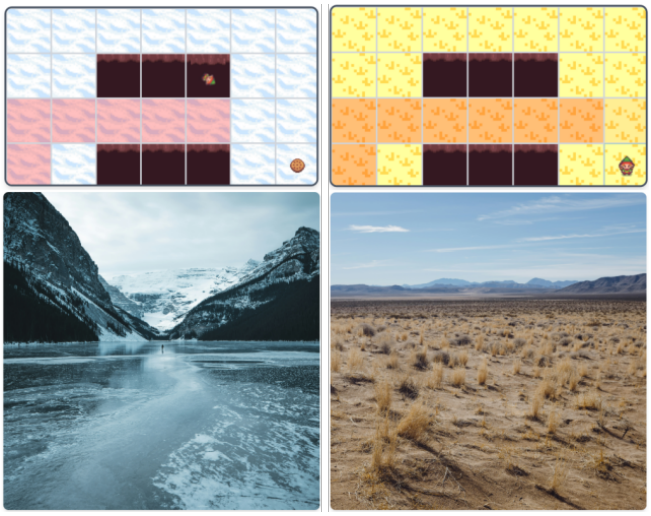


Figure 4: Visual contexts (Left: Icy/Slippery; Right: Grassy/Stable). Top-down maps represent what participants saw in the mini-games while close-up images (from [Unsplash, 2026]) provided details to create a mental model of the dynamics. This is the only time when Elfie moved visually.

**Mediator Measurement.** Participants repeated the risk assessment from the baseline phase (Fig. 5). However, unlike $C_1$, the Mediator ($M$) was derived only from the standardized score of the cliff-adjacent plan. $C_1$ required both plans to establish the participants' baseline ability to discriminate between safety and danger. $M$, on the other hand, focused on the shift in perceived risk induced by the intervention. We excluded the safe-zone plan from this calculation to isolate the relevant signal.

**Feedback Behaviour.** Participants played 3 games (Correction, Comparison, Off) in a randomized order, each consisting of a practice map and 5 experimental rounds. The final non-randomized Demonstration game was excluded to prevent order effects. We derived a single continuous measure of feedback for each type using

$$Y_{raw} = \frac{\bar{d}_{stable} + (125 - \bar{d}_{slippery})}{2},$$

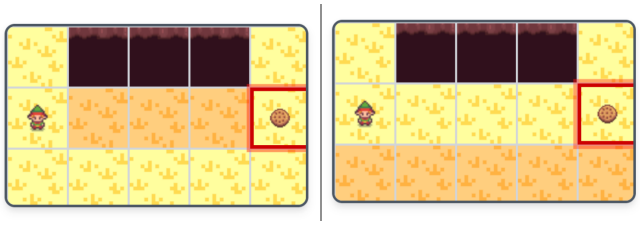


Figure 5: For $C_1$ and $M$, participants view two plans: one close to the cliff edges and the other away from the cliff edges. They rated likelihood of slipping for each plan on a 5-point Likert scale (1 = Least, 5 = Most Likely).

where $\bar{d}_{stable}$ and $\bar{d}_{slippery}$ are the 5-round average DTW distances between participant trajectories and trajectories sampled from $\pi_{stable}$ and $\pi_{slippery}$ respectively for each map. The constant 125 is the DTW distance between the straight-path and perimeter-hugging paths (calculated via `dtw-python`, `step_pattern=symmetric2` [Giorgino, 2009]). High values indicate proximity to paths under slippery rather than stable dynamics understanding. Scores were Z-standardized across the sample to form the final outcome variable $Y$.

**Randomized Assignment.** Participants were randomized into one of four conditions (2 Visuals $\times$ 2 Instructions) using a block randomization with dropout replacement. The order of Correction, Comparison, and Off games was randomized with the Demonstration games fixed at the beginning and the end. Every participant saw all 20 maps in a random order (5 per game).

## 5 Data Analysis and Results

Analysis was carried out with Statistical Package for the Social Sciences (SPSS) [IBM Corp., 2023].

**Participants.** Following two pilots (N=19, 14), the final sample size was N=211 valid records. Recruited via university pools (66.3%, granted course credit) and online ads (compensated $5 each), the group was composed of undergrads (81.4%), graduate students (15.3%), and professionals (3.3%). Most were Arizona State University affiliates (97.2%) and physically present in AZ, USA (98.1%), majoring mainly in business (34.0%) and Computer Science

(25.6%). Participants (ages 18–50; M=20.33, SD=3.98; 53% male, 47% female, 0% other) largely lacked prior AI/RL familiarity (70.7%).

**Descriptive Stats.** The total duration of the experiment averaged 25.22 minutes ($SD = 10.07$, Mdn $= 22.67$, Range: 9.87 to 71.75). Among the feedback types, the Off Intervention took the longest to complete ($M = 5.42$ min, $SD = 3.53$), while the final Demonstration took the shortest amount of time ($M = 2.71$ min, $SD = 1.46$).

**Multilevel Mediation.** We conceptualized our study as a 2-2-1 multilevel mediation design [Zhang *et al.*, 2009] (see Fig. 6). $X_2$ (Safe vs. Danger) and $M$ varied between participants (2 levels) while Feedback Behaviour $Y$ was a repeated outcome across three feedback trials nested within participants (1 level). Since participants were split along Visual Context ($X_1$, Grassy vs. Icy), we adopted a split-sample strategy similar to that employed by [Jensen-Campbell *et al.*, 1995]. Consequently, we estimated the mediation separately for each visual context ($X_1$). For Path $a$ ($X \rightarrow M$), we used a univariate general linear model to regress Dynamics Understanding on Instructions while controlling for baseline covariates. For Path $b$ ($M \rightarrow Y$) and the direct effect $c'$ ($X_2 \rightarrow Y$), we fit a repeated-measures mixed model to predict Feedback from Understanding and Instructions (with trial position and baseline covariates as controls). We also tested whether these effects differed by feedback type by including the interaction terms $X_2\times$ Feedback Type and $M\times$ Feedback Type.

**Path $a$.** Across both visual contexts, $X_2$ significantly predicted $M$, controlling for $C_1$ and $C_2$. In the Stable context, the effect of $X_2$ on $M$ was large and statistically significant ($\hat{a} = -1.326$, $SE = 0.123$, $t = -10.823$, $p < .001$). In the Slippery context, the manipulation effect was also significant but smaller in magnitude ($\hat{a} = -0.918$, $SE = 0.168$, $t = -5.451$, $p < .001$). $C_1$ was positively related to $M$ in both contexts (Stable: $B = 0.376$, $SE = 0.086$, $p < .001$; Slippery: $B = 0.198$, $SE = 0.097$, $p = .045$), whereas $C_2$ was not statistically significant in either context (Stable: $B = 0.061$, $p = .431$; Slippery: $B = 0.173$, $p = .115$). $C_1$ functioned as a reliable covariate for predicting $M$, while $C_2$ did not.

**Paths $b$ and $c'$.** In the Stable context, $M$ significantly predicted $Y$ ($\hat{b} = 0.448$, $SE = 0.111$, $t = 4.056$, $p < .001$). Additionally, the direct effect of $X_2$ on $Y$ remained significant after controlling for $M$ ($\hat{c}' = -0.475$, $SE = 0.203$, $p = .021$), indicating that the $X_2$ manipulation influenced $Y$ both indirectly through $M$ and also through additional pathways not captured by $M$. The same pattern held in the Slippery context. $M$ significantly predicted $Y$ ($\hat{b} = 0.356$, $SE = 0.085$, $t = 4.191$, $p < .001$), and $X_2$ had a strong direct effect controlling for $M$ ($\hat{c}' = -0.662$, $SE = 0.166$, $p < .001$).

**Trial Order.** To account for potential learning or fatigue, we included Trial Order (i.e., the chronological order of the games) as a control. While non-significant in the Stable context ($F(2, 204) = 0.923$, $p = .399$), Trial Order significantly predicted behaviour in the Slippery context ($F(2, 202) = 3.551$, $p = .030$). Specifically, feedback values were higher in Trial 1 than Trial 3 (estimate $= 0.168$, $SE = 0.064$, $p = .009$).

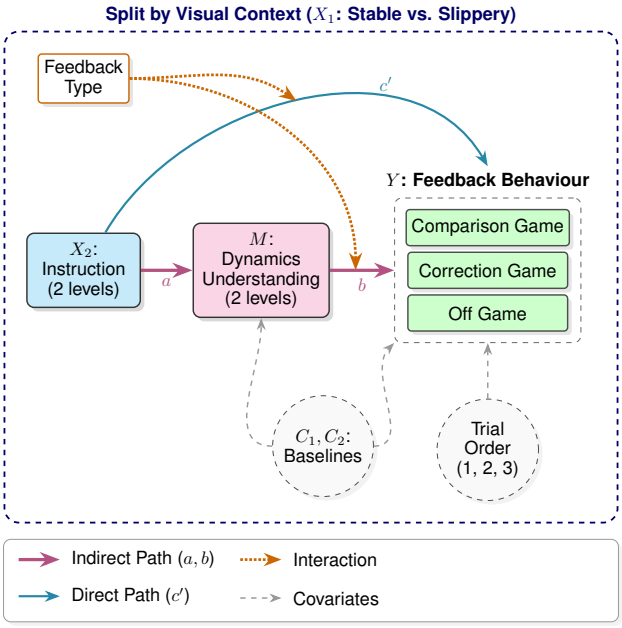



Figure 6: Causal Mediation model. Instructions ($X_2$) predict Dynamics Understanding ($M$), which predicts Feedback Behaviour ($Y$). Feedback Type is included as a moderator of the $b$ and $c'$ paths. The model is split by Visual Context ($X_1$).

**Feedback Type (Direct & Moderating Effects).** We found no main effect of feedback type on behaviour in either context (Stable: $p = .339$; Slippery: $p = .227$). Neither the direct effect of instructions ($X_2 \rightarrow Y$) nor the mediator relationship ($M \rightarrow Y$) was moderated by feedback type in the Stable (all $p \geq .428$) or Slippery (all $p \geq .120$) contexts.

**Indirect Effect Estimation.** To test the mediation (indirect effect $a \times b$), we used the Monte Carlo Method for Assessing Mediation (MCMAM) [Preacher and Selig, 2012]. We drew 20,000 random values of the $a$ and $b$ coefficients from distributions centred on their estimates (with spread given by their standard errors), computed $ab$ for each draw, and used the resulting distribution of $ab$ values to form a 95% confidence interval. In the Stable context, the estimated indirect effect was $ab = -0.593$ with a 95% Monte Carlo CI of $[-0.920, -0.292]$, which excludes zero. In the Slippery context, the indirect effect was also significant ($ab = -0.327$, 95% CI $[-0.547, -0.149]$).

**Visual Context.** To test whether the mediation effect differed across visual contexts, we computed the indirect effect separately in each context and then took their difference ($ab_{\text{diff}} = ab_{\text{stable}} - ab_{\text{slippery}}$) across the 20,000 simulations. The 95% CI for this difference included zero ($ab_{\text{diff}} = -0.267$, 95% CI $[-0.641, 0.102]$), providing no evidence of difference. The average indirect effect across contexts was significant ($ab_{\text{avg}} = -0.460$, 95% CI $[-0.649, -0.283]$), indicating overall consistency.

## 6 Hypothesis Verification

**H1.** Results support the hypothesised mediation. Instructions updated internal models (Path $a$; **H1a**), which significantly predicted behaviour (Path $b$; **H1b**). Significant indirect effects ($ab$) confirmed mediation in both visual contexts, and significant direct paths ($c'$) indicated partial mediation.

**H2.** Non-significant interactions between Feedback Type and mediation paths suport the hypothesis that mediation is consistent across off 3 feedback types.

**H3.** Although effects appeared larger in the Stable context, the difference was not statistically significant (CI included zero). This confirms the hypothesis that the mediation is robust to visual priming.

## 7 Discussion

Our study provides empirical evidence against a common and often implicit assumption in RLHF: that human feedback reflects the human's desired behaviour evaluated under the environment's true dynamics model ($T_{env}$). We observed a change in participants' feedback behaviour by manipulating their understanding of environment dynamics through instructions, even though the goal (reaching a cookie without falling in holes) and the visuals of the environments were held fixed. This supports our central claim, illustrated in Fig. 1, that feedback signals are confounded signals that reflect not just the latent reward function but also the human's dynamics understanding (or misunderstanding) of the environment.

A key finding of this work is that the feedback's dependence on dynamics understanding persists across three feedback types and two visual contexts. Prior work often treats variation in feedback types as differences in cognitive loads or bounded rationality [Ghosal *et al.*, 2023]. While we do not contradict that perspective as participants did spend different amounts of time in different tasks, our findings indicate an additional source of systematic variation. The feedback provider's latent transition function $T^H_{env}$ shapes behaviour across all modes of feedback. In practice, this means that simply switching feedback types (e.g., from preferences to corrections) would not remove the risk of incorrect feedback signals if $T^H_{env} \neq T_{env}$. Furthermore, overriding of visual contexts suggests that dynamics misconceptions drive feedback independently of surface-level visual features.

Ultimately, this work provides the necessary empirical foundation for theoretical approaches that treat the human's internal model as a distinct latent variable [Gong and Zhang, 2020; Viano *et al.*, 2021]. Through this investigation we attempt to move beyond the hypothesis that beliefs influence feedback to demonstrating that they do so. This finding supports the necessity of algorithms that explicitly account for $T^H_{env} \neq T_{env}$.

It is also worth discussing these findings in the context of interactive settings such as Cooperative Inverse Reinforcement Learning (CIRL) [Hadfield-Menell *et al.*, 2016], Shared Autonomy [Reddy *et al.*, 2018], or Human-Robot Cross-Training [Nikolaidis and Shah, 2013]. In these settings, the human and agent interact, co-adapt, and/or learn from each other. This can theoretically mitigate any mismatch in dynamics understanding through iterative correction. However, while methods like this promote mental-model convergence, they remain susceptible to sustained mismatches. If a user exhibits constant, misunderstanding-driven, beliefs (e.g., avoiding a state due to false danger beliefs) and the system does not explicitly model dynamics mismatch, the agent may misidentify this behaviour as a true preference. We argue that, even for interactive methods, it is crucial to explicitly determine if behaviour meant to convey preferences stems from inaccurate dynamics beliefs.

### 7.1 Limitations and Future Work

Our study utilized a simple grid-world environment. However, physical robotic tasks may introduce more complex understanding mismatches not captured here. Additionally, our manipulation of dynamics understanding was explicit and drastic (Safe vs Danger) while dynamics misconceptions in real-world scenarios may be more subtle. In more complex environments, such as autonomous driving, the transition dynamics could be far less intuitive or obvious than discrete grid movement. Although we believe this will carry over to more complex environments, providing such empirical evidence is deferred to future work. Also, since the tasks involved mostly mental planning, a potential source of confound that we did not explore could be a disparity in the perceived difficulty of navigating in danger vs safe environment, i.e., if the cognitive difficulty of task plays a role.

A notable design choice in this study was the restriction of the task to planning without execution. Participants did not observe the agent physically slipping or succeeding. In practice, observing an agent's execution could bridge the gap between $T^H_{env}$ and $T_{env}$. However, we argue that observation is not an end-all-be-all remedy. A risk-averse user could prevent the agent from ever attempting the action that would reveal the true dynamics, preventing the belief update required to align the models. Risk psychology suggests that beliefs often resist contradictory evidence [Slovic, 1987], implying that observation alone may not bridge the gap between $T^H_{env}$ and $T_{env}$. Hence, future work should investigate if observing agent acting in the environment can correct latent misconceptions, or if humans preserve beliefs by attributing unexpected outcomes to other factors. Finally, this manuscript does not propose mitigation strategies. Readers are encouraged to refer to Sec. 2.3 for some existing approaches.

## 8 Conclusions

We have provided, to our knowledge, the first empirical evidence that human feedback behaviour in sequential decision-making is mediated by the human's internal dynamics understanding. Through a randomized controlled trial, we show that manipulating dynamics understanding (while keeping latent reward fixed) systematically shifts feedback behaviour. This persists across visual contexts and remains across three distinct feedback types. These findings confirm a pitfall in current reward learning approaches that assume a matched dynamics model ($T^H_{env} \equiv T_{env}$). Ignoring this mismatch can lead to agents learning confounded rewards. Ultimately, demonstrating this relationship supports a refocusing of value alignment. RLHF approaches must explicitly model the user's latent beliefs with rewards.

## Ethical Statement

This research was approved by the Institutional Review Board of Arizona State University. Participants provided informed consent prior to data collection and were compensated for their time. We identify no ethical issues.

## Acknowledgments

This research is supported in part by the NSF grant 2047186. We also acknowledge support from the 2025 ASU Graduate Student Government JumpStart Grant, and the use of Google's Gemini 2.5 Pro and Gemini 3.0 Pro for assistance with code generation and grammar revision.

## References

[Abbeel and Ng, 2004] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 1, New York, NY, USA, July 2004. Association for Computing Machinery.

[Armstrong and Mindermann, 2018] Stuart Armstrong and Sören Mindermann. Occam's razor is insufficient to infer the preferences of irrational agents. *Advances in Neural Information Processing Systems*, 31, 2018.

[Bajcsy *et al.*, 2017] Andrea Bajcsy, Dylan P. Losey, Marcia K. O'Malley, and Anca D. Dragan. Learning Robot Objectives from Physical Human Interaction. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 217–226. PMLR, October 2017.

[Baker *et al.*, 2011] Chris L Baker, Rebecca R Saxe, and Joshua B Tenenbaum. Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 33:33, 2011.

[Battaglia *et al.*, 2013] Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, November 2013.

[Bobu *et al.*, 2020] Andreea Bobu, Dexter R.R. Scobee, Jaime F. Fisac, S. Shankar Sastry, and Anca D. Dragan. LESS is more: Rethinking probabilistic models of human behavior. *ACM/IEEE International Conference on Human-Robot Interaction*, pages 429–437, March 2020.

[Casper *et al.*, 2023] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J. K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*, September 2023.

[Christiano *et al.*, 2017] Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

[Dandekar *et al.*, 2025] Sylee Dandekar, Shripad Deshmukh, Frank Chiu, W. Bradley Knox, and Scott Niekum. A Descriptive and Normative Theory of Human Beliefs in RLHF. arXiv preprint arXiv:2506.01692, June 2025.

[Gabriel and Ghazavi, 2023] Iason Gabriel and Vafa Ghazavi. The challenge of value alignment: From fairer algorithms to AI safety. In *Oxford Handbook of Digital Ethics*. Oxford University Press, December 2023.

[Ghosal *et al.*, 2023] Gaurav R. Ghosal, Matthew Zurek, Daniel S. Brown, and Anca D. Dragan. The effect of modeling human rationality level on learning rewards from multiple feedback types. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37:5983–5992, 6 2023.

[Giorgino, 2009] Toni Giorgino. Computing and visualizing dynamic time warping alignments in r: the dtw package. *Journal of statistical Software*, 31:1–24, 2009.

[Gleave and Toyer, 2022] Adam Gleave and Sam Toyer. A Primer on Maximum Causal Entropy Inverse Reinforcement Learning. arXiv preprint arXiv:2203.11409, March 2022.

[Gong and Zhang, 2020] Ze Gong and Yu Zhang. What is it you really want of me? generalized reward learning with biased beliefs about domain dynamics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:2485–2492, 4 2020.

[Gong and Zhang, 2022] Ze Gong and Yu Zhang. Explicable policy search. *Advances in Neural Information Processing Systems*, 35:38859–38872, 2022.

[Hadfield-Menell *et al.*, 2016] Dylan Hadfield-Menell, Stuart J. Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[Hadfield-Menell *et al.*, 2017] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 220–227. AAAI Press, 2017.

[Hanni and Zhang, 2021] Akkamahadevi Hanni and Yu Zhang. Active Explicable Planning for Human-Robot Teaming. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 176–180, Boulder CO USA, March 2021. ACM.

[Herman *et al.*, 2016] Michael Herman, Tobias Gindele, Jörg Wagner, Felix Schmitt, and Wolfram Burgard. Inverse Reinforcement Learning with Simultaneous Estimation of Rewards and Dynamics. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 102–110. PMLR, May 2016.

[Ho and Ermon, 2016] Jonathan Ho and Stefano Ermon. Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[Ibarz *et al.*, 2018] Borja Ibarz, Geoffrey Irving, Jan Leike, Shane Legg, Tobias Pohlen, and Dario Amodei. Reward learning from human preferences and demonstrations in Atari. *Advances in Neural Information Processing Systems*, 2018-December:8011–8023, November 2018.

[IBM Corp., 2023] IBM Corp. *IBM SPSS Statistics for Windows, Version 29.0.2.0*. IBM Corp., 2023.

[Jensen-Campbell *et al.*, 1995] Lauri A. Jensen-Campbell, William G. Graziano, and Stephen G. West. Dominance, prosocial orientation, and female preferences: Do nice guys really finish last? *Journal of Personality and Social Psychology: Interpersonal Relations and Group Processes*, 68(3):427–440, March 1995.

[Jeon *et al.*, 2020] Hong Jun Jeon, Smitha Milli, and Anca Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems*, 2020-December, 2 2020.

[Kahneman and Tversky, 1979] Daniel Kahneman and Amos Tversky. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–291, 1979.

[Kaufmann *et al.*, 2024] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A Survey of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*, June 2024.

[MacKinnon and Tofighi, 2012] David P. MacKinnon and Davood Tofighi. Statistical Mediation Analysis. In *Handbook of Psychology, Second Edition*, chapter 25. John Wiley & Sons, Ltd, 2012.

[Ng and Russel, 2000] Andrew Y. Ng and Stuart Russel. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, 6 2000.

[Nikolaidis and Shah, 2013] Stefanos Nikolaidis and Julie Shah. Human-robot cross-training: Computational formulation, modeling and evaluation of a human team training strategy. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 33–40, March 2013.

[Powell *et al.*, 2015] Thomas E Powell, Hajo G Boomgaarden, Knut De Swert, and Claes H De Vreese. A clearer picture: The contribution of visuals and text to framing effects. *Journal of communication*, 65(6):997–1017, 2015.

[Preacher and Selig, 2012] Kristopher J. Preacher and James P. Selig. Advantages of Monte Carlo Confidence Intervals for Indirect Effects. *Communication Methods and Measures*, 6(2):77–98, April 2012.

[Ramachandran and Amir, 2007] Deepak Ramachandran and Eyal Amir. Bayesian Inverse Reinforcement Learning. *International Joint Conferences on Artificial Intelligence*, 7:2586–2591, January 2007.

[Reddy *et al.*, 2018] Siddharth Reddy, Anca Dragan, and Sergey Levine. Shared Autonomy via Deep Reinforcement Learning. In *Robotics: Science and Systems XIV*. Robotics: Science and Systems Foundation, June 2018.

[Sakoe and Chiba, 2003] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 2003.

[Simon, 1955] Herbert A. Simon. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69(1):99–118, 1955.

[Skalse and Abate, 2023] Joar Skalse and Alessandro Abate. Misspecification in inverse reinforcement learning. volume 37 of *AAAI'23/IAAI'23/EAAI'23*, pages 15136–15143. AAAI Press, February 2023.

[Slovic, 1987] Paul Slovic. Perception of risk. *Science*, 236(4799):280–285, 1987.

[Slovic, 2020] P. Slovic. The construction of preference. *Shaping Entrepreneurship Research: Made, as Well as Found*, pages 104–119, 5 2020.

[Smith and Vul, 2013] Kevin A. Smith and Edward Vul. Sources of Uncertainty in Intuitive Physics. *Topics in Cognitive Science*, 5(1):185–199, 2013.

[Unsplash, 2026] Unsplash. Beautiful Free Images & Pictures | Unsplash. https://unsplash.com/, 2026.

[Valente *et al.*, 2020] Matthew J. Valente, Judith J.M. Rijnhart, Heather L. Smyth, Felix B. Muniz, and David P. MacKinnon. Causal Mediation Programs in R, Mplus, SAS, SPSS, and Stata. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(6):975–984, November 2020.

[Viano *et al.*, 2021] Luca Viano, Yu-Ting Huang, Parameswaran Kamalaruban, Adrian Weller, and Volkan Cevher. Robust Inverse Reinforcement Learning under Transition Dynamics Mismatch. In *Advances in Neural Information Processing Systems*, volume 34, pages 25917–25931. Curran Associates, Inc., 2021.

[Wirth *et al.*, 2017] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18:1–46, 2017.

[Zhang *et al.*, 2009] Zhen Zhang, Michael J Zyphur, and Kristopher J Preacher. Testing multilevel mediation using hierarchical linear models: Problems and solutions. *Organizational Research Methods*, 12(4):695–719, 2009.

[Zhang *et al.*, 2017] Yu Zhang, Sarath Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. Plan explicability and predictability for robot task planning. In *IEEE international conference on robotics and automation*. IEEE, 2017.

[Ziebart *et al.*, 2008] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, pages 1433–1438, Chicago, Illinois, July 2008.