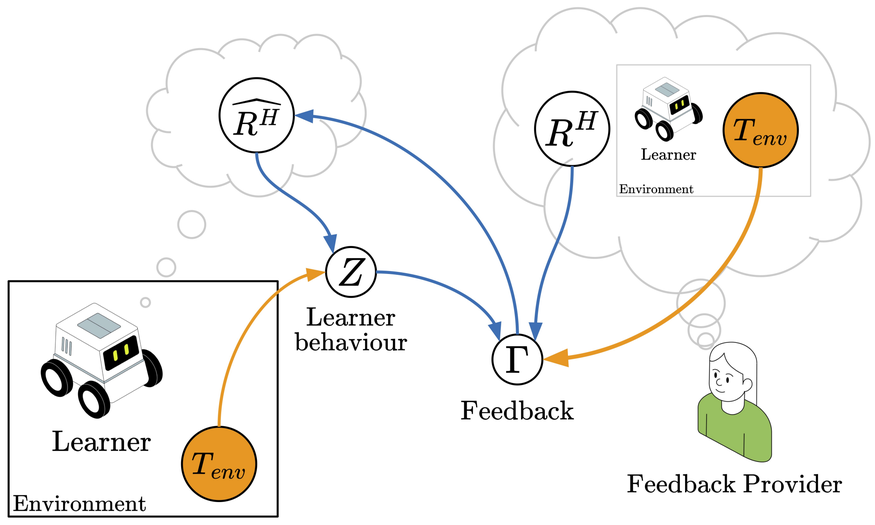
Empirical Evidence & Analysis of a Critical Pitfall in Reward Learning from Human Feedback (RLHF)

Taha Shaheen Stephen G. West Yu Zhang

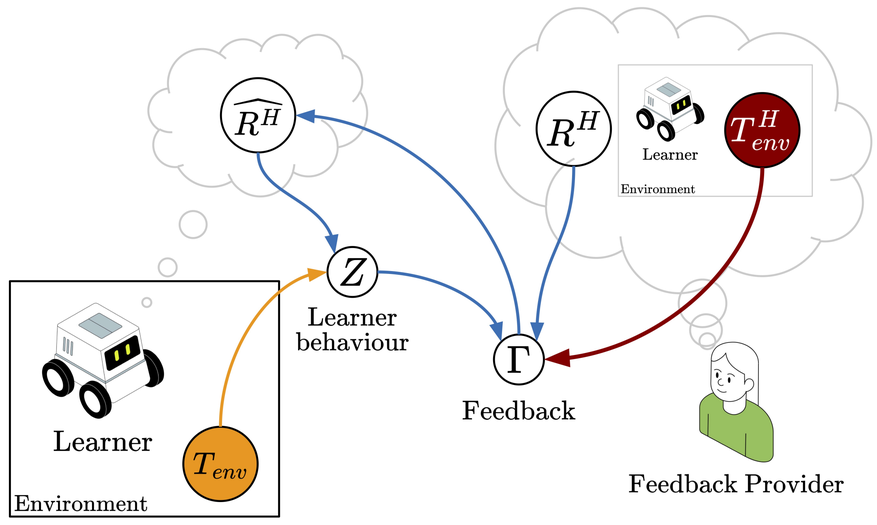
Arizona State University, USA
taha.shaheen@asu.edu

tahashaheen.github.io

Research Question: *Does a **perturbed** understanding of dynamics **impact** feedback?*



Standard view: Feedback Γ reflects human's latent reward R^H. Human knows true domain dynamics T_{env}.



Our view: Feedback Γ is a confounded signal impacted by perturbations in the human's dynamics understanding T^H_{env}.

Key Takeaways

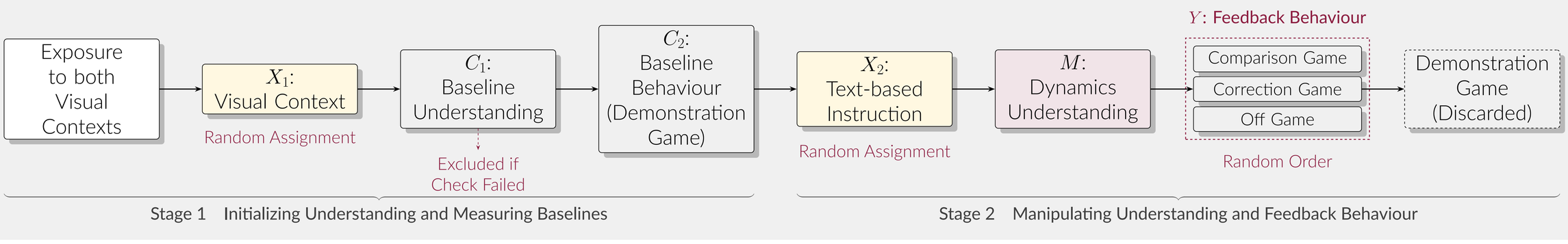
- **Pitfall in RLHF:** Methods that assume humans have perfect dynamics understandings ($T^H_{env} \equiv T_{env}$) risk learning rewards that encode dynamics misunderstandings.
- **Contribution:** Empirical evidence for link between human's dynamics model and feedback signal, showing that *dynamics understanding impacts feedback.*

Hypotheses

- **H1:** Internal dynamics understanding mediates effect of instruction framing on feedback.
- **H2:** The mediation persists across 3 feedback types: corrective interventions, pairwise preferences, off-style-interventions.
- **H3:** The mediation persists across 2 visual contexts.

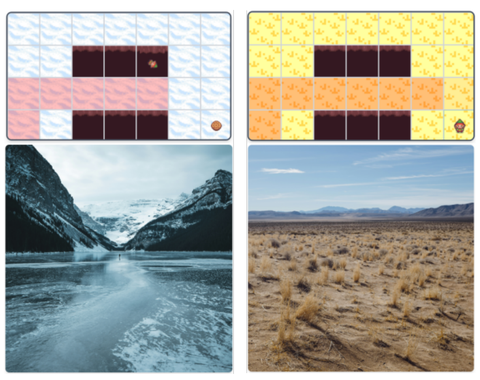
Study Design

Randomized controlled trial with $N = 211$ participants.



Randomized Control Trial

X_1: Exposure to both *visual contexts* before randomzation.



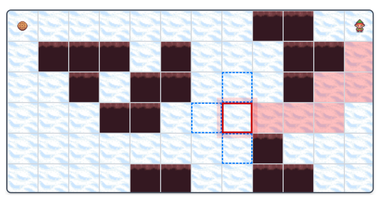
Icy/Slippery context Grassy/Stable context

$C_1 \& M$: Measuring *baseline* and *dynamics understanding*. Participants rated the slipping risk of two plans.

Cliff-adjacent plan: dynamics understanding question.

Away-from-cliffs plan: safety check question.

C_2: *Baseline behaviour* collected by playing a demonstration game.

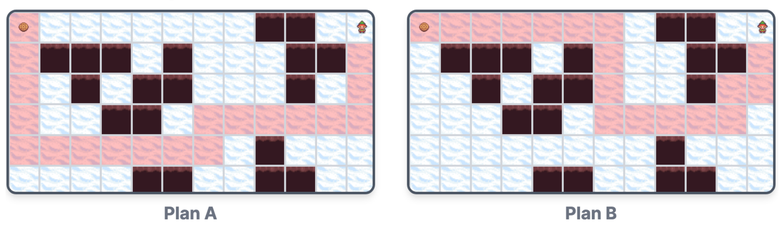


Demonstration: create trajectories by selecting tiles.

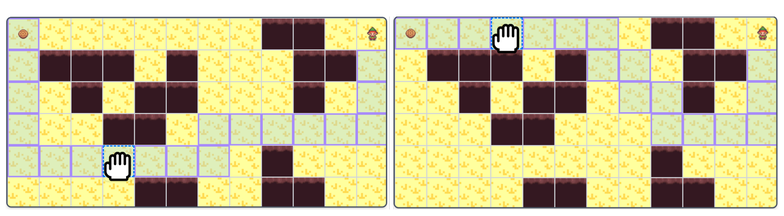
X_2: Updating dynamics using *text-based instruction framing*:

- **Safe:** ground is stable; accidental slipping is not possible.
- **Danger:** tiles are slippery and slope toward cliffs.

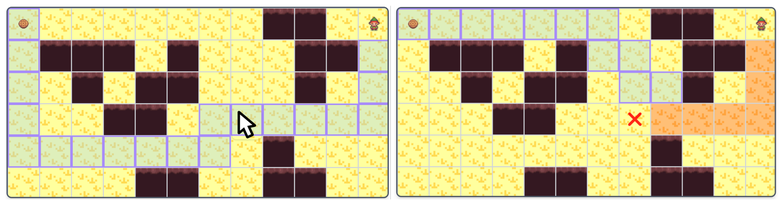
Y: Measuring feedback across three *feedback types*.



Preference: choose between two candidate trajectories.



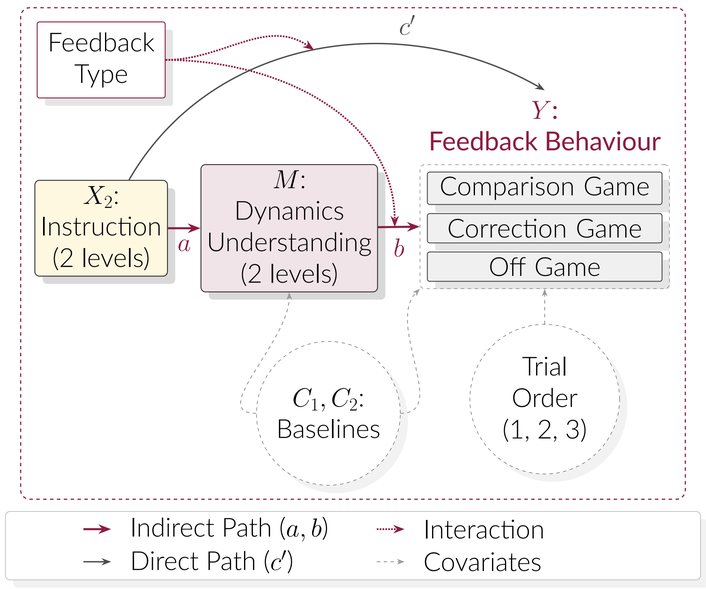
Correction: edit a proposed trajectory by dragging points.



Intervention: place obstacles to force replanning.

Causal Mediation Model

Split by Visual Context (X_1: Stable vs. Slippery)



What the Results Show

- Instructions changed the internal dynamics model.
- The changed internal model shifted feedback behaviour.
- Held in both stable and slippery visual contexts.
- Feedback type did not significantly moderate the effect.

Mediation Results

Effect	Stable Context	Slippery Context
$X_2 \to M$ (a)	-1.326^{***}	-0.918^{***}
$M \to Y$ (b)	0.448^{***}	0.356^{***}
Direct $X_2 \to Y$ (c')	-0.475^{*}	-0.662^{***}
Indirect ab	-0.593	-0.327
95% CI for ab	$[-0.920, -0.292]$	$[-0.547, -0.149]$

$^{*}p < .05$, $^{***}p < .001$. Indirect effects used 20,000 Monte Carlo draws.

Robustness

- **Across feedback types:** No main effect or interaction (all tests $p \geq .120$ in both visual contexts).
- **Across visual contexts:** Indirect-effect difference included zero: $ab_{diff} = -0.267$, 95% CI $[-0.641, 0.102]$.
- **Overall effect:** Average indirect effect remained significant: $ab_{avg} = -0.460$, 95% CI $[-0.649, -0.283]$.

Limitations & Next Steps

- Used simple grid world but real world settings have complex dynamics misconceptions.
- Manipulation was explicit but real-world misconceptions may be harder to detect.
- Safe/Danger framing may have changed perceived task difficulty or cognitive load. Did not measure.
- Future work should test whether observing execution corrects internal dynamics beliefs or leaves them intact.
- Algorithms should treat human internal dynamics as a latent variable, not only as random noise or noisy rationality.

Acknowledgements: This research is supported in part by NSF grant 2047186 and the 2025 ASU Graduate Student Government JumpStart Grant. We thank the reviewers for their comments and feedback.