UNDERSTANDING AI TEAMMATES THROUGH GROUNDED DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

Effective human–AI coordination requires people to understand and anticipate AI teammates' behavior. Yet plausible post-hoc explanations may not faithfully reflect the policy that selected an action. We present PolicyLens, an interactive system that provides evidence-grounded explanations of AI teammates' actions and the possible consequences of alternative human actions. Before deployment, its regularity-constrained policy distillation module extracts compact executable summaries of neural behavior; during collaboration, PolicyLens receives open-ended questions and constructs answers from interaction records, available program paths, or isolated simulations using the same fixed controller. We implement PolicyLens in Warehouse, Cooperative Pong, and an Overcooked-inspired Cooperative Kitchen, matching explanations to the controller used in each environment. In a 60-participant study across the three environments, 53.3% of the explanation group gave positive ratings (5–7 on a seven-point scale) on the item about understanding when their teammate waited or changed plans, compared with 30.0% of the control group, a 77.8% higher proportion. This overall result provides preliminary support for improving participants' perceived understanding of AI teammates. PolicyLens offers a concrete approach to supporting human–AI collaboration through dialogue grounded in verifiable evidence.

1 INTRODUCTION

Human–AI teams must coordinate their actions to achieve shared goals, making effective collaboration a central objective for interactive AI (Carroll et al., 2019). However, involving people throughout reinforcement learning or collecting human training data can be costly (Zhao et al., 2023). A widely studied alternative is to train agents with artificial teammates through self-play, then pair the trained agents with humans at deployment (Carroll et al., 2019; Hu et al., 2020). This creates a fundamental coordination problem: how can a newly introduced human collaborate with an AI teammate whose strategies were learned through interaction with other agents?

Self-play can establish coordination conventions that work well among training partners but are unfamiliar to humans (Carroll et al., 2019; Hu et al., 2020). A new teammate must infer when the AI will wait, which role it will take, and how it may respond to a different human action. Motivated by this problem, we focus on supporting the human's understanding of a fixed AI teammate during collaboration. Questions such as "Why did you wait?" and "What if I had moved instead?" require both an account of the observed decision and evidence about a possible alternative. Our central question is whether dialogue grounded in the teammate's behavior can support this understanding and subsequent coordination without updating the deployed controller.

Policy extraction and programmatic reinforcement learning offer executable representations of policy behavior (Bastani et al., 2018; Verma et al., 2018), while conversational systems let users explore model explanations through natural language (Slack et al., 2023). For teammate dialogue, these capabilities must connect to the specific interaction being discussed. A neural proposal may be adjusted by coordination rules, and an attempted action may be blocked by the environment. A plausible description can therefore misidentify the process responsible for an observed outcome, reflecting the distinction between plausibility and faithfulness (Jacovi & Goldberg, 2020). An answer must identify the relevant decision, its supporting evidence, and the conditions under which an alternative is evaluated.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

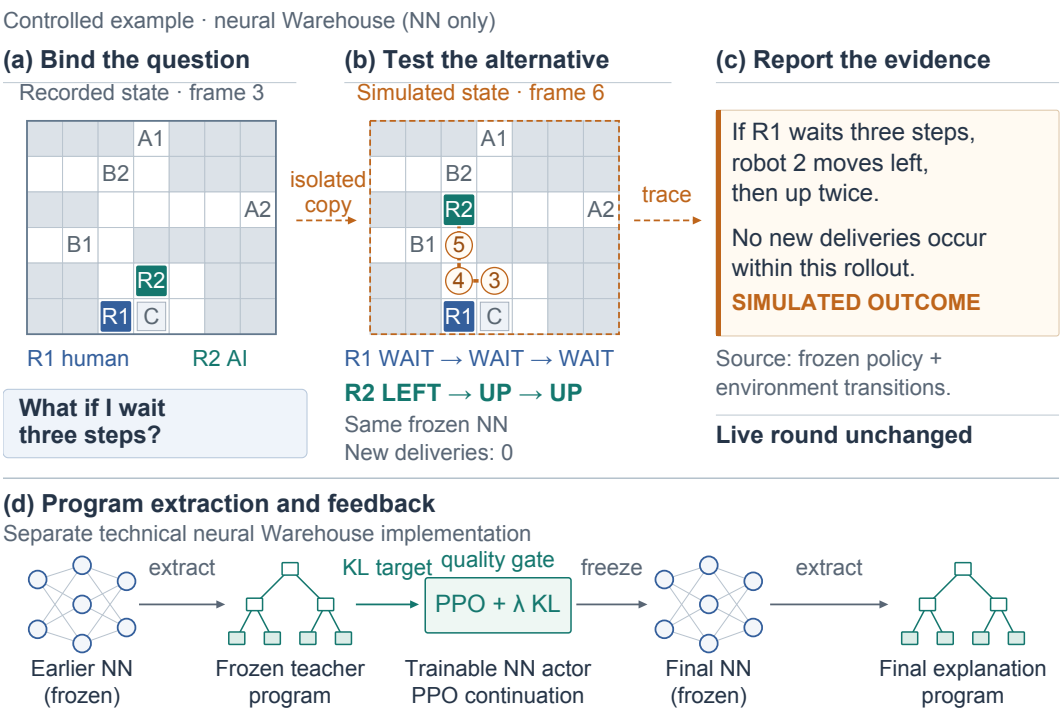

Figure 1: **From an interaction question to a checkable answer.** (a–c) A controlled neural Warehouse example, with a condensed logged answer. Orange dashed elements denote simulation; path markers show frames 3–5 and R2 is at frame 6. A/B denote pickup/delivery points; C denotes charging. (d) A frozen teacher program provides gated training feedback; the final explanation program is extracted after freezing the policy. Network/tree glyphs are schematic. Runtime dialogue updates neither the live state nor network parameters.

We present PolicyLens, an interactive system that explains AI teammates' actions and explores alternative human actions using evidence from the controller and environment (Figure 1). For neural policies, its regularity-constrained policy distillation module extracts compact executable summaries before deployment. A separate frozen program can provide gated feedback during offline policy training; the final explanation program summarizes the resulting policy without replacing it. During collaboration, a language model maps open-ended questions to supported evidence requests tied to the selected interaction. PolicyLens retrieves interaction and controller records, uses available program paths when the program's predicted action matches the neural policy's action, and evaluates supported human-action alternatives through isolated simulations of the same fixed controller. Answers distinguish recorded behavior, approximate program conditions, and simulated outcomes.

We instantiate the dialogue interface in Warehouse, Cooperative Pong, and an Overcooked-inspired Cooperative Kitchen, and evaluate program learning separately in neural Warehouse. Our 60-participant study compares explanation access with a no-explanation control. Within each environment and cohort, both groups use the same AI teammate and scoring rules. Only the explanation group receives guided prompts and question answering in the intervention task; neither group has explanations in the baseline or subsequent task. This design evaluates the explanation interface as a whole, including its guidance and access, while keeping the teammate controller fixed across conditions.

In the final questionnaire, 53.3% of the explanation group and 30.0% of controls rated their understanding of when the teammate waited or changed plans positively (5–7 on a seven-point scale). These descriptive results provide preliminary support for perceived understanding, while baseline-adjusted task-performance comparisons remain inconclusive. Our contributions are a dialogue framework that connects questions to controller evidence, a program-learning component for neural policies, and a three-environment evaluation of explanation access during human–AI collaboration.

2 Related Work

Policy extraction and behavioral summaries. Policy representations encode different objects. Distillation transfers teacher outputs to neural students (Rusu et al., 2016); Linear Model U-Trees approximate neural Q-values (Liu et al., 2019), while differentiable trees undergo reinforcement learning before discretization (Silva et al., 2020). PROPEL projects neural updates into programs through imitation (Verma et al., 2019); Deep Symbolic Policy optimizes symbolic controllers using returns (Landajuela et al., 2021). Behavioral summaries instead aggregate states into policy graphs (Topin & Veloso, 2019), select trajectories around action-value gaps (Amir & Amir, 2018), or identify visitation and transition patterns (Sequeira & Gervasio, 2020). Programs describe decision conditions across states; trajectories document particular executions. Our Warehouse distillation runs use programs as approximations and feedback targets; runtime records additionally capture coordination-rule adjustments and execution outcomes.

Interactive and conversational explanations. Question taxonomies connect user needs to explanation methods (Lim et al., 2009; Liao et al., 2020), while dialogue protocols organize clarification, follow-ups, and challenges (Madumal et al., 2019). Robot-policy interfaces answer behavioral questions using predicate summaries of observed executions (Hayes & Shah, 2017). Glass-Box constructs constrained counterfactual queries, whereas LLMCheckup parses requests into interpretability operations (Sokol & Flach, 2018; Wang et al., 2024). ConvXAI supports contextual exploration of heterogeneous explanations (Shen et al., 2023); EMCEE learns dialogue about static explanations from filtered synthetic conversations (Zhang et al., 2025). Our questions refer to saved joint interactions. Binding requests to actors and decision times distinguishes neural proposals, rule adjustments, and physical outcomes, which require different evidence. Incoming actions, next decisions, and hypothetical continuations require different records or simulations. Evidence validation constrains what is rendered, but does not guarantee answer completeness.

Contrastive and counterfactual explanations. Intervention targets determine evidential scope. Visual methods generate action-changing observations (Olson et al., 2021) or learn masks for action-specific input features (Zuo et al., 2025). Other approaches simulate user-derived foil policies (van der Waa et al., 2018), relate future features to action preferences (Lin et al., 2021), or revise a human's symbolic model until the robot's plan is optimal within it (Chakraborti et al., 2017). Recourse distinguishes feature changes from executable interventions (Karimi et al., 2021); sequential counterfactual optimization and counterfactual off-policy evaluation require explicit causal assumptions (Tsirtsis et al., 2021; Oberst & Sontag, 2019). We instead vary human commands in an environment copy while preserving the frozen teammate controller. These bounded continuations depend on the selected state, commands, and simulated dynamics, rather than predicting actual human choices.

Human–AI coordination and evaluation. Legibility separates goal inference from motion prediction (Dragan et al., 2013); mutual-adaptation models represent changing human strategies (Nikolaidis et al., 2017). Explanation studies examine behavioral prediction (Hase & Bansal, 2020), error correction (Poursabzi-Sangdeh et al., 2021), trust calibration (Zhang et al., 2020), and overreliance on AI recommendations (Buçinca et al., 2021). These outcomes are not interchangeable: proxy tasks and preferences can disagree with decision-task performance (Buçinca et al., 2020). Across three tasks with comparable human and AI accuracy, explanations added no significant performance benefit over recommendations and confidence (Bansal et al., 2021). These studies motivate separating policy approximation, behavioral understanding, and coordination. In Warehouse, Cooperative Pong, and Cooperative Kitchen, the question is whether evidence about a fixed teammate supports subsequent joint action.

3 Grounded Dialogue for AI Teammates

An explanation of a joint action must identify both the decision it describes and the evidence supporting it. PolicyLens associates questions with saved interactions, separates neural proposals from controller adjustments and physical outcomes, and evaluates supported alternatives in independent environment copies. The interface supports Warehouse, Cooperative Pong, and Cooperative Kitchen.

3.1 Interaction Setting and Controller Evidence

A teammate's proposed action may be changed by a coordination rule or prevented from executing by another robot or the map. We therefore distinguish the neural proposal, the controller's selected action, and the observed outcome. Let x_t describe the interaction at turn t, including the environment state, relevant history, controller memory, and random state. Given observation o_t, the frozen neural policy $\pi_{\bar{\theta}}$ proposes $\hat{a}_t$, and the fixed coordinator selects $a_t = C(x_t, \hat{a}_t)$. The full controller K packages this action, a pending memory update, and supporting evidence into the decision record $d_t = K(x_t)$. Together with the human's action u_t, this record determines the joint transition: $(x_{t+1}, e_t) = F(x_t, u_t, d_t)$. The transition function commits the memory update and records what actually happened through execution events e_t. Where available, the saved proposals, rule adjustments, and execution outcomes let an explanation distinguish what the network suggested, what the controller selected, and what the environment allowed.

Time binding determines which record answers the question. For $t > 0$, displayed frame x_t results from the preceding joint action. "Why did you do that?" refers to the saved decision at x_{t-1}, whereas a next-action question at a nonterminal frame refers to $K(x_t)$ (Figure 2). The server retrieves the incoming transition rather than recomputing its explanation from the displayed state. It binds requests to an authorized session, domain instance, run, task, and turn, and records release identifiers and state hashes. Authorization is checked again before an answer becomes available, preventing a delayed response from reopening an expired explanation phase.

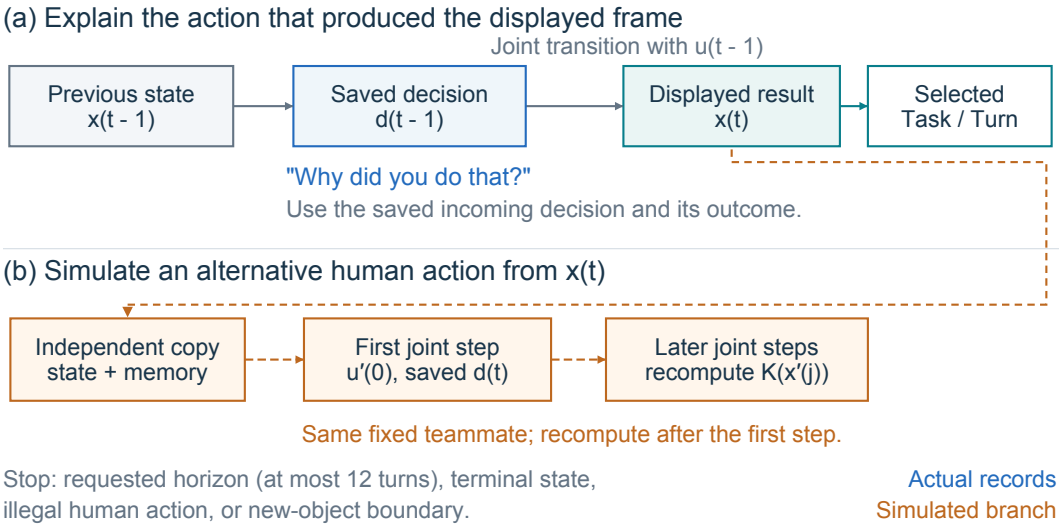

Figure 2: **Explaining recorded actions and simulating alternatives.** (a) To explain a past action, the system retrieves the saved decision and execution outcome that produced the displayed state x_t (solid arrows). (b) Alternative human actions are tested in an independent copy of x_t (dashed arrows). The first step preserves the robot's saved decision d_t, including its action and pending memory update. Subsequent decisions are computed by the same fixed controller K on simulated states. Position interventions recompute the initial decision (Section 3.4).

3.2 Executable Policy Summaries

An executable decision tree summarizes the neural policy through conditions that can be inspected for a particular decision (Bastani et al., 2018; Verma et al., 2018). In the shared backend, we fit this tree to the action probabilities of a specified frozen policy $\pi_{\bar{\theta}}$. Training observations come from interactions with that policy and simulations of alternative human actions. The same neural policy supplies all target distributions, including at intervened states. The tree and Actor receive the same public features in the same order, including verified information about the preceding transition. We

fit the tree greedily under depth and leaf limits, using the weighted squared-error objective

$$\min_{T \in \mathcal{T}_{d,L}} \sum_i w_i \big\| T(o_i) - \pi_{\bar{\theta}}(\cdot \mid o_i) \big\|_2^2, \qquad w_i = (1 + 8\Delta_i) m_i. \tag{1}$$

Here Δ_i is the difference between the two largest neural action probabilities. The multiplicity m_i counts the original observation and its occurrences in retained intervention pairs that produce physical changes. Thus, observations with a clearer neural action preference or more retained occurrences receive more weight. Leaf outputs are normalized into action distributions.

We use different scenes and episodes for fitting and development evaluation, although the same observation may occur in both sets. Each predefined candidate tree must meet development agreement thresholds for ordinary actions and the required role and interaction categories. Eligible candidates are ranked by complexity, then KL divergence, then a fixed index to break ties. Complexity is measured by depth, leaf count, and the number of distinct split features; this selection criterion does not contribute an Actor gradient. When program explanations are enabled, we compare the tree's action with the neural policy's action for the queried decision. The shared backend presents the corresponding condition path as an approximate explanation only if they agree; otherwise, that path is withheld.

The extracted program also supplies offline training feedback. During a bounded PPO continuation (Schulman et al., 2017), a separate teacher program T^- remains frozen. We apply a probability floor to its outputs, normalize them, and detach the resulting target distributions q_{T^-} from gradient computation. The Actor minimizes

$$\mathcal{L}_{\text{actor}} = \mathcal{L}_{\text{PPO,ent}} + \tfrac{\lambda}{2} \mathbb{E}_{o \sim B_N} D_{\text{KL}}(\pi_\theta \| q_{T^-}) + \tfrac{\lambda}{2} \mathbb{E}_{o \sim B_P} D_{\text{KL}}(\pi_\theta \| q_{T^-}), \tag{2}$$

with all distributions conditioned on o. The two KL terms apply program feedback to different observations: B_N contains PPO samples collected under neural control, while B_P contains equally weighted endpoints of sampled intervention pairs. These auxiliary endpoints contribute only to the program-feedback term; they supply neither policy-gradient actions nor critic targets. If a required capability check or teacher-reliability check fails, feedback is disabled for the rest of that continuation. Setting $\lambda = 0$ recovers native PPO. Once the resulting Actor is frozen, we fit a new tree for explanation. The teacher tree guides training, and the final tree summarizes decisions; neither selects runtime actions in place of the controller.

3.3 State-Grounded Question Answering

PolicyLens answers questions using evidence from the selected interaction and its history. This includes saved decisions, execution outcomes, coordination plans, and environment rules. A language model interprets the question and produces a plan that selects evidence or requests a supported simulation. The plan identifies whose behavior is being discussed and whether the question concerns a past action, the current state, or a hypothetical continuation. The model selects these operations rather than directly writing factual claims.

Before answering, the system checks that the selected evidence matches the requested actor, time, and question type. For example, a ball's actual assignment must come from its recorded plan; reachability alone cannot establish that assignment. Detected ambiguities prompt clarification, while unresolved validation or service failures return an unavailable response.

Answers combine factual text with actual simulation outputs, retain their evidence identifiers, and distinguish recorded events from simulated outcomes. These checks keep answers tied to available evidence.

3.4 Counterfactual Rollouts

PolicyLens uses simulation to answer "What would happen if I acted differently?" It copies the state selected by the user and executes alternative human actions in that copy, leaving the live game unchanged. The simulation preserves existing coordination commitments and the random state. It runs for one joint turn by default and at most twelve turns.

For human-action interventions, the first step preserves the robot's saved decision, including its action and pending memory update. This decision was made before the alternative human command

was available. From the second step onward, the same fixed controller selects actions based on the simulated states. The system also supports position interventions: placing the human at another valid position and recomputing the robot's initial decision. These interventions can inspect a hypothetical decision without advancing game time and therefore differ from changing a human action.

4 Experiments

4.1 Questions and Tasks

We evaluate the reliability of PolicyLens's evidence and its value for collaboration during explanation access and after withdrawal. The human study covers Warehouse, Cooperative Pong, and Cooperative Kitchen (Figure 3); technical checks concern one frozen Warehouse neural policy.

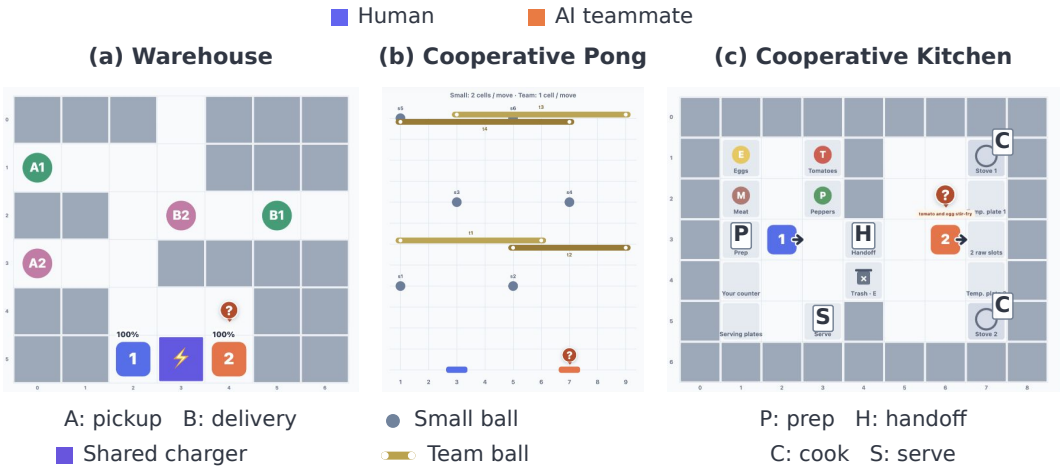

Figure 3: **Study tasks.** Warehouse: share routes and a charger to deliver parcels. Pong: allocate ordinary balls and jointly catch cooperative balls. Kitchen: prepare, hand off, cook, and serve food. These are controlled-preview screenshots, not participant trajectories.

4.2 Program and Replay Evaluation

The Warehouse neural policy was frozen after 3.95 million training steps. We fitted readable decision rules to its action distributions; the selected model, C0, matched 34,824 of 37,700 recommendations (92.37%; 95% interval, 91.89–92.80%). The simulator and independent neural-policy/physics replay executed the same one- or three-step sequences. Physical states and action traces matched in all 3,556 comparisons, without changing the live game or its random state.

Among 215 pairs from 62 scenes with different physical outcomes and next neural actions, C0 matched both actions in 123 (57.21%; 95% interval, 49.74–64.95%) and the simulator in all 215 (Figure 4). See Appendix A for protocols and metrics, and Section 5.3 for error analysis.

4.3 Human-Study Design

The formal Prolific study yielded 87 participation records (74 main, 13 replacement): 66 completed all three tasks and 21 did not. After excluding six completed records, we retained 60 participants (54 main, six quota replacements), with ten per domain and group. Main-study allocation was randomized; replacements filled quotas.

Participants completed Tasks 1–3 in fixed order within one domain. Task 1 had no explanations. In Task 2, A received event-triggered question prompts and question answering; the first guided question and answer acknowledgment were mandatory, with later requests optional. B had no explanations. Task 3 used a new scenario and removed access and previous answers. Reading and

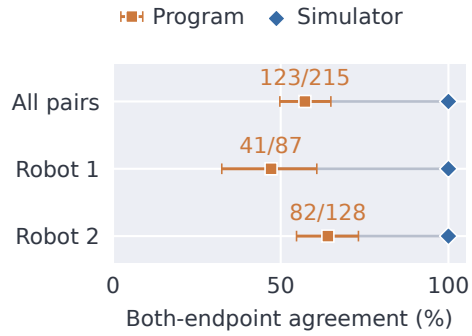

Figure 4: **Paired action agreement.** Both actions must match the reference. Error bars: 95% scene-bootstrap intervals. Pooled results combine both robots.

asking did not advance game turns. Within each domain and cohort, both groups used the same AI teammate and scoring rules. Both groups rated understanding during Task 2 and completed a final questionnaire.

Warehouse and Kitchen used final team points. Pong scored caught ball value divided by all scheduled ball value, multiplied by 100%; ordinary balls were worth one point and cooperative balls three. We compared A and B separately by domain on Tasks 2 and 3, adjusting for Task 1 scores. Effects use HC3 standard errors, allowing unequal score variability, pointwise 95% confidence intervals, and Holm correction across six tests. Each participant is one analysis unit; domains are not pooled. Full methods appear in Appendix B.

4.4 Human-Study Results

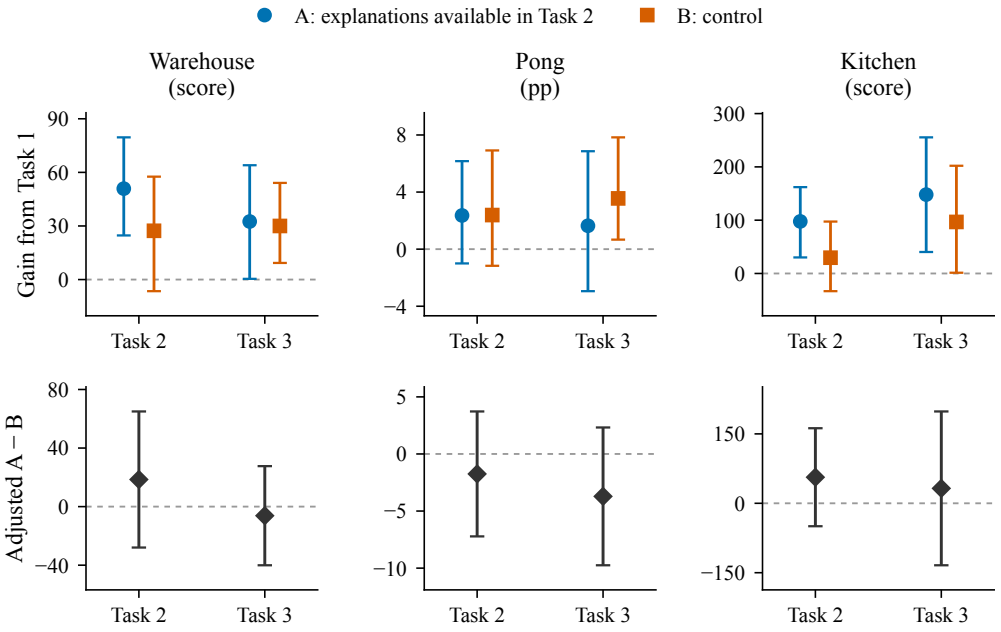

Figure 5: **Human performance.** Top: mean gains from Task 1, with 95% intervals from 10,000 within-group participant resamples. Bottom: adjusted A-minus-B effects with pointwise HC3 t_{17} 95% intervals. Each domain–group cell has ten participants. Units are points in Warehouse/Kitchen and percentage points in Pong; axes differ. A had explanations only in Task 2.

Adjusted A-minus-B differences were 18.5 points in Warehouse, −1.7 percentage points in Pong, and 56.3 points in Kitchen at Task 2; at Task 3, they were −6.2, −3.7, and 32.3, respectively (Figure 5). On the final seven-point item about understanding when the teammate waited or changed plans, 16/30 in A (53.3%) and 9/30 in B (30.0%) rated 5–7, a descriptive above-midpoint summary. Complete score and questionnaire results appear in Appendices B.4 and B.6.

5 Analysis

5.1 Gains During Access and After Withdrawal

Group A showed larger mean score gains in Warehouse and Kitchen, both during explanation access and after withdrawal (Figure 5). In Task 2, gains from Task 1 were 50.9 points in Warehouse and 97.7 points in Kitchen for A, compared with 27.3 and 29.6 points for B. In Task 3, after explanations were removed, gains from Task 1 remained larger for A: 32.5 versus 30.0 points in Warehouse, and 148.0 versus 96.7 points in Kitchen. These descriptive patterns highlight PolicyLens's potential to support human–AI collaboration.

Participants' feedback also provides encouraging evidence about perceived understanding. At the end of the study, 53.3% of A gave positive ratings (5–7 on a seven-point scale) on the item about understanding when their teammate waited or changed plans, compared with 30.0% of B. The proportion of positive ratings was therefore 77.8% higher in A. This self-reported result provides preliminary support for PolicyLens helping users feel that they better understand their AI teammates.

5.2 Why Pong May Differ

Pong showed nearly identical Task 2 gains (2.36 versus 2.39 percentage points) and a smaller Task 3 gain in A (1.64 versus 3.56). Visible ball positions may let participants judge local actions without explanations. Prior work relates explanation use to relative cognitive cost, self-confidence, and intuition (Vasconcelos et al., 2023; Chong et al., 2022; Chen et al., 2023).

Baseline ordinary-ball catch rates were 75.3% in A and 79.5% in B, but cooperative-ball catch rates were 3.5% in both. These rates are compatible with easier local decisions but difficult coordination. They cannot show that participants ignored the AI; that would require action-trajectory evidence.

5.3 What Rule Summaries and Replay Provide

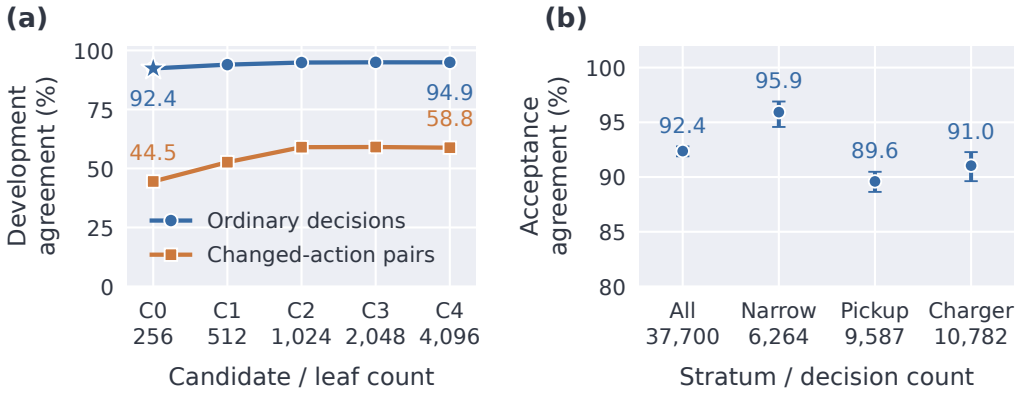

Figure 6: **Program diagnostics.** (a) Five configurations on 37,534 development decisions and 1,432 action-changing pairs; the star marks C0. Lines connect configurations, not training steps. (b) Acceptance by overlapping scene categories; the vertical axis starts at 80%. Whiskers show 95% intervals from 10,000 scene resamples, conditional on the fixed neural policy and program.

Figure 6 diagnoses the rule approximation. Candidates vary in fitting weights and tree size, so this is not a capacity-only comparison. Ordinary and paired agreement use different samples and units. In

one Warehouse test, changing the teammate's command changed the neural recommendation from waiting to moving up, but both states reached the same rule, which still recommended waiting. Rules can miss distinctions needed for alternative actions; replay checks their outcomes under specified commands. Its short-sequence consistency does not validate future human choices or answer completeness (Appendix C).

6 CONCLUSION

We presented PolicyLens, a system for asking "why" and "what if" questions about AI teammates. Across Warehouse, Cooperative Pong, and Cooperative Kitchen, it connects questions to controller records and simulations without changing the live interaction. Answers distinguish neural proposals, controller actions, and observed outcomes, while keeping approximate program summaries separate from execution evidence.

At the end of our 60-participant study, the proportion reporting positive teammate understanding (5–7 on a seven-point scale) was 53.3% in the explanation group and 30.0% in the control group, a 77.8% higher proportion. In descriptive comparisons, the explanation group's mean score gains from Task 1 were 86.4% larger in Warehouse and 230.1% larger in Kitchen than the control group's during Task 2, when explanations were available. After withdrawal in Task 3, the corresponding differences were 8.3% and 53.1%. Pong showed no corresponding advantage in either task, suggesting a possible role for self-confidence, intuition, and the potentially lower learning demands of its local rules.

PolicyLens provides a concrete, traceable basis for reasoning about AI teammates. Future work will combine objective understanding tests and delayed transfer with larger randomized studies that balance scenario order, and adapt explanations to users' uncertainty and coordination needs.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language refinement, Chinese–English translation, literature search and organization, and limited verification of the feasibility of selected code implementations. The tool was not used to formulate or determine the research questions, methodological design, experimental results, or scientific conclusions. All AI-assisted text, bibliographic information, and code-related suggestions were independently reviewed and verified by the authors against the original sources, research records, and actual execution results where applicable. The authors take full responsibility for the accuracy, integrity, and final content of this work.

ETHICS STATEMENT

The human-participant study was approved by the relevant institutional review board. Identifying details are omitted for double-blind review. Participants provided informed consent before participation and were informed about the tasks, recorded data, and use of an external AI service where applicable. Participation was voluntary, and participants could withdraw without penalty. Compensation followed the terms communicated during recruitment.

Research records were stored under pseudonymous participant identifiers. Access to identifiable recruitment and compensation information was restricted to the research team. Reporting excludes directly identifying participant information. Any sharing of participant-level records is subject to the approved protocol and participants' consent.

REFERENCES

Dan Amir and Ofra Amir. HIGHLIGHTS: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1168–1176. International Foundation for Autonomous Agents and Multiagent Systems, 2018. URL `https://ifaamas.org/Proceedings/aamas2018/pdfs/p1168.pdf`.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 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*, CHI '21, pp. 1–16. Association for Computing Machinery, 2021. doi: 10.1145/3411764.3445717.

Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable reinforcement learning via policy extraction. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/e6d8545daa42d5ced125a4bf747b3688-Paper.pdf.

Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, pp. 454–464. Association for Computing Machinery, 2020. doi: 10.1145/3377325.3377498.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021. doi: 10.1145/3449287. URL http://dx.doi.org/10.1145/3449287.

Micah Carroll, Rohin Shah, Mark Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human–AI coordination. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/f5b1b89d98b7286673128a5fb112cb9a-Paper.pdf.

Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 156–163, 2017. doi: 10.24963/ijcai.2017/23. URL https://www.ijcai.org/proceedings/2017/23.

Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. Understanding the role of human intuition on reliance in human–AI decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):370:1–370:32, 2023. doi: 10.1145/3610219. URL https://doi.org/10.1145/3610219.

Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, 127:107018, 2022. doi: 10.1016/j.chb.2021.107018. URL https://doi.org/10.1016/j.chb.2021.107018.

Anca D. Dragan, Kenton C.T. Lee, and Siddhartha S. Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 301–308. IEEE, 2013. doi: 10.1109/hri.2013.6483603. URL http://dx.doi.org/10.1109/HRI.2013.6483603.

Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5540–5552. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.491. URL http://dx.doi.org/10.18653/v1/2020.acl-main.491.

Bradley Hayes and Julie A. Shah. Improving robot controller transparency through autonomous policy explanation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 303–312. Association for Computing Machinery, 2017. doi: 10.1145/2909824.3020233. URL https://dspace.mit.edu/entities/publication/cb3d847e-18cc-4dbe-90df-92442f145dfb.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "other-play" for zero-shot coordination. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4399–4410. PMLR, 2020. URL https://proceedings.mlr.press/v119/hu20a.html.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, jul 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 353–362, 2021. doi: 10.1145/3442188.3445899.

Mikel Landajuela, Brenden K. Petersen, Sookyung Kim, Claudio P. Santiago, Ruben Glatt, Nathan Mundhenk, Jacob F. Pettit, and Daniel Faissol. Discovering symbolic policies with deep reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5979–5989. PMLR, 2021. URL https://proceedings.mlr.press/v139/landajuela21a.html.

Q. Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–15. Association for Computing Machinery, 2020. doi: 10.1145/3313831.3376590. URL https://doi.org/10.1145/3313831.3376590.

Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2119–2128. Association for Computing Machinery, 2009. doi: 10.1145/1518701.1519023. URL https://doi.org/10.1145/1518701.1519023.

Zhengxian Lin, Kin-Ho Lam, and Alan Fern. Contrastive explanations for reinforcement learning via embedded self predictions. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ud3DSz72nYR.

Guiliang Liu, Oliver Schulte, Wang Zhu, and Qingcan Li. Toward interpretable deep reinforcement learning with linear model U-trees. In *Machine Learning and Knowledge Discovery in Databases: ECML PKDD 2018, Part II*, volume 11052 of *Lecture Notes in Computer Science*, pp. 414–429. Springer, 2019. doi: 10.1007/978-3-030-10928-8_25. URL https://doi.org/10.1007/978-3-030-10928-8_25.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. A grounded interaction protocol for explainable artificial intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1033–1041. International Foundation for Autonomous Agents and Multiagent Systems, 2019. URL https://ifaamas.org/Proceedings/aamas2019/pdfs/p1033.pdf.

Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research*, 36 (5-7):618–634, 2017. doi: 10.1177/0278364917690593. URL http://dx.doi.org/10.1177/0278364917690593.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with Gumbel-Max structural causal models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4881–4890. PMLR, 2019. URL https://proceedings.mlr.press/v97/oberst19a.html.

Matthew L. Olson, Roli Khanna, Lawrence Neal, Fuxin Li, and Weng-Keen Wong. Counterfactual state explanations for reinforcement learning agents via generative deep learning. *Artificial Intelligence*, 295:103455, 2021. doi: 10.1016/j.artint.2021.103455.

Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pp. 1–52. ACM, 2021. doi: 10.1145/3411764.3445315. URL http://dx.doi.org/10.1145/3411764.3445315.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.06295.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Pedro Sequeira and Melinda Gervasio. Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations. *Artificial Intelligence*, 288:103367, 2020. doi: 10.1016/j.artint.2020.103367. URL https://doi.org/10.1016/j.artint.2020.103367.

Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. ConvXAI: Delivering heterogeneous AI explanations via conversations to support human-AI scientific writing. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, pp. 384–387. Association for Computing Machinery, 2023. doi: 10.1145/3584931.3607492. URL https://doi.org/10.1145/3584931.3607492.

Andrew Silva, Taylor Killian, Ivan Jimenez, Sung-Hyun Son, and Matthew Gombolay. Optimization methods for interpretable differentiable decision trees applied to reinforcement learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1855–1865. PMLR, 2020. URL https://proceedings.mlr.press/v108/silva20a.html.

Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence*, 5(8):873–883, 2023. doi: 10.1038/s42256-023-00692-8. URL https://doi.org/10.1038/s42256-023-00692-8.

Kacper Sokol and Peter Flach. Glass-Box: Explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 5868–5870. International Joint Conferences on Artificial Intelligence Organization, 2018. doi: 10.24963/ijcai.2018/865. URL https://www.ijcai.org/proceedings/2018/865.

Nicholay Topin and Manuela Veloso. Generation of policy-level explanations for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2514–2521, 2019. doi: 10.1609/aaai.v33i01.33012514. URL https://ojs.aaai.org/index.php/AAAI/article/view/4097.

Stratis Tsirtsis, Abir De, and Manuel Gomez-Rodriguez. Counterfactual explanations in sequential decision making under uncertainty. In *Advances in Neural Information Processing Systems*, volume 34, pp. 30127–30139, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/fd0a5a5e367a0955d81278062ef37429-Abstract.html.

Jasper van der Waa, Jurriaan van Diggelen, Karel van den Bosch, and Mark Neerincx. Contrastive explanations for reinforcement learning in terms of expected consequences. In *IJCAI/ECAI Workshop on Explainable Artificial Intelligence*, 2018. URL https://arxiv.org/abs/1807.08706.

Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7 (CSCW1):129:1–129:38, 2023. doi: 10.1145/3579605. URL https://doi.org/10.1145/3579605.

Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. Programmatically interpretable reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5045–5054. PMLR, 2018. URL https://proceedings.mlr.press/v80/verma18a.html.

Abhinav Verma, Hoang M. Le, Yisong Yue, and Swarat Chaudhuri. Imitation-projected programmatic reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/5a44a53b7d26bb1e54c05222f186dcfb-Abstract.html.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Josef van Genabith, Leonhard Hennig, and Sebastian Möller. LLMCheckup: Conversational examination of large language models via interpretability tools and self-explanations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pp. 89–104. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.hcinlp-1.9. URL https://aclanthology.org/2024.hcinlp-1.9/.

Tong Zhang, Mengao Zhang, Wei Yan Low, X. Jessie Yang, and Boyang Albert Li. Conversational explanations: Discussing explainable AI with non-AI experts. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 409–424. Association for Computing Machinery, 2025. doi: 10.1145/3708359.3712143. URL https://doi.org/10.1145/3708359.3712143.

Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 295–305. ACM, 2020. doi: 10.1145/3351095.3372852. URL http://dx.doi.org/10.1145/3351095.3372852.

Rui Zhao, Jinming Song, Yufeng Yuan, Haifeng Hu, Yang Gao, Yi Wu, Zhongqian Sun, and Wei Yang. Maximum entropy population-based training for zero-shot human–AI coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6145–6153, 2023. doi: 10.1609/aaai.v37i5.25758. URL https://ojs.aaai.org/index.php/AAAI/article/view/25758.

Rui Zuo, Simon Khan, Zifan Wang, Garrett Ethan Katz, and Qinru Qiu. Why the agent made that decision: Contrastive explanation learning for reinforcement learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 655–663, 2025. doi: 10.24963/ijcai.2025/74. URL https://www.ijcai.org/proceedings/2025/74.

A Technical Evaluation Details

Table 1: Evaluation units and reference evidence. Decisions from one scene and language variants of one fixture are not independent replications.

Evaluation	Units	Reference
Program acceptance	37,700 decisions; 64 scenes	Frozen Warehouse neural policy
Simulator replay	3,556 comparisons	Independent neural-policy/physics replay
Structured dialogue	326 bilingual fixtures	Supplied plans and specified evidence checks

A.1 Data and Program Selection

The Warehouse policy was frozen after 3.95 million joint training steps. Extraction uses 197 ordered public features and five actions. Fitting, development, and acceptance use 192, 64, and 64 scenes from the existing reinforcement-learning training manifest, respectively. Each scene is paired with self-play and three scripted partners. Scene and episode identities are separate, but these are not unseen training environments. Twelve identical public observations occur in both fitting and development. A recorded amendment retained them and froze all five candidates before development metrics were read.

The fitting pool contains 203,879 unique rows and 72,589 intervention pairs; 44,140 physically effective pairs add endpoint observations. The skilled partner minimizes joint pickup distance and uses a collision-free joint route; the assertive partner greedily approaches its nearest pickup. Both

account for charging. The noisy partner follows the skilled policy, with probability 0.1 of choosing uniformly among all five commands. Self-play uses the frozen policy for both robots.

Three overlapping critical categories use public state. Narrow passages have at most two passable neighbors and shortest-path distance between robots at most three. Shared pickups have an available pickup within distance four of both robots. Shared chargers are within distance three of both, or within five when minimum team battery is at most 30. Non-WAIT membership uses the neural policy's most probable action.

The ordinary selection rule was recorded before reading development results. It requires agreement of at least 90% overall and 85% for non-WAIT actions, each robot, and each critical category and its non-WAIT subset, with at least ten scenes per required category. Eligible candidates minimize

$$C(T) = 0.5 \min(1, d/20) + 0.3 \min(1, \ell/4096) + 0.2 \min(1, p/4095),$$

where d, ℓ, and p are depth, leaf count, and distinct split-feature count. KL divergence and candidate index break ties. This selects C0 before acceptance scoring.

The original stricter rule remained in parallel: ordinary agreement and agreement across all records, including intervention samples, must reach 90% overall and 85% on non-WAIT and critical subsets, with KL at most 0.35. Both-endpoint agreement must reach 85% overall and in every critical category.

Table 2: Frozen candidates on 37,534 ordinary development decisions and 1,432 action-changing pairs. c multiplies fitting weights for changed-pair endpoints; η weights a hard-action target. C0 uses only neural probability targets. These are not separate training-seed replications.

	Depth	Leaves	c	η	Ordinary (%)	Pair (%)	KL
C0	12	256	1	0	92.37	44.48	.170
C1	14	512	2	.5	93.98	52.65	.139
C2	16	1,024	4	.5	94.87	59.01	.114
C3	18	2,048	4	1	94.96	59.08	.098
C4	20	4,096	4	1	94.95	58.80	.086

A.2 Agreement and Replay Measures

Let $\hat{a}_i$ and a_i^T be the most probable neural and program actions for observation o_i. Ordinary agreement is

$$\widehat{A}(T) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[a_i^T = \hat{a}_i]. \tag{3}$$

Write $p_i(a) = \pi_{\bar{\theta}}(a \mid o_i)$ and $q_i(a) = T(o_i)_a$. The saved evaluator computes

$$\widehat{D}_\epsilon = \frac{1}{N} \sum_{i=1}^{N} \sum_{a \in \mathcal{A}} p_i(a) \log \frac{\max\{p_i(a), \epsilon\}}{\max\{q_i(a), \epsilon\}}, \qquad \epsilon = 10^{-8}. \tag{4}$$

Logs are natural. Only log arguments are floored; weights are not renormalized.

An effective pair has two nonterminal endpoints with different physical states and different next neural actions. Both branches must match:

$$\widehat{A}_{\text{pair}}(M) = \frac{1}{|\mathcal{P}_{\text{eff}}|} \sum_{j \in \mathcal{P}_{\text{eff}}} \prod_{b \in \{0,1\}} \mathbf{1}[a_{jb}^M = a_{jb}^{\text{ref}}]. \tag{5}$$

Here M is the program or simulator, b indexes branches, and the reference is independent replay's next neural action.

Replay compares an initial human movement with waiting, followed by WAIT commands over one or three joint steps. Simulation and independent replay start from the same saved state. Checks cover commands, executed actions, physical and random states, and endpoint neural outputs. All 3,556 comparisons match; the live game and its random state remain unchanged.

The original extraction acceptance data contain 67,999 rows and 24,233 pairs, including 1,519 physically effective, neural-action-changing pairs with 42.73% program agreement. The separate replay corpus contains 215 effective pairs from 62 scenes: C0 matches both actions in 123 (57.21%; 95% interval, 49.74–64.95%), and the simulator in all 215. These are different populations. Analysis recomputes saved traces, rather than rerunning all historical neural rollouts.

A.3 Uncertainty and Additional Results

We check all 256 acceptance context files against the frozen plan. All ordinary rows remain, and robot results are pooled by decision count. Bootstrap intervals use 10,000 resamples of 64 physical scenes with replacement (seed 260923), retaining every selected scene's robot decisions and partner profiles. We recompute pooled agreement and take its 2.5th and 97.5th percentiles. Intervals describe scene variation conditional on one frozen policy and program; they exclude training uncertainty. The simulator's all-pass checks are not assigned a confidence interval.

Overall acceptance agreement is 34,824/37,700 (92.37%; 95% interval, 91.89–92.80%), with mean KL 0.169. Agreement is 91.97% for Robot 1 and 92.77% for Robot 2. Narrow-passage, shared-pickup, and shared-charger agreement is 95.93%, 89.60%, and 91.04%. Non-WAIT agreement is 91.87% over 31,907 decisions. Categories overlap. These rates measure neural action recommendations, including inactive roles, rather than successful movement.

B Human Study

B.1 Participants and Procedure

Formal recruitment yielded 87 participation records: 74 main and 13 replacement. Of these, 66 completed all tasks and 21 did not. Six main-study completers were excluded: five recorded as invalid or problematic participation and one after an 11,632-second interruption. The available notes do not establish a uniform quantitative exclusion threshold. We preserve these decisions without adding score-based exclusions. The 25 earlier pilot records are separate.

The retained sample comprises 54 main and six replacement participants, with ten per domain–group cell. Allocation records show 46 balanced-block assignments, eight randomized vacancy assignments, and six quota replacements. Replacements were one Warehouse A, two Pong B, one Kitchen A, and two Kitchen B. This is an analysis of retained completers, not all assigned participants (intention-to-treat). Optional explanation use does not change group membership; noncompletion is not scored zero.

Participants consented in English and declared an age of at least 21. Consent described voluntary withdrawal, pseudonymous interaction records, and external AI processing; no audio or video was collected. Completion paid GBP 3.00 plus up to GBP 0.20 per task.

Everyone completed Tasks 1–3 in fixed order. Task 1 had no explanations. In Task 2, A's first prompt required selecting a question, opening its answer, and acknowledging it before continuing (turn 0 in Warehouse, 4 in Pong, and 5 in Kitchen). Later event-triggered explanations were optional; a separate channel accepted free-text or example questions. B had neither channel. Task 3 introduced a new within-domain scenario and removed explanations and answer history. Reading and asking did not advance game turns. Both groups rated understanding during Task 2 and completed a final questionnaire.

Within each domain and cohort, groups used the same controller and scoring rules. Main and replacement cohorts used different releases. Main-release source code was inspected; replacement settings and events were checked from exports, without independent verification of source equivalence.

B.2 Tasks and Scoring

Warehouse allowed 120 turns per task and used seeds 52000, 51000, and 51500 for Tasks 1–3 in both groups. Pong allowed 60 turns with 60 ordinary and 20 cooperative balls in Task 1; Tasks 2–3 allowed 90 turns with 90 ordinary and 30 cooperative balls. Task 3 exchanged initial human/AI lanes

and mirrored contact positions. Kitchen allowed 240 turns and three orders per task, changing menus and some starting positions. Realized enrollment-seed frequencies differed between groups in Pong and Kitchen; Kitchen shared only seed 2004. Warehouse scenes did not depend on enrollment seed. Baseline adjustment cannot remove all later scenario differences.

The outcome is the saved final team score, checked against run and final-frame records. Warehouse scoring is

$$S_W = 10D - 10C - 5U - 5B - T_{\text{pen}} - 2H,$$

where D counts deliveries, C collisions, U shutdowns, B shared-charger obstructions, T_{pen} time units charged to the score, and H human-detour units. Early battery shutdown charges remaining time as well as elapsed time. Kitchen scoring is

$$S_K = 100O - T - 5W_i - 20W_d,$$

where O counts timely served dishes, T turns, W_i discarded ingredients/components, and W_d discarded combined dishes. Negative and zero scores and valid early endings are retained. Pong scoring is

$$S_P = 100\frac{n_o + 3n_c}{N_o + 3N_c},$$

where n counts catches and N scheduled balls; subscripts o, c denote ordinary and cooperative balls. Denominators are 120 in Task 1 and 180 in Tasks 2–3. They count scheduled value, not a guaranteed achievable maximum. Pong differences are percentage points; other domains use native points.

B.3 Statistical Analysis

For each domain d and task $t \in \{2, 3\}$, we fit

$$S_{it} = \alpha_{dt} + \tau_{dt}G_i + \beta_{dt}(S_{i1} - \overline{S}_{d1}) + \varepsilon_{it}, \tag{6}$$

where $G_i = 1$ for A and zero for B. The coefficient τ_{dt} is the adjusted A-minus-B difference. Each fit uses one outcome per participant; Task 3 is not adjusted for Task 2. Domains are analyzed separately. HC3 standard errors allow unequal residual variability, using leverage-adjusted squared residuals $e_i^2/(1 - h_{ii})^2$, where e_i is the residual and h_{ii} the leverage. Pointwise 95% intervals and two-sided tests use t_{n-3} (17 degrees of freedom in the primary sample). Holm correction covers the six primary tests; the intervals themselves are not corrected for multiple comparisons.

We also report the descriptive gain contrast

$$\delta_{dt} = \overline{(S_t - S_1)}_{A,d} - \overline{(S_t - S_1)}_{B,d}. \tag{7}$$

It fixes the baseline coefficient to one and can differ from τ_{dt}. A single baseline task cannot establish causal learning trajectories.

Bootstrap summaries use 10,000 resamples: seed 20260925 for the primary sample, 20260926 for 54 retained main participants, and 20260927 for all 66 completers, including excluded records. Resampling occurs within domain and group, keeping each participant's three-task record together. The 2.5th and 97.5th percentiles define intervals. All adjusted-effect resamples were valid. Questionnaire summaries use available item responses; missing or inapplicable responses are not zero. The study was not preregistered. The adjusted model follows an earlier proposed analysis; supplementary and behavioral summaries are exploratory.

B.4 Performance Results

Tables 3–5 report task scores, adjusted effects, and within-participant gains. All six primary confidence intervals include zero, and all Holm-adjusted p values are 1.000. A larger raw gain does not necessarily imply a larger adjusted effect: in Warehouse Task 3, A gained 32.5 points and B gained 30.0, but the adjusted A-minus-B estimate was -6.2 points.

B.5 Sensitivity and Model Checks

We repeated the six models in two samples: the 54 retained main-recruitment participants, and all 66 formal completers, including the six excluded records. The 60-person sample remains primary.

Table 3: Task scores: participant mean ± sample SD.

Domain	Group	Task 1	Task 2	Task 3
Warehouse	A	-136.1 ± 52.2	-85.2 ± 54.3	-103.6 ± 31.4
Warehouse	B	-124.0 ± 41.9	-96.7 ± 54.6	-94.0 ± 36.9
Pong	A	39.42 ± 4.70	41.78 ± 4.21	41.06 ± 5.38
Pong	B	41.50 ± 6.76	43.89 ± 4.68	45.06 ± 5.72
Kitchen	A	-199.9 ± 97.1	-102.2 ± 105.0	-51.9 ± 145.5
Kitchen	B	-176.8 ± 118.8	-147.2 ± 118.0	-80.1 ± 168.3

Each cell has 10 participants. Warehouse and Kitchen use points; Pong uses weighted catch percentage. Higher scores are better. A had explanations only in Task 2.

Table 4: Baseline-adjusted A-minus-B differences.

Domain	Task	Effect	HC3 SE	HC3 95% CI	p	Bootstrap 95% CI
Warehouse	2	18.5	22.0	[-28.0, 65.0]	0.412	[-18.3, 57.5]
Warehouse	3	-6.2	16.0	[-40.1, 27.6]	0.704	[-32.5, 23.8]
Pong	2	-1.7	2.6	[-7.2, 3.7]	0.509	[-5.5, 2.3]
Pong	3	-3.7	2.9	[-9.7, 2.3]	0.211	[-8.3, 1.0]
Kitchen	2	56.3	50.2	[-49.6, 162.2]	0.277	[-24.9, 144.1]
Kitchen	3	32.3	78.8	[-133.9, 198.5]	0.687	[-97.2, 177.1]

Each model uses 20 participants and a t_{17} reference distribution. All six Holm-adjusted p values are 1.000. Effects are points for Warehouse/Kitchen and percentage points for Pong. Bootstrap intervals are supplementary; each uses 10,000 valid resamples and no singular fits.

Table 5: Within-participant gains and A-minus-B gain differences.

Domain	Change	A mean ± SD	B mean ± SD	A−B	95% CI
Warehouse	2−1	50.9 ± 46.6	27.3 ± 55.0	23.6	[-17.5, 66.8]
Warehouse	3−1	32.5 ± 53.9	30.0 ± 37.6	2.5	[-36.7, 40.2]
Pong	2−1	2.36 ± 6.12	2.39 ± 6.81	-0.03	[-5.64, 5.19]
Pong	3−1	1.64 ± 8.42	3.56 ± 6.32	-1.92	[-8.14, 4.22]
Kitchen	2−1	97.7 ± 113.5	29.6 ± 112.0	68.1	[-28.3, 161.9]
Kitchen	3−1	148.0 ± 184.5	96.7 ± 175.2	51.3	[-99.4, 199.0]

Changes retain their sign. The interval describes the between-group difference in gains, using 10,000 participant resamples. Each group has 10 participants. Pong changes are percentage points.

Neither alternative includes outcomes for the 21 noncompleters. No sensitivity comparison is significant after its six-test Holm correction. Pong Task 3 in the 54-person sample has a pointwise interval excluding zero, but its corrected p value is 0.266.

We checked baseline overlap, leverage, Cook's distance, residual error, and estimates obtained by leaving out one participant at a time. These checks did not trigger any new exclusions. Actual scene seeds were uneven in Pong and Kitchen. Baseline adjustment cannot remove all later scene differences, and the small overlap does not support a reliable model with many seed effects.

B.6 Questionnaires and Overall Understanding Ratings

During Task 2, participants rated understanding on a 1–5 scale at reached checkpoints of 20%, 40%, 60%, 80%, and 100% of the turn budget. Early endings still triggered an end rating. The 277 responses comprise five ratings from 46 participants, four from five, and three from nine. All 60 participants supplied an end rating. Unreached checkpoints were not filled in, and repeated ratings were not counted as separate participants.

The final questionnaire used a 1–7 agreement scale. Common items covered predicting the teammate's actions, understanding when it waited or changed plans, knowing how to coordinate, mental effort, and smooth collaboration. A-only items asked whether explanations answered the question, were easy to understand, and helped choose the next action. One Kitchen B participant omitted coor-

Table 6: Sensitivity analyses using alternative participant samples.

Sample	Domain	Task	n_A/n_B	A−B	HC3 95% CI	p_{Holm}
Main-54	Warehouse	2	9/10	22.4	[-24.6, 69.5]	1.000
Main-54	Warehouse	3	9/10	-0.2	[-35.2, 34.7]	1.000
Main-54	Pong	2	10/8	-3.8	[-7.9, 0.4]	0.351
Main-54	Pong	3	10/8	-5.0	[-9.9, -0.1]	0.266
Main-54	Kitchen	2	9/8	25.0	[-90.9, 141.0]	1.000
Main-54	Kitchen	3	9/8	40.1	[-164.2, 244.3]	1.000
All-66	Warehouse	2	11/10	10.9	[-36.1, 58.0]	1.000
All-66	Warehouse	3	11/10	-10.6	[-44.0, 22.8]	1.000
All-66	Pong	2	10/12	0.1	[-5.2, 5.4]	1.000
All-66	Pong	3	10/12	-1.4	[-7.9, 5.2]	1.000
All-66	Kitchen	2	11/12	61.1	[-29.7, 151.9]	1.000
All-66	Kitchen	3	11/12	38.1	[-112.0, 188.2]	1.000

Main-54 removes quota replacements. All-66 includes the six excluded completers and is a sensitivity sample only. Models use t_{n-3} intervals and a separate six-test Holm correction for each sample.

Table 7: Influence diagnostics for the six primary models.

Domain	Task	Max. leverage	Max. Cook's D	Leave-one-out A−B	RMSE
Warehouse	2	0.296	0.406	[9.5, 28.1]	48.4
Warehouse	3	0.296	0.191	[-12.6, -0.2]	32.5
Pong	2	0.461	0.613	[-3.1, -0.4]	4.5
Pong	3	0.461	0.471	[-5.3, -2.6]	5.7
Kitchen	2	0.444	0.524	[37.6, 75.3]	101.1
Kitchen	3	0.444	0.345	[8.0, 58.5]	160.7

Leave-one-out ranges describe refits after removing one participant at a time; they are not confidence intervals. No participant was excluded using these checks.

Table 8: Baseline score ranges and their overlap.

Domain	A range	B range	Overlap	A inside	B inside
Warehouse	[-225.0, -56.0]	[-193.0, -55.0]	[-193.0, -56.0]	9/10	9/10
Pong	[30.0, 47.5]	[26.7, 48.3]	[30.0, 47.5]	10/10	8/10
Kitchen	[-300.0, -35.0]	[-295.0, 93.0]	[-295.0, -35.0]	9/10	9/10

All participants remain in the analysis, including those outside the overlapping range. Pong values are percentages.

Table 9: Recorded scenario seeds and participant counts.

Domain	Group	Seed (count)
Pong	A	731101 (1), 731102 (1), 731103 (1), 731104 (1), 731105 (1), 731106 (1), 731107 (1), 731109 (1), 731111 (1), 731112 (1)
Pong	B	731100 (2), 731101 (2), 731102 (1), 731103 (1), 731104 (1), 731105 (1), 731106 (1), 731108 (1)
Kitchen	A	2004 (1), 2006 (2), 2007 (1), 2010 (1), 2012 (1), 2017 (1), 2018 (1), 2019 (1), 2020 (1)
Kitchen	B	2000 (2), 2001 (1), 2002 (2), 2003 (1), 2004 (1), 2005 (1), 2008 (1), 2009 (1)

Each listed distribution occurs in all three tasks; task-specific scene changes still apply. Warehouse uses 52000, 51000, and 51500 for Tasks 1–3, respectively, with 10 participants per group at each seed.

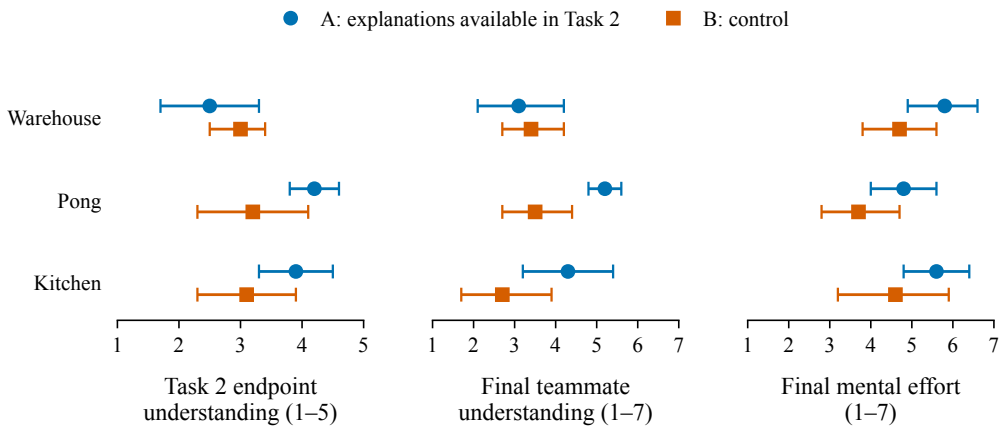

Figure 7: **Perceived understanding and effort.** Means and 95% intervals from 10,000 participant resamples within each domain and group ($n = 10$). Left: Task 2 end understanding (1–5). Middle: final teammate understanding (1–7). Right: final mental effort (1–7). Higher effort scores mean more burden.

dination. Missing and not-applicable answers were excluded for that item, never replaced with zero. No scored objective-understanding test was available. Twelve people supplied open comments; this self-selected subset is not treated as a representative qualitative sample.

Table 10: Task 2 perceived understanding on a 1–5 scale.

Domain	Group	End mean ± SD	End 95% CI	Checkpoint mean ± SD
Warehouse	A	2.5 ± 1.4	[1.7, 3.3]	2.7 ± 1.2
Warehouse	B	3.0 ± 0.8	[2.5, 3.4]	3.0 ± 0.7
Pong	A	4.2 ± 0.6	[3.8, 4.6]	3.8 ± 0.7
Pong	B	3.2 ± 1.5	[2.3, 4.1]	3.1 ± 0.8
Kitchen	A	3.9 ± 1.0	[3.3, 4.5]	3.5 ± 0.9
Kitchen	B	3.1 ± 1.4	[2.3, 3.9]	2.4 ± 1.0

Each row has 10 participants. The last column first averages each participant's available checkpoints, then summarizes those participant averages. End intervals use 10,000 participant resamples.

Overall positive understanding. For the final item "I understood when my teammate waited or changed its plan," we count ratings of 5–7 as positive. All 60 answered, considering all three tasks. A had 16/30 positive responses (53.3%) and B had 9/30 (30.0%): a difference of 23.3 percentage points, or a 77.8% higher proportion relative to B. The latter is $(16 - 9)/9 \times 100\%$, not a 77.8% improvement in understanding ability. This descriptive threshold was added during manuscript revision. Equal group sizes give each domain equal weight, but the direction differs by domain; it is not evidence of a uniform effect or objective understanding.

B.7 Behavior and Task Endings

We describe Warehouse deliveries and shutdowns, Pong ordinary and cooperative catches, and Kitchen completed orders and waste. These measures were fixed for this analysis after initial outcome inspection and are exploratory. Early endings can reduce opportunities. In Kitchen, resolving every order may include expiry rather than successful service. All completed runs, including valid early failures and zero outcomes, retain their scores.

Table 11: All final questionnaire items (1–7 scale).

Domain	Item	n_A	A mean ± SD	n_B	B mean ± SD
Warehouse	Predictability	10	3.2 ± 2.0	10	3.6 ± 1.1
Warehouse	Teammate understanding	10	3.1 ± 1.9	10	3.4 ± 1.3
Warehouse	Coordination	10	3.3 ± 2.0	10	3.7 ± 1.2
Warehouse	Mental effort	10	5.8 ± 1.4	10	4.7 ± 1.5
Warehouse	Smoothness	10	3.2 ± 2.4	10	3.8 ± 1.6
Warehouse	Explanation relevance	9	4.7 ± 1.8	–	–
Warehouse	Explanation clarity	9	4.2 ± 2.3	–	–
Warehouse	Explanation helpfulness	9	4.2 ± 1.9	–	–
Pong	Predictability	10	5.1 ± 1.4	10	4.3 ± 1.4
Pong	Teammate understanding	10	5.2 ± 0.6	10	3.5 ± 1.4
Pong	Coordination	10	4.9 ± 1.2	10	4.2 ± 1.4
Pong	Mental effort	10	4.8 ± 1.4	10	3.7 ± 1.7
Pong	Smoothness	10	5.3 ± 1.4	10	4.4 ± 1.3
Pong	Explanation relevance	9	5.2 ± 1.0	–	–
Pong	Explanation clarity	10	5.3 ± 1.6	–	–
Pong	Explanation helpfulness	9	5.0 ± 1.9	–	–
Kitchen	Predictability	10	4.2 ± 2.3	10	3.0 ± 2.1
Kitchen	Teammate understanding	10	4.3 ± 1.9	10	2.7 ± 1.9
Kitchen	Coordination	10	3.9 ± 1.7	9	3.7 ± 2.2
Kitchen	Mental effort	10	5.6 ± 1.3	10	4.6 ± 2.4
Kitchen	Smoothness	10	3.1 ± 1.6	10	3.1 ± 2.1
Kitchen	Explanation relevance	9	5.1 ± 1.8	–	–
Kitchen	Explanation clarity	9	4.2 ± 1.6	–	–
Kitchen	Explanation helpfulness	9	4.1 ± 1.7	–	–

Higher mental effort means more burden; other items have a positive direction. A dash means that the item was not asked in B. Missing or not-applicable responses are omitted, never scored zero.

Table 12: Positive responses to the teammate-understanding item.

Domain	A positive/n	A (%)	B positive/n	B (%)
Warehouse	2/10	20.0	3/10	30.0
Pong	9/10	90.0	3/10	30.0
Kitchen	5/10	50.0	3/10	30.0
All domains	16/30	53.3	9/30	30.0

Positive means 5–7 on the seven-point scale. Each person contributes one response about all three tasks.

Table 13: Descriptive behavior: mean ± sample SD.

Domain	Measure	Group	Task 1	Task 2	Task 3
Warehouse	Deliveries	A	6.6 ± 3.4	10.7 ± 4.6	9.0 ± 2.8
Warehouse	Deliveries	B	8.7 ± 4.5	9.4 ± 3.9	8.3 ± 3.8
Warehouse	Shutdowns	A	0.8 ± 0.4	0.6 ± 0.5	0.7 ± 0.5
Warehouse	Shutdowns	B	0.6 ± 0.5	0.6 ± 0.5	0.9 ± 0.3
Pong	Ordinary catch rate	A	0.753 ± 0.119	0.819 ± 0.083	0.801 ± 0.100
Pong	Ordinary catch rate	B	0.795 ± 0.131	0.838 ± 0.112	0.854 ± 0.111
Pong	Cooperative catch rate	A	0.035 ± 0.041	0.017 ± 0.032	0.020 ± 0.028
Pong	Cooperative catch rate	B	0.035 ± 0.047	0.040 ± 0.086	0.047 ± 0.106
Kitchen	Completed orders	A	0.6 ± 0.7	1.4 ± 0.8	1.9 ± 1.1
Kitchen	Completed orders	B	0.8 ± 0.9	1.1 ± 1.0	1.6 ± 1.3
Kitchen	Waste events	A	7.2 ± 2.1	4.7 ± 3.0	4.0 ± 3.0
Kitchen	Waste events	B	6.2 ± 2.7	5.1 ± 2.3	4.0 ± 3.2

Each cell has 10 participants. Pong rates are proportions (0–1) with separate denominators for ordinary and cooperative balls. Other measures are counts. These measures describe performance components, not understanding mechanisms.

Table 14: Early-ending counts among 10 runs per group and task.

Domain	Group	Recorded reason	Task 1	Task 2	Task 3
Warehouse	A	Battery shutdown	8	6	7
Warehouse	B	Battery shutdown	6	6	9
Kitchen	A	All orders resolved	5	8	6
Kitchen	B	All orders resolved	5	5	6

The remaining Warehouse/Kitchen runs reached their turn limits. Resolved Kitchen orders may have expired rather than been served. Pong has no saved termination-reason field; all 60 Pong runs are completed. All 180 study runs retain their final scores.

B.8 Explanation Use

The 653 event-triggered prompts ended in 489 skips without a request, 95 "understood" responses without a request, and 69 requested and displayed explanations. The 69 include 30 mandatory guides and 39 optional requests from 17 A participants. There were 58 reading acknowledgments; the other 11 displayed explanations were later skipped without acknowledgment. Prompt triggers, displays, and acknowledgments are different events, and none alone measures attention or understanding.

The separate question-answering channel received 66 requests from 25 A participants: 62 answers, three clarifications, and one unavailable response. Display timestamps exist for 51 answers and all three clarifications; 11 answered requests have no display timestamp. Requests could come from typed text or interface examples. Sixteen requests included simulation, and 14 logs recorded a successful repair attempt; these are processing events, not an independently judged accuracy rate. Provider metadata was available for 65 requests, with configured and returned model names `deepseek-v4-flash` and `deepseek-flash`. No explanation exposure was recorded outside A's Task 2.

Table 15: Explanation requests and recorded displays in Group A.

Domain	Required	Optional	Template shown	Q&A requests	Q&A shown
Warehouse	10	12	22	26	19
Pong	10	8	18	25	23
Kitchen	10	19	29	15	12
Total	30	39	69	66	54

Required/optional columns count template requests. Q&A shown includes answered and clarification responses. A display record does not establish that the answer was read.

C Additional Technical Analysis

For each effective pair, count correct endpoint actions:

$$c_j = \sum_{b \in \{0,1\}} \mathbf{1}[a_{jb}^{T} = a_{jb}^{\mathrm{ref}}], \qquad j \in \mathcal{P}_{\mathrm{eff}}. \tag{8}$$

The program matches 335/430 endpoints, but both actions in only 123/215 pairs. These fractions have different units; ordinary agreement also uses a different population.

In the first missed switch under fixed ordering, changing Robot 2's command from WAIT to UP changes Robot 1's reference from WAIT to UP. Both states reach one leaf, so the program returns WAIT twice despite a changed charger-distance feature. This reveals a lost distinction, not the network's internal reason. The program switches in 153 pairs, including 30 with an incorrect endpoint.

Replay consistency covers specified commands, saved states, fixed controllers, and modeled dynamics; it does not establish predictions of adaptive human behavior.

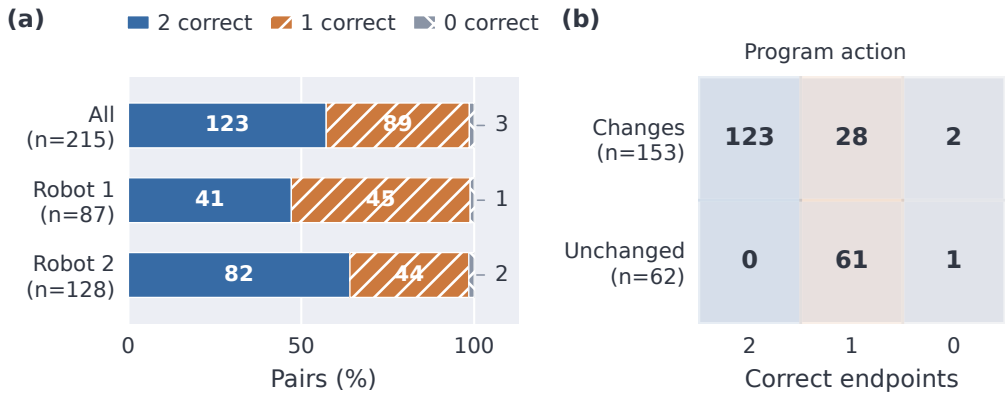

Figure 8: **Program errors on 215 effective pairs.** (a) Pairs with two, one, or zero correct actions, pooled and by robot. (b) Counts by whether the program changes its action. These describe the same technical tests, not additional trials or human-study outcomes.