Explanation Mechanism Influences Human Reliance on Reinforcement Learning Agents

Xuying Zhong
University of Bath
United Kingdom
xz2206@bath.ac.uk

Daniel Beechey
University of Bath
United Kingdom
djeb20@bath.ac.uk

Crescent Jicol
University of Bath
United Kingdom
cj406@bath.ac.uk

Janina Hoffmann
University of Bath
United Kingdom
jah253@bath.ac.uk

Özgür Şimşek
University of Bath
United Kingdom
o.simsek@bath.ac.uk

Abstract

In decision support, user reliance on AI is reflected in whether to adopt or override AI recommendations. Executing an action recommended by a reinforcement learning agent affects future states and can have delayed, uncertain consequences. In two user studies with 170 participants, we examine how explanation mechanisms influence whether users adopt these recommendations. For all explanation mechanisms tested, *appropriate adoption*, defined as adoption of optimal recommendations with authentic explanations, was higher than a no-explanation baseline. Action Advantage Attribution (AAA), an advantage-decomposition mechanism we introduce, yielded the highest appropriate adoption and the largest gap between appropriate adoption and adoption of suboptimal recommendations. Compared with no explanation, AAA increased appropriate adoption by 89%, but it also increased adoption of suboptimal recommendations. Saliency-based explanations promoted adoption indiscriminately, even for suboptimal recommendations. Shapley-based attributions yielded lower adoption of both optimal and suboptimal recommendations compared to the other mechanisms. These findings suggest that explanation choice should depend on the costs of accepting a bad recommendation or rejecting a good one.

1 Introduction

Reinforcement learning is increasingly applied across a range of domains, including clinical support [1, 2], air traffic control [3], industrial process control [4], robotic control [5–7], and autonomous navigation [8]. When reinforcement learning agents support human decision-making, users repeatedly face a consequential choice: adopt the recommended action or override it. The quality of human–AI collaboration depends on *calibrated reliance:* deferring to appropriate recommendations while overriding erroneous ones [9, 10].

Judging a reinforcement learning recommendation requires evaluating an action whose consequences extend beyond the current state. The action changes the distribution of subsequent states and rewards, and its quality depends on expected cumulative return [11]. A locally costly or unintuitive action can therefore be optimal because it enables greater future reward, whereas an immediately attractive action can prevent the agent from reaching better outcomes later. Users must assess recommendations in light of these delayed consequences; misplaced reliance may affect subsequent decisions and outcomes [12].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Studies of decision support based on supervised learning show that feature-based explanations can raise perceived trust without preventing over-reliance or improving error correction [9, 13]. These findings motivate direct evaluation in reinforcement learning, where users must evaluate actions with delayed and uncertain consequences. Explanations of reinforcement learning agents can provide information on aspects such as policies, action values, reward trade-offs, and the expected consequences of actions [14–16]. Whether different explanation mechanisms help users adopt optimal recommendations and reject suboptimal ones, rather than merely increase acceptance, remains unclear.

Previous user studies of reinforcement learning explanations relied on proxy outcomes such as trust ratings [17–19] or mental model accuracy [19–21], rather than the adopt-or-override decisions that directly measure advice taking [22]. When comparing explanations, mechanism frequently co-varies with modality (how explanatory information is communicated) or explanation scope (whether explanations describe individual decisions or the overall policy) [23, 24, 21], making it difficult to determine whether differences in reliance are due to the explanation mechanism, its modality, or its scope. Moreover, placebo effects such as whether explanation presence alone increases adoption regardless of content are rarely investigated [25, 26]. These gaps leave fundamental questions unanswered: Do explanation mechanisms influence reliance behaviour differently? If so, in what ways? As argued in a recent position paper [27], progress in explainable AI risks becoming disconnected from downstream human impact unless such behavioural questions are addressed through controlled user studies.

Our objectives are to quantify the influence of explanation mechanisms on reliance behaviour in decision support that is based on reinforcement learning, to examine factors associated with these effects, and to evaluate their consequences for calibrated reliance. We ask four research questions: (i) To what extent do the presence and type of explanation mechanisms influence adoption behaviour? (ii) What factors drive this relationship? (iii) Does the mechanism maximising appropriate adoption improve calibrated reliance relative to no explanation? (iv) Does providing such an explanation risk inducing over-reliance on flawed recommendations?

To answer these questions, we introduce a placebo-controlled behavioural evaluation framework that compares explanation mechanisms while controlling for agent policy, task, and explanation scope, and systematically varying presentation format, action optimality, and explanation veracity. *Reliance* is measured directly as the observed adopt-or-override decision. We define *appropriate adoption* as the adoption rate for optimal actions with authentic explanations (or optimal actions without explanations in the no-explanation baseline) and *over-reliance* as the rate of adopting a suboptimal action regardless of explanation veracity.

We compare three action-level explanation mechanisms in Study 1 with 100 participants, then benchmark the mechanism yielding the highest appropriate adoption against a no-explanation baseline in Study 2 with 70 participants.

All three explanation mechanisms yielded higher appropriate adoption than the no-explanation baseline.

With saliency-based explanations from SARFA [28], over-reliance exceeded appropriate adoption on average. Qualitative feedback described the maps as visually prominent but difficult to connect to navigation decisions.

Shapley-based attributions from SVERL [29] elicited cautious adoption that varied between authentic and misaligned explanations. Feedback suggested that participants withheld adoption when a heatmap conflicted with their task understanding, even for optimal recommendations. This is the first human evaluation of Shapley-based attributions for reinforcement learning agents.

Action Advantage Attribution (AAA), a reward-decomposition mechanism that we introduce here, which allocates action advantage across reward components, increased appropriate adoption by 89% relative to no explanation, but also increased over-reliance.

Study 1 showed distinct reliance patterns: saliency-based explanations yielded the highest over-reliance, Shapley-based attributions the lowest adoption rates, and AAA the highest appropriate adoption and largest gap between appropriate adoption and over-reliance. Explanation understanding was positively associated with appropriate adoption only for AAA. These comparisons provide the first controlled behavioural evidence that explanation mechanism influences reliance on the same reinforcement learning policy.

2 Background

We train a reinforcement learning agent to solve a Markov decision process (S, A, P, R, γ), where S is a set of states, A a set of actions, P the transition probability matrix, R the reward function, and γ the discount factor [30]. We use Proximal Policy Optimisation (PPO) [31], an actor-critic method whose actor outputs the action probabilities $\pi(a|s)$ for $a \in A$ in state $s \in S$. We use the advantage function to compare taking a particular action with following the policy from the current state. The action-value function $Q^\pi(s, a)$ gives the expected discounted return from taking action a in state s and following policy π thereafter. The state-value function $V^\pi(s)$ gives the expected discounted return from following policy π from state s. Their difference defines the action advantage:

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s). \tag{1}$$

The PPO critic provides a learned estimate $\widehat{V}^\pi(s)$ of the state-value function $V^\pi(s)$.

Several different types of explanations have been proposed for reinforcement learning agents, targeting global policy summaries [32, 33], individual decisions [28, 34], or intermediate representations [35, 36]. We focus on action-level explanations because actions in specific states are the fundamental building blocks of agent behaviour and directly related to the adopt-or-override decision that users need to make. Five families have been proposed, distinguished by the question each answers. *Saliency-based explanations* produce feature importance values via attention scores [37, 38], gradient-based methods [39], perturbation-based approaches [40, 28], or integrated gradients [41]. *Shapley-based attributions* [42, 29, 43, 44] estimate each feature's weighted average marginal contribution across feature subsets. Saliency and Shapley both assign values to individual features but they follow different rationales. A highly salient feature can have low Shapley value, and vice versa. *Reward decomposition explanations* [34, 45] assume reward is the sum of multiple semantically meaningful components and then break it down, making trade-offs among reward components explicit. *Counterfactual explanations* identify minimal state changes that would alter the selected action [46, 47], supporting contrastive reasoning but requiring explicit causal models and trajectory simulation [48]. *Example-based explanations* [24, 49, 32] retrieve similar past behaviours via nearest-neighbour search, prototype learning, or trajectory selection; selecting representative examples without bias remains challenging [50].

Explanations are often presented visually in spatial reinforcement learning domains such as navigation and game-playing [15], where heatmaps efficiently convey feature attributions [51, 52]. Other approaches use rules [53], decision trees [54], or interactive interfaces [55] to explain model decisions. In this study, we focus on visual explanations presented as heatmaps and bar charts to reduce cognitive load and produce explanations accessible to participants without a technical background.

3 Related Work

Research in human-AI interaction distinguishes trust as an attitudinal construct from reliance as observable behaviour [56, 57]. *Calibrated reliance* requires deferring when AI is correct and overriding when erring [9, 10]. In supervised learning, explanations have been extensively studied as tools that influence calibrated trust or reliance [22]. Feature-based explanations often increase trust without reducing over-reliance and can also increase over-reliance [13, 9]. Example-based explanations improve collaboration performance [58]. Confidence scores more consistently support calibration [59]. Global explanations support calibrated reliance better than purely local explanations [60]. An explanation's effect on user trust also depends on model accuracy and how well the explanation reflects the underlying model [61]. A study of human reliance on AI predicting occupations [62] further showed that explanation effects are mechanism-specific, with reliance depending on whether explanations attribute decisions to task-relevant rationale or to social-stereotype features. To our knowledge, no prior study has compared how different explanation mechanisms affect users' adoption of reinforcement learning recommendations.

Evaluations of explainable reinforcement learning have used technical metrics such as faithfulness [63, 64] or user-study outcomes such as action-anticipation accuracy [19, 21, 65], strategy comprehension [66, 20], and perceived collaboration performance [67]. These user-study outcomes assess users' understanding of the agent or perceptions of collaboration but do not directly measure reliance behaviour in response to action recommendations [68, 47]. In addition, existing explainable reinforcement learning comparisons co-vary mechanism with the scope of the explanation, from

individual decisions to the overall policy [21] or presentation format [23, 24], and rarely control for placebo effects [25]. These limitations in evaluation methodology mirror broader concerns in explainable AI, where poorly controlled user studies and proxy outcomes can obscure downstream human impact [27]. Our framework addresses these limitations by (i) controlling for policy, task, and explanation scope while assessing the effects of presentation format, (ii) crossing action optimality with explanation veracity, and (iii) using adoption decisions as the primary outcome.

4 Action Advantage Attribution

Reward decomposition explains action preferences through meaningful reward components, such as reaching a goal or avoiding a collision. Existing methods learn component-specific action values and use them to compare actions [34, 69, 70]. Our studies require explanations of one fixed PPO policy. We therefore developed Action Advantage Attribution (AAA), which allocates action advantage across reward components using the policy's existing critic, without training separate component value functions. AAA can therefore be applied to an existing PPO agent without retraining it.

AAA explains the advantage of taking an action relative to following the policy from the current state, as defined in Equation 1. Let $R(s,a) = \sum_c r_c(s,a)$ be the total immediate reward, where $r_c(s,a)$ is the immediate reward for component c. To present the allocation rule, consider a deterministic transition from state s to state s' under action a. AAA retains each immediate reward component and distributes the value-difference term $\gamma V^\pi(s') - V^\pi(s)$ in proportion to the absolute component magnitudes:

$$A_c^\pi(s,a) = r_c(s,a) + \frac{|r_c(s,a)|}{\sum_{c'} |r_{c'}(s,a)|}\big(\gamma V^\pi(s') - V^\pi(s)\big), \tag{2}$$

where $V^\pi(s)$ is the state-value function, the expected return from state s under policy π; γ is the discount factor; and $A_c^\pi(s,a)$ is the contribution allocated to component c. When at least one immediate reward component is non-zero, the allocation weights sum to one, giving $\sum_c A_c^\pi(s,a) = A^\pi(s,a)$.

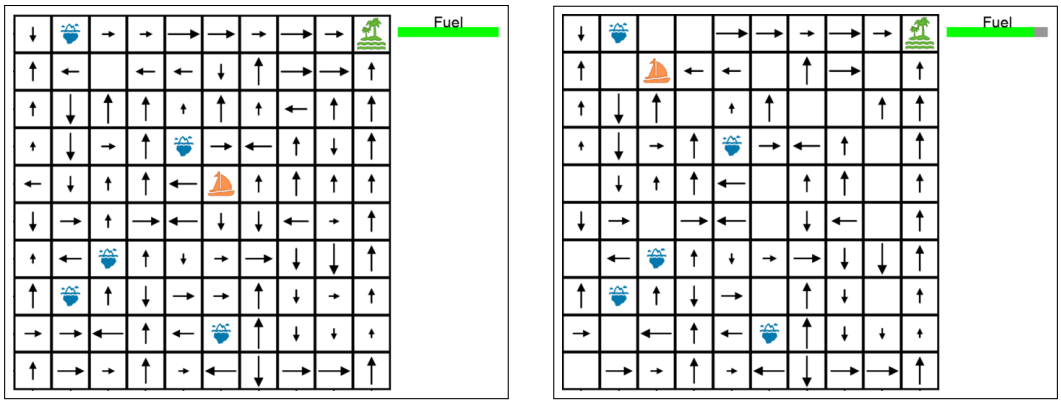
The shared state-value baseline shows how much each action improves or reduces expected return relative to following the policy. AAA reuses the trained PPO critic to estimate this advantage and displays its allocation across reward components for each candidate action. In the stochastic navigation task, we apply the rule to sampled transitions using a one-step temporal-difference estimate. The component contributions sum exactly to that estimate. The implementation is described in Appendix B.3.

5 Methodology

We designed two user studies to examine how different action-level explanation mechanisms influence reliance on action recommendations from reinforcement learning agents. Study 1 compared reliance across three explanation mechanisms in four conditions: SARFA [28] represented as a heatmap, SVERL [29] represented as a heatmap and a barplot, and AAA represented as a barplot. Study 2 compared the condition yielding the highest appropriate adoption in Study 1 against a no-explanation baseline. Both studies used within-participant designs to compare explanation conditions while controlling for individual differences in reliance [71].

5.1 Study 1: Mechanism comparison

The task environment is a ship navigation task we designed for this study. An example state transition is shown in Figure 1. To maximise return, users navigated a 10×10 grid to reach an island while avoiding icebergs and stochastic ocean currents that pushed the ship in the indicated directions. Grid maps were constructed by first defining two viable island-bound current routes, one short and one long, before filling the remaining cells with randomly directed currents. Each current cell received a fixed extinction probability sampled uniformly from $[0.2, 0.8]$ at episode start; after every decision step, the current vanished with that preassigned probability (independently from other current cells). All cells and probabilities were visible, so the task was a fully observable Markov decision process; an ideal decision-maker with the same observations as a participant corresponded to the optimal policy. The reward function comprises three components: fuel penalty (-0.01 per decision action),



(a) Initial state (b) State after executing action `left` from initial state

Figure 1: Example state transition in the stochastic ship navigation environment. The agent moves on a 10×10 grid with stochastic currents (arrow direction shows flow; smaller arrows correspond to higher extinction probability) and reward components fuel penalty, collision penalty, and goal reward. (a) Initial state. (b) Resulting state after executing action `left`. Such individual actions induce state transitions under uncertainty, forming the fundamental elements of what needs to be explained.

collision penalty (-0.1), and goal reward ($+1$). To ensure clear action preferences, test states were selected so that the relative advantage gap $(A_{(1)} - A_{(2)})/(A_{(1)} - A_{(4)})$, where $A_{(1)}$, $A_{(2)}$, and $A_{(4)}$ are the highest, second-highest, and lowest advantages among the four actions, exceeded 0.25. The display was deliberately easy to inspect, while identifying the optimal action still required reasoning forward through stochastic transitions. This controlled moderate difficulty lets participants make an informed independent judgement without making recommendations either obvious or impossible to evaluate [72]. Full environment specification is available in Appendix A.

The agent was trained with PPO for 40 million decision steps. Networks had two hidden layers of 256 units with $\tanh$ activations. Training used learning rate 5×10^{-4}, $\gamma = 0.99$, GAE-$\lambda = 0.95$, clip ratio 0.3, mini-batches of 2048 samples over four epochs. Training ran on an Apple M1, 8-core CPU/GPU, 16GB RAM, completing in around 150 minutes. A pilot study ($n = 20$) confirmed a competence gap: the agent reached the goal in 99% of episodes ($M = 4.5$ steps), human novices at 87% ($M = 6.1$ steps). This one trained policy was used in all experiments. We presented the same agent under a different name in each explanation condition to reduce the influence of participants' impressions from earlier conditions. Further details are provided in Appendix D.2.

As mechanisms to be tested in the studies, we selected one representative from each of three action-level explanation families. These mechanisms can be applied after policy training and can generate explanations in real time. Implementation details are provided in Appendix B, and example explanations are provided in Appendix C. Counterfactual and example-based methods were excluded because their trajectory-level operation would change the explanation scope [73, 74]. All conditions used the same visual modality: SARFA was shown as a map, SVERL as both a map and a bar chart to estimate a map-versus-bar effect, and AAA as a bar chart. Presentation format was therefore varied systematically and included in the statistical model. The four experimental conditions were as follows:

SARFA Map: Specific and Relevant Feature Attribution (SARFA) [28] produces action-contrastive saliency. Compared with other saliency methods [40, 39], SARFA shows greater interpretability in user studies by explicitly isolating action-specific features from those that uniformly affect all actions [28].

SVERL Map: Shapley Values for Explaining Reinforcement Learning (SVERL) [29, 43] is a principled Shapley-based explanation for reinforcement learning, satisfying four mathematical axioms and resolving conceptual problems in earlier applications [15]. Here we presented the feature attributions as spatial heatmaps.

SVERL Bar: The same SVERL attributions were presented as bar charts showing the 10 contributors with the largest absolute attribution magnitudes, to test format variation. SARFA

and SVERL maps have matched visual morphology but are generated by fundamentally different mechanisms, enabling a comparison that holds the map format fixed.

AAA Bar: We selected AAA from the reward-decomposition family and presented its component contributions and total advantage estimate for each candidate action as a bar chart.

We used placebo trials to test whether participants distinguished optimal from suboptimal recommendations and authentic from misaligned explanations. We crossed action optimality (A: optimal; $\bar{A}$: suboptimal) with explanation veracity (E: authentic; $\bar{E}$: misaligned). An optimal action coincided with the shorter route; suboptimal actions did not. Suboptimal trials used states in which the trained policy itself recommended a suboptimal action. Authentic explanations were generated from the same policy for the recommended action. For misaligned explanations, we changed the content while keeping the display visually similar. For SARFA and SVERL, we sampled values from the distribution of authentic feature attributions and randomly assigned them to features in the current state. For AAA Bar, we shuffled the fuel, collision, and goal contributions within each action while preserving each action's total advantage estimate and the ranking of actions. The construction is described in Appendix B.4.

Each participant saw four trial types under every explanation condition: optimal recommendation with an authentic explanation (AE), suboptimal recommendation with an authentic explanation ($\bar{A}E$), optimal recommendation with a misaligned explanation ($A\bar{E}$), and suboptimal recommendation with a misaligned explanation ($\bar{A}\bar{E}$). The last three trial types served as placebo trials.

The procedure began with consent and screening, after which participants completed five practice trials and baseline questionnaires covering demographics, gaming experience [75], perceived task competence [76], technology literacy [77], AI attitude [78], baseline trust [79], and an attention check. The questionnaire items are listed in Appendix E. In the main phase, participants encountered all four explanation conditions in counterbalanced Latin square order, each preceded by a mechanism-specific tutorial that explained the visual encoding and included an example. Within each condition, 30 trials were presented in random order with fixed composition. The trial counts are given in Appendix Table D.1. Trial counts were deliberately unbalanced, with AE trials dominating to reflect ideal decision-support scenarios where the agent mostly recommends optimal actions, maintaining reliance incentives.

Each trial proceeded in two stages. Participants first inspected the state and entered an independent *initial* action after a one-second delay to prevent impulsive responses; the interface then revealed the agent's recommended action with explanation, and participants made a *final* adopt-or-override decision after a one-second delay. To avoid order effects, the two response keys were counterbalanced across participants. To ensure adoption reflected a deliberate behavioural shift rather than coincidental alignment, when initial choice matched the agent recommendation, trials were recorded as *agreement* and excluded from adoption measurement. After each condition, participants completed trust [79] and explanation satisfaction [80] scales, qualitative feedback, and three mechanism-specific comprehension questions. The post-experimental phase included a three-item reinforcement learning comprehension check, personality assessment [81], and a debriefing explaining the use of different agent names and the trial types. The median session duration was 63.5 minutes. The experiment was implemented in PsychoPy [82] and conducted in a controlled laboratory setting on identical workstations.

We collected three types of metrics: outcome variables for each experimental condition, participant-level derived metrics, and self-reported measures.

Outcome variables capture independent task performance, adoption behaviour, and comprehension. *Independent accuracy* is the proportion of trials where the initial decision, made before seeing any agent recommendation, matches the optimal action. *Adoption rate* is the proportion of initially disagreeing trials where a participant switches to adopt the recommendation. We use Adoption$(\cdot)$ to denote the adoption rate conditioned on a specific trial type. *Appropriate adoption* is Adoption(AE). *Over-reliance* is Adoption$(\bar{A})$, across $\bar{A}E$ and $\bar{A}\bar{E}$. Within each explanation condition, an adoption rate was undefined for a participant if their initial choices matched the agent's recommendations on all trials used to calculate that rate. Availability and exclusion-bias analyses are reported in Appendix F. All other trial-type adoption rates are referred to directly via the Adoption$(\cdot)$ notation hereafter. *Com-*

prehension accuracy equals the proportion correct on checks of *explanation understanding*, assessed after each condition, and *reinforcement learning understanding*, assessed after the experiment.

Participant-level derived metrics describe how participants use explanations. *Selective reliance* equals Adoption(AE) − Adoption$(\bar{A})$, the difference between appropriate adoption and over-reliance. *Explanation sensitivity* equals Adoption(AE) − Adoption$(A\bar{E})$, comparing adoption of optimal actions accompanied by authentic versus misaligned explanations. *Decision efficiency* equals Adoption(AE) / $\overline{\text{RT}}$, where $\overline{\text{RT}}$ is mean final-decision *response time* in seconds, relating appropriate adoption to the time taken to decide. Selective reliance, explanation sensitivity, and decision efficiency were calculated only when all adoption rates required for the metric were available. We calculated each derived metric for each participant before averaging across participants.

Self-reported measures capture trust and satisfaction. They were administered using their original Likert scales and rescaled to $[0, 1]$ for analysis. *Trust* has four subscales: perceived risk, competence, benevolence, and reciprocity. *Satisfaction* has nine items.

We recruited 100 participants (73 female, 26 male, 1 non-binary; M_{age}=21.18, SD=4.52) with normal colour perception, through a university research participation scheme in exchange for course credit. The study involved minimal risk (a low-stakes computer task). Participants were informed of the risks and provided informed consent. Ethical approval was obtained from the university research ethics committee (review reference 6229-16881). A priori power analysis using G*Power [83] indicated 48 participants for medium effects (d=0.50, α=0.05, power=0.80, within-participant); we oversampled for robustness. Pre-experimental assessments were administered using their original Likert scales and rescaled to $[0, 1]$ for comparison: gaming experience (M=0.560, SD=0.274), perceived task competence (M=0.648, SD=0.151), technology literacy (AI: M=0.695, SD=0.162; explainable AI: M=0.583, SD=0.179; reinforcement learning: M=0.664, SD=0.181; explainable reinforcement learning: M=0.587, SD=0.162), AI attitude (M=0.497, SD=0.160).

5.2 Study 2: Baseline validation

Study 2 compared AAA Bar, the condition that maximised appropriate adoption in Study 1, against the baseline of having no explanation (NEB), in which participants saw the agent's recommended action but no explanation. The task, agent, trial structure, and metrics were identical to those in Study 1, with two within-participant conditions: AAA Bar and NEB, each consisting of 30 trials, in counterbalanced order. A different sample of 70 participants (51 female, 18 male, 1 non-binary; M_{age}=21.43, SD=4.33) was recruited under the same eligibility criteria; participant characteristics closely matched those of Study 1. The procedure mirrored Study 1, with the NEB condition skipping the tutorial, comprehension check, and explanation satisfaction questionnaire. The median session duration was 26.2 minutes. The same placebo design was applied within the AAA Bar condition (AE, $\bar{A}E$, $A\bar{E}$, $\bar{A}\bar{E}$). NEB varied only action optimality (A, $\bar{A}$). For NEB, appropriate adoption is Adoption(A). Selective reliance equals Adoption(A) − Adoption$(\bar{A})$. Decision efficiency uses Adoption(A) as its numerator.

Cross-study comparisons assessed each remaining explanation condition against no explanation. We compared SARFA Map, SVERL Map, and SVERL Bar in Study 1 with NEB in Study 2 using two-sided Welch independent-samples t-tests. Bonferroni correction covered nine comparisons: three conditions across appropriate adoption, over-reliance, and selective reliance.

Mixed-effects models examined appropriate adoption with a random intercept per participant, adjusting for participant characteristics and testing mechanism, format, and explanation-understanding effects. Pairwise comparisons were Bonferroni-corrected unless stated otherwise. Model specifications and full results are provided in Appendix J. Qualitative feedback was analysed using reflexive thematic analysis [84].

6 Results

Independent accuracy was moderate across conditions (M=0.457, SD=0.087), well above the 0.25 chance level but below ceiling. The participant-level distribution of independent accuracy is shown on the left hand side, in Figure 2. Experimental order did not significantly affect independent accuracy (Friedman $\chi^2(3)$=3.296, p=.348). The explanation mechanism effects remained significant after controlling for independent accuracy.

Comprehension checks in Study 1 showed that at least two of the three mechanism-specific questions were answered correctly by 91% of participants for SARFA Map, 91% for SVERL Map, 96% for SVERL Bar, and 89% for AAA Bar. On the three-item reinforcement learning comprehension check, 98% answered at least two items correctly. These results indicate that most participants had the basic understanding needed to evaluate the agent's recommendations and explanations.

All explanation mechanisms increased appropriate adoption relative to no explanation. In Study 2, the appropriate adoption for AAA Bar was 0.647 (SD=0.295), an 88.63% increase over NEB (M=0.343, SD=0.162). Selective reliance increased from -0.004 under NEB to 0.346 under AAA Bar. Trust was higher under AAA Bar than under NEB, with higher competence, benevolence, and reciprocity ratings and lower perceived risk (all $p < .001$). Trust subscale comparisons are provided in Appendix Table I.4.

In the between-participant cross-study comparisons, SARFA Map, SVERL Map, and SVERL Bar each showed higher appropriate adoption than NEB (all $p < .001$). SVERL Map also showed higher selective reliance ($t(162.07)$=3.81, p=.002, Hedges' g=0.57), whereas neither SVERL Bar nor SARFA Map differed significantly from NEB on this metric (p=.510 and .105, respectively).

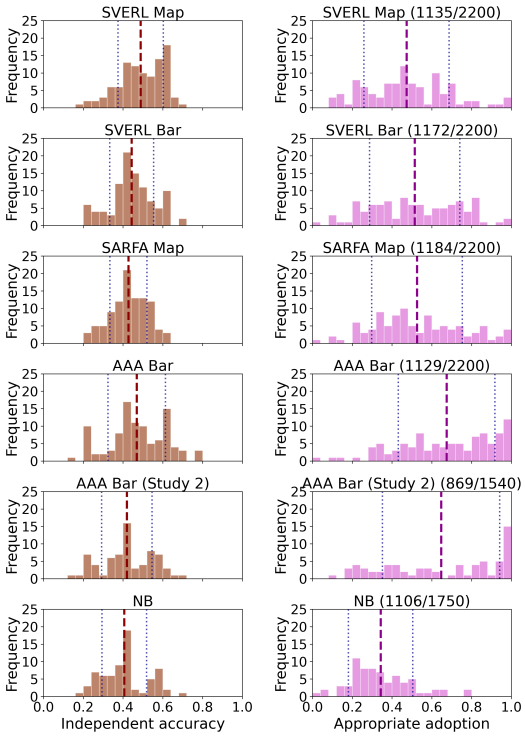


Figure 2: Participant-level distributions of independent accuracy (left) and appropriate adoption (right). Lines show means (dashed) and one standard deviation (dotted); parentheses give non-agreement/total trial counts.

Over-reliance also increased under AAA Bar and SARFA Map relative to no explanation. AAA Bar showed over-reliance of 0.495 (SD=0.375) in Study 2, 45% higher than NEB (M=0.341, SD=0.248). A similar rate was observed under AAA Bar in Study 1 (M= 0.494, SD=0.361). SARFA Map showed the highest over-reliance in Study 1 (M=0.639, SD=0.341) and higher over-reliance than NEB in the cross-study comparison ($t(161.88)$=6.49, $p < .001$, Hedges' g=0.97). Over-reliance under SVERL Bar (M=0.414, SD=0.363) and SVERL Map (M=0.287, SD=0.288) did not differ significantly from that under NEB (both $p = 1.000$).

Appropriate adoption differed across explanation mechanisms in Study 1. AAA Bar elicited the highest appropriate adoption (M=0.674, SD=0.243), 28.14% higher than SARFA Map, 31.13% higher than SVERL Bar and 42.50% higher than SVERL Map. The participant-level distributions of appropriate adoption are shown in Figure 2, on the right hand side.

The appropriate adoption of SVERL Bar did not significantly differ from SVERL Map (Wilcoxon p=.593). In the mixed-effects model, mechanism terms were significant whereas the measured map-versus-bar term was not. The coefficients are reported in Appendix Table J.3. Mechanism effects also remained significant after controlling for gaming experience, perceived competence, AI attitude, independent accuracy, and reinforcement learning understanding (all $p < .001$). Among covariates, higher independent accuracy predicted lower adoption ($p < .001$); this relationship is shown in Figure 6 in the Appendix. Higher technology literacy predicted higher adoption (p=.025).

Selective reliance was highest under AAA Bar in Study 1, with M=0.201 and SD=0.228. The participant-level distributions of selective reliance are shown on the left of Figure 3. SARFA Map produced negative selective reliance (M=-0.114, SD=0.216), driven by Adoption($\bar{A}E$) of 0.704, which exceeded the other conditions by at least 19% (Wilcoxon p=.006). Qualitative feedback from

Study 1 showed that 27 participants found SARFA Map visually prominent but semantically opaque and could not connect it to navigation decisions.

Explanation sensitivity was negative under SVERL Map, with M=−0.138 and SD=0.424. The participant-level distributions of explanation sensitivity are shown in Figure 3, on the right hand side. In qualitative feedback, participants said that the heatmap made individual features easy to understand but was difficult to interpret as a whole, while the bar format was precise but demanding. The explanation sensitivities of the other conditions were all positive. Across Study 1, adoption of suboptimal recommendations (M=0.457) was 12.29% higher than adoption of recommendations with misaligned explanations (Adoption($\bar{E}$)=0.407; t=2.03, p=.043).

AAA Bar had the longest response time (6.29 s) yet the highest decision efficiency (0.215). Only under AAA Bar was explanation understanding positively correlated with appropriate adoption (r=0.223, p=.027). Qualitative feedback from Study 1 showed that 46 participants described AAA Bar as providing a *reason* for the recommendation. The SVERL Map × explanation understanding interaction was significantly negative in the mixed-effects model (b=−0.170, p=.002).

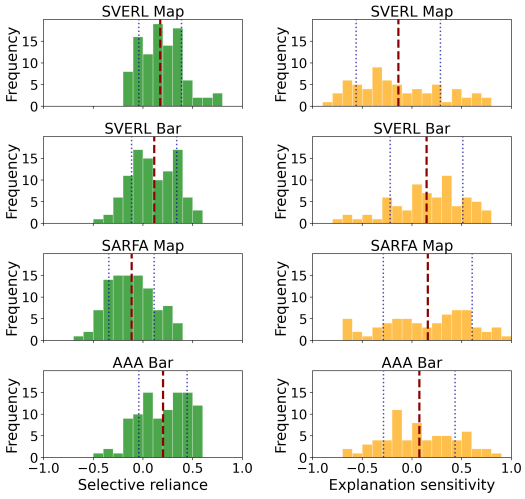


Figure 3: Participant-level distributions of selective reliance (left) and explanation sensitivity (right) across explanation conditions. The dashed lines show the mean, the dotted lines one standard deviation above and below the mean.

AAA Bar produced the highest mean trust in the agent's recommendations and satisfaction with the explanations, followed by SARFA Map. The two SVERL conditions produced lower mean scores on both measures. For both metrics, the difference between SVERL Bar and SVERL Map was not significant. Pairwise comparisons are in Appendix Tables I.2 and I.3.

7 Discussion

Explanations help users decide whether to rely on recommendations from reinforcement learning agents. The comparison with no explanation shows that explanations can increase appropriate adoption. Selective reliance also matters because it captures the gap between appropriate adoption and over-reliance. Even when this gap grows, users may still accept suboptimal recommendations, consistent with prior findings that explanations alone do not eliminate over-reliance [9]. The choice of mechanism should therefore depend on the costs of accepting a bad recommendation and rejecting a good one. Confidence scores can complement explanations when users decide whether to adopt the agent's recommendation, as prior work has shown that they help users calibrate trust [59].

AAA helps users compare actions by showing their advantages and reward-component contributions. The positive relationship between explanation understanding and appropriate adoption, together with participants' descriptions of AAA as providing a reason for the recommendation, suggests that the information provided by AAA supported deliberation. The combination of response time and decision efficiency is also consistent with users spending time evaluating the recommendation. This interpretation aligns with prior work emphasising the importance of distinguishing explanations that justify individual decisions from those that merely describe policy behaviour [63].

Feature attribution mechanisms help users judge recommendations by highlighting features of the current state. The high over-reliance under SARFA, together with feedback about difficulty connecting its maps to navigation decisions, suggests that some participants may have judged recommendations by how plausible the explanations appeared. This interpretation is consistent with automation bias, where users accept automated advice without adequately checking it [85]. Feedback about SVERL suggests that participants sometimes chose not to follow the agent's recommendation when the

explanation did not match their understanding of the task. The comparison with NEB suggests that cautious adoption under SVERL Map can still support selective reliance. SARFA Map and SVERL Map produced different reliance patterns despite sharing the map format. Appropriate adoption did not differ significantly between SVERL Map and SVERL Bar. These comparisons suggest that explanation mechanism, rather than presentation format alone, shapes reliance. These findings extend evidence from supervised learning that explanation mechanisms can shape reliance differently [13, 9].

Self-reported trust and satisfaction should be assessed alongside reliance behaviour. SARFA Map received the second-highest mean ratings on both measures but produced the highest over-reliance in Study 1. This contrast shows that positive ratings alone do not establish calibrated reliance.

Explanation veracity also affects whether users adopt recommendations. We distinguish responses to explanation content from responses to recommendation quality by varying explanation veracity and action optimality separately. Users may reject an optimal recommendation when its explanation is misaligned, so evaluations need to consider explanation veracity as well as recommendation quality. This builds on work showing placebo effects of perceptually plausible yet unfaithful explanations on trust in supervised learning [25] by examining reliance behaviour in reinforcement learning.

The navigation task enabled a controlled comparison of how explanations of the same reinforcement learning policy shape reliance behaviour. Its simple visual presentation helped participants understand the task, while stochastic transitions and action consequences required forward reasoning. The evidence comes from a single low-stakes domain and a predominantly student sample. Applying the same framework to realistic navigation, clinical, or process-control tasks with domain experts would test how the observed reliance patterns change with workload, accountability, and expertise.

AAA extends reward-based explanations to a fixed trained policy by allocating action advantage across reward components. Its weights depend on absolute immediate reward magnitudes: components with zero immediate reward receive no share of the adjustment, even when their delayed rewards may affect the total advantage. In the navigation task, only the fuel penalty is non-zero on non-terminal steps, so the full advantage is allocated to fuel. Future variants could use separately trained component critics [69] to distinguish contributions from delayed rewards. Applying the allocation rule to tasks in which all immediate reward components can be zero would also require a fallback allocation. A further direction is to test explanations that link actions to reachable future states and delayed rewards, using the behavioural framework to assess whether these displays improve calibrated reliance.

Using our placebo-controlled behavioural evaluation framework, we show that explanation mechanisms shape how users adopt or override recommendations from the same reinforcement learning policy. All three tested mechanisms yielded higher appropriate adoption than no explanation, but both appropriate adoption and over-reliance varied across mechanisms. Saliency-based explanations encouraged adoption of both optimal and suboptimal recommendations, while Shapley-based attributions led to more cautious adoption. AAA, the advantage-decomposition mechanism we introduced, yielded the highest appropriate adoption and selective reliance, while still increasing over-reliance relative to no explanation. Explanation mechanisms should therefore be chosen according to the costs of accepting suboptimal recommendations and rejecting optimal ones. Their effects should be evaluated through reliance behaviour alongside self-reported trust and satisfaction metrics.

Acknowledgments and Disclosures

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/X025470/1]. We thank the members of the Bath Reinforcement Learning Laboratory for their constructive feedback and acknowledge support from the University of Bath. Xuying Zhong is a research student in the UKRI Centre for Doctoral Training in Accountable, Responsible and Transparent AI (EP/S023437/1) at the University of Bath.

References

[1] Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24 (11):1716–1720, 2018.

[2] Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1): 16–18, 2019.

[3] Supriyo Ghosh, Sean Laguna, Shiau Hong Lim, Laura Wynter, and Hasan Poonawala. A deep ensemble method for multi-agent reinforcement learning: a case study on air traffic control. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 31, pages 468–476, 2021. doi: 10.1609/icaps.v31i1.15993.

[4] Ammar N Abbas, Chidera W Amazu, Joseph Mietkiewicz, Houda Briwa, Andres Alonso Perez, Gabriele Baldissone, Micaela Demichela, Georgios C Chasparis, John D Kelleher, and Maria Chiara Leva. Analyzing operator states and the impact of AI-enhanced decision support in control rooms: a human-in-the-loop specialized reinforcement learning framework for intervention strategies. *International Journal of Human–Computer Interaction*, 41(12):7218–7252, 2025.

[5] Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 40(4-5):698–721, 2021.

[6] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: world models for physical robot learning. In *Conference on Robot Learning*, pages 2226–2240. PMLR, 2023.

[7] Yifan Xu, Xiao Zhan, Akilu Yunusa Kaltungo, Ming Shan Ng, Tsukasa Ishizawa, Kota Fujimoto, and Clara Cheung. Dialogue based interactive explanations for safety decisions in human robot collaboration, 2026. arXiv preprint arXiv:2604.05896.

[8] Gregory Stein. Generating high-quality explanations for navigation in partially-revealed environments. *Advances in Neural Information Processing Systems*, 34:17493–17506, 2021.

[9] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel 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*, pages 1–16, 2021.

[10] 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.

[11] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: an introduction*. MIT Press, 2 edition, 2018. ISBN 9780262039246.

[12] Claire Glanois, Paul Weng, Matthieu Zimmer, Dong Li, Tianpei Yang, Jianye Hao, and Wulong Liu. A survey on interpretable reinforcement learning. *Machine Learning*, 113(8):5847–5890, 2024.

[13] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: a case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 29–38, 2019.

[14] Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. Explainable reinforcement learning: a survey and comparative review. *ACM Computing Surveys*, 56(7):1–36, 2024.

[15] Yanzhe Bekkemoen. Explainable reinforcement learning (XRL): a systematic literature review and taxonomy. *Machine Learning*, 113(1):355–441, 2024.

[16] 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 2018 Workshop on Explainable AI*, 2018.

[17] Aaquib Tabrez, Shivendra Agrawal, and Bradley Hayes. Explanation-based reward coaching to improve human performance via reinforcement learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 249–257. IEEE, 2019.

[18] Yotam Amitai, Ofra Amir, and Guy Avni. ASQ-IT: interactive explanations for reinforcement-learning agents. *Artificial Intelligence*, 335:104182, 2024. doi: 10.1016/j.artint.2024.104182.

[19] Pradyumna Tambwekar and Matthew Gombolay. Towards reconciling usability and usefulness of policy explanations for sequential decision-making systems. *Frontiers in Robotics and AI*, 11:1375490, 2024.

[20] Joe McCalmon, Thai Le, Sarra Alqahtani, and Dongwon Lee. CAPS: comprehensible abstract policy summaries for explaining reinforcement learning agents. In *Proceedings of the 2022 International Conference on Autonomous Agents and Multiagent Systems*, pages 889–897, 2022.

[21] Risal Shahriar Shefin, Md Asifur Rahman, Thai Le, and Sarra Alqahtani. xSRL: safety-aware explainable reinforcement learning-safety as a product of explainability. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 1932–1940, 2025.

[22] Emma Casolin, Flora D Salim, and Ben Newell. Evaluating the influences of explanation style on human-AI reliance, 2024. arXiv preprint arXiv:2410.20067.

[23] Mark Towers, Yali Du, Christopher Freeman, and Timothy J. Norman. A comparative user evaluation of XRL explanations using goal identification. In *ECAI 2025 Workshop on Evaluating Explainable AI and Complex Decision-Making*, 2025.

[24] Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. How can I explain this to you? an empirical study of deep neural network explanation methods. In *Advances in Neural Information Processing Systems*, volume 33, pages 4211–4222. Curran Associates, Inc., 2020.

[25] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. The impact of placebic explanations on trust in intelligent systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, Glasgow Scotland Uk, 2019. ACM. ISBN 978-1-4503-5971-9. doi: 10.1145/3290607.3312787.

[26] Himabindu Lakkaraju and Osbert Bastani. "how do I fool you?" manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 79–85, 2020.

[27] Matej Pičulin, Bernarda Petek, Irena Ograjenšek, and Erik Strumbelj. Position: explainable AI cannot advance without better user studies. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 81977–81994. PMLR, 2025.

[28] Nikaash Puri, Sukriti Verma, Piyush Gupta, Dhruv Kayastha, Shripad Deshmukh, Balaji Krishnamurthy, and Sameer Singh. Explain your move: understanding agent actions using specific and relevant feature attribution. In *International Conference on Learning Representations*, 2020.

[29] Daniel Beechey, Thomas MS Smith, and Özgür Şimşek. Explaining reinforcement learning with Shapley values. In *International Conference on Machine Learning*, pages 2003–2014. PMLR, 2023.

[30] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[31] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. arXiv preprint arXiv:1707.06347.

[32] Dan Amir and Ofra Amir. Highlights: summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, pages 1168–1176, 2018.

[33] Isaac Lage, Daphna Lifschitz, Finale Doshi-Velez, and Ofra Amir. Exploring computational user models for agent policy summarization. In *IJCAI: Proceedings of the Conference*, volume 28, page 1401, 2019.

[34] Zoe Juozapaitis, Anurag Koul, Alan Fern, Martin Erwig, and Finale Doshi-Velez. Explainable reinforcement learning via reward decomposition. In *IJCAI/ECAI Workshop on Explainable Artificial Intelligence*, pages 47–53, 2019.

[35] Giang Dao, Indrajeet Mishra, and Minwoo Lee. Deep reinforcement learning monitor for snapshot recording. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 591–598. IEEE, 2018.

[36] Omer Gottesman, Joseph Futoma, Yao Liu, Sonali Parbhoo, Leo Celi, Emma Brunskill, and Finale Doshi-Velez. Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions. In *International Conference on Machine Learning*, pages 3658–3667. PMLR, 2020.

[37] Hidenori Itaya, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi, and Komei Sugiura. Visual explanation using attention mechanism in actor-critic-based deep reinforcement learning. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE, 2021.

[38] Qiyuan Zhang, Xiaoteng Ma, Yiqin Yang, Chenghao Li, Jun Yang, Yu Liu, and Bin Liang. Learning to discover task-relevant features for interpretable reinforcement learning. *IEEE Robotics and Automation Letters*, 6(4):6601–6607, 2021.

[39] Rahul Iyer, Yuezhang Li, Huao Li, Michael Lewis, Ramitha Sundar, and Katia Sycara. Transparency and explanation in deep reinforcement learning neural networks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 144–150, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 978-1-4503-6012-8. doi: 10.1145/3278721.3278776.

[40] Samuel Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. Visualizing and understanding Atari agents. In *International Conference on Machine Learning*, pages 1792–1801. PMLR, 2018.

[41] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[42] Erik Strumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11:1–18, 2010.

[43] Daniel Beechey, Thomas Smith, and Özgür Şimşek. A theoretical framework for explaining reinforcement learning with Shapley values, 2025. arXiv preprint arXiv:2505.07797.

[44] Daniel Beechey and Özgür Şimşek. Approximating Shapley explanations in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pages 22278–22312. Curran Associates, Inc., 2025. doi: 10.52202/085713-0752.

[45] Jianhao Wang, Zhizhou Ren, Beining Han, Jianing Ye, and Chongjie Zhang. Towards understanding cooperative multi-agent Q-learning with value factorization. In *Advances in Neural Information Processing Systems*, volume 34, pages 29142–29155. Curran Associates, Inc., 2021.

[46] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2493–2500, 2020. doi: 10.1609/aaai.v34i03.5631.

[47] 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.

[48] Jasmina Gajcin and Ivana Dusparic. Redefining counterfactual explanations for reinforcement learning: overview, challenges and opportunities. *ACM Computing Surveys*, 56(9):1–33, 2024.

[49] Ronilo Ragodos, Tong Wang, Qihang Lin, and Xun Zhou. ProtoX: explaining a reinforcement learning agent via prototyping. *Advances in Neural Information Processing Systems*, 35:27239–27252, 2022.

[50] Antonin Poché, Lucas Hervier, and Mohamed-Chafik Bakkay. Natural example-based explainability: a survey. In *World Conference on eXplainable Artificial Intelligence*, pages 24–47. Springer, 2023.

[51] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 275–285, 2020.

[52] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation, 2019. arXiv preprint arXiv:1902.00006.

[53] Jianfeng Huang, Plamen P Angelov, and Chengliang Yin. Interpretable policies for reinforcement learning by empirical fuzzy sets. *Engineering Applications of Artificial Intelligence*, 91:103559, 2020.

[54] Dweep Trivedi, Jesse Zhang, Shao-Hua Sun, and Joseph J Lim. Learning to synthesize programs as interpretable and generalizable policies. *Advances in Neural Information Processing Systems*, 34:25146–25163, 2021.

[55] Christian Arzate Cruz and Takeo Igarashi. Interactive explanations: diagnosis and repair of reinforcement learning based agent behaviors. In *2021 IEEE Conference on Games (CoG)*, pages 01–08. IEEE, 2021.

[56] John D Lee and Katrina A See. Trust in automation: designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004.

[57] Kevin Anthony Hoff and Masooda Bashir. Trust in automation: integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, 2015.

[58] 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):1–32, 2023.

[59] Yunfeng Zhang, Q Vera Liao, and Rachel KE 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*, pages 295–305, 2020.

[60] Carina Newen, Daniel Bodemer, Sonja Glantz, Emmanuel Müller, Magdalena Wischnewski, and Lenka Schnaubert. Uncertainty awareness and trust in explainable AI-on trust calibration using local and global explanations. In *2025 IEEE International Conference on Data Mining (ICDM)*, pages 623–631. IEEE, 2025.

[61] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. How model accuracy and explanation fidelity influence user trust in AI. In *IJCAI 2019 Workshop on Explainable Artificial Intelligence*, 2019.

[62] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. Explanations, fairness, and appropriate reliance in human-AI decision-making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024.

[63] Mythreyi Velmurugan, Chun Ouyang, Catarina Moreira, and Renuka Sindhgatta. Developing a fidelity evaluation approach for interpretable machine learning, 2021. arXiv preprint arXiv:2106.08492.

[64] Miquel Miró-Nicolau, Antoni Jaume-i Capó, and Gabriel Moyà-Alcover. Assessing fidelity in XAI post-hoc techniques: a comparative study with ground truth explanations datasets. *Artificial Intelligence*, 335: 104179, 2024.

[65] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Matthew Olson, Alan Fern, and Margaret Burnett. Mental models of mere mortals with explanations of reinforcement learning. *ACM Trans. Interact. Intell. Syst.*, 10(2):15:1–15:37, 2020. doi: 10.1145/3366485.

[66] Tobias Huber, Katharina Weitz, Elisabeth André, and Ofra Amir. Local and global explanations of agent behavior: integrating strategy summaries with saliency maps. *Artificial Intelligence*, 301:103571, 2021.

[67] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. Expanding explainability: towards social transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2021.

[68] Herman Yau, Chris Russell, and Simon Hadfield. What did you think would happen? explaining agent behaviour through intended outcomes. In *Advances in Neural Information Processing Systems*, volume 33, pages 18375–18386. Curran Associates, Inc., 2020.

[69] Harm Van Seijen, Mehdi Fatemi, Joshua Romoff, Romain Laroche, Tavian Barnes, and Jeffrey Tsang. Hybrid reward architecture for reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[70] Stuart J Russell and Andrew Zimdars. Q-decomposition for reinforcement learning agents. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 656–663, 2003.

[71] Vivian Lai, Yiming Zhang, Chacha Chen, Q Vera Liao, and Chenhao Tan. Selective explanations: leveraging human input to align explainable AI. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–35, 2023.

[72] Maximilian Moll and John Dorsch. A systematic review of human-centered explainability in reinforcement learning: transferring the RCC framework to support epistemic trustworthiness. *Human-Intelligent Systems Integration*, pages 1–10, 2025.

[73] Shuyang Dong, Shangtong Zhang, and Lu Feng. Counterfactual explanations for continuous action reinforcement learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 5039–5046, 2025.

[74] Shripad Vilasrao Deshmukh, Arpan Dasgupta, Balaji Krishnamurthy, Nan Jiang, Chirag Agarwal, Georgios Theocharous, and Jayakumar Subramanian. Explaining RL decisions with trajectories. In *The Eleventh International Conference on Learning Representations*, 2023.

[75] Anca Salagean, Daniel A Ledger, Dominic Potts, Tarini Sehgal, Sarah Binesmael, Koner Kalkanel, Hussain Karim, Michelle Wu, George Fletcher, Eamonn O'Neill, et al. Virtual selves, real emotions: the impact of photorealistic avatar personalization on emotion in virtual reality. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–29, 2026.

[76] Center for Self-Determination Theory. Intrinsic Motivation Inventory (IMI), 2022.

[77] Simone Grassini. A psychometric validation of the PAILQ-6: perceived artificial intelligence literacy questionnaire. In *Proceedings of the 13th Nordic Conference on Human-Computer Interaction*, pages 1–10, 2024.

[78] Cornelia Sindermann, Peng Sha, Min Zhou, Jennifer Wernicke, Helena S Schmitt, Mei Li, Rayna Sariyska, Maria Stavrou, Benjamin Becker, and Christian Montag. Assessing the attitude towards artificial intelligence: introduction of a short measure in German, Chinese, and English language. *KI-Künstliche Intelligenz*, 35(1):109–118, 2021.

[79] Siddharth Gulati, Sonia Sousa, and David Lamas. Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, 38(10):1004–1015, 2019. doi: 10.1080/0144929X.2019.1656779.

[80] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, 5:1096257, 2023.

[81] Raul Costa Mastrascusa, Matheus Loli de Oliveira Fenili Antunes, Nathalia Saraiva de Albuquerque, Sara Luísa Virissimo, Marcela Foletto Moura, Bibiana Vieira Marques Motta, Wagner de Lara Machado, Carmen Moret-Tatay, and Tatiana Quarti Irigaray. Evaluating the complete (44-item), short (20-item) and ultra-short (10-item) versions of the Big Five Inventory (BFI) in the Brazilian population. *Scientific Reports*, 13(1):7372, 2023.

[82] Jonathan Peirce, Jeremy R Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*, 51:195–203, 2019.

[83] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191, 2007.

[84] Virginia Braun, Victoria Clarke, Nikki Hayfield, Louise Davey, and Elizabeth Jenkinson. Doing reflexive thematic analysis. In *Supporting Research in Counselling and Psychotherapy: Qualitative, Quantitative, and Mixed Methods Research*, pages 19–38. Springer, 2023.

[85] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012.

Appendix

A Environment specification

The ship navigation task is formalised as a finite-horizon Markov decision process (MDP) on a 10×10 grid. Each cell is one of four types: *Plain water*: safe cells where the ship can choose an action, *Iceberg*: terminal hazard; entering an iceberg cell ends the episode with a collision penalty, *Ocean current*: once the ship enters a current, it is automatically pushed one cell at a time in a fixed direction (up, right, down, or left) until it reaches plain water, the goal, or an iceberg, *Island*: terminal goal; reaching it yields a large positive reward and ends the episode.

The ship always starts at coordinate $(5, 5)$ and is given 15 fuel points. At each decision state (plain water), an action $a \in \{\text{up}, \text{right}, \text{down}, \text{left}\}$ is selected, which consumes 1 fuel. After the action, the ship moves one cell in the chosen direction; if that cell is a current, it is pushed along that current until reaching a non-current cell, the goal, or an iceberg. Each current cell has a fixed per-action extinction probability drawn uniformly from $[0.2, 0.8]$ at the beginning of an episode; when a current extinguishes, it becomes plain water for the remainder of the episode.

The reward signal is additively decomposed into three components:

$$r_{\text{fuel}} = -0.01, \qquad r_{\text{iceberg}} = -0.1, \qquad r_{\text{island}} = +1,$$

where r_{fuel} is incurred for every decision action, r_{iceberg} for a collision with an iceberg, and r_{island} for reaching the island. The episode terminates when the ship reaches the island, hits an iceberg, or runs out of fuel. All cells are fully observable, and transitions are stochastic due to the random extinction of currents.

B Explanations

All explanations were generated post-hoc from one trained PPO policy. We denote the set of grid cells by $\mathcal{F}$.

B.1 SARFA

SARFA [28] generates action-specific saliency maps by perturbing each editable cell $i \in \mathcal{F}$ and measuring the impact on the probability distribution over actions. For current cells we test two perturbations: removing the current (turning it into plain water) and reversing its direction. For plain-water cells we test four perturbations: conversion to currents in each direction.

For a perturbation at cell i, let $s^{(i)}$ denote the perturbed state and $\hat{a}$ the recommended action. Following SARFA [28], relevance is defined as

$$K_i = \frac{1}{1 + D_{\mathrm{KL}}(P'_{\neg\hat{a}} \,\|\, P_{\neg\hat{a}})},$$

where $P_{\neg\hat{a}}$ and $P'_{\neg\hat{a}}$ are the distributions over the remaining actions before and after perturbation, renormalised after excluding $\hat{a}$, and D_{KL} is the Kullback–Leibler divergence. Specificity is measured by $\Delta p_i = \pi(\hat{a} \mid s) - \pi(\hat{a} \mid s^{(i)})$.

The original SARFA score combines specificity and relevance as $2K_i\Delta p_i/(K_i + \Delta p_i)$. To retain both supporting and opposing features in the displays, we use the signed extension

$$\mathrm{sal}(i) = \mathrm{sgn}(\Delta p_i)\, \frac{2K_i|\Delta p_i|}{K_i + |\Delta p_i|}.$$

Positive values indicate features whose perturbation decreases the probability of the recommended action (the feature supports the action), while negative values indicate features whose perturbation increases it (the feature opposes the action).

For each cell, we select the maximum absolute saliency value across all tested perturbations. To account for stochastic dynamics, current cells are weighted by their location-specific extinction probability; non-current cells are weighted by the mean extinction probability across all currents. Finally, scores are normalised to the range $[-1, 1]$ by dividing by the maximum absolute value across all cells. The resulting values are displayed as a spatial heatmap (*SARFA Map*).

B.2 SVERL

Shapley Values for Explaining Reinforcement Learning (SVERL) [29, 43] formulates feature attribution as a cooperative game played by the features. In our environment, we represent the grid cells in $\mathcal{F}$ as the individual players.

A Shapley value quantifies a feature's average marginal contribution to the agent's action probability across all possible combinations of known and unknown features. Let $\pi(s) = (\pi(a \mid s))_{a \in A}$ denote the vector of action probabilities under the trained policy at state s. SVERL formalises the attributions by defining a characteristic function, $\tilde{\pi}_s(\mathcal{C})$, which represents the expected action-probability vector for a state s given only a known subset of features (a coalition) $\mathcal{C} \subseteq \mathcal{F}$:

$$\tilde{\pi}_s(\mathcal{C}) = \mathbb{E}[\pi(S) \mid S_{\mathcal{C}} = s_{\mathcal{C}}]$$

The Shapley value $\phi_i(s)$ for a grid cell i is then defined as its weighted marginal contribution across all possible subsets $\mathcal{C} \subseteq \mathcal{F} \setminus \{i\}$:

$$\phi_i(s) = \sum_{\mathcal{C} \subseteq \mathcal{F} \setminus \{i\}} \frac{|\mathcal{C}|!(|\mathcal{F}| - |\mathcal{C}| - 1)!}{|\mathcal{F}|!} \big[\tilde{\pi}_s(\mathcal{C} \cup \{i\}) - \tilde{\pi}_s(\mathcal{C})\big]$$

Scalable approximation is needed because exact computation requires evaluating all $2^{|\mathcal{F}|}$ possible coalitions, which is computationally intractable. We therefore use the scalable parametric approximation method introduced in FastSVERL [44].

First, we approximate the characteristic function by training a dedicated characteristic model, $\hat{\pi}(s|\mathcal{C})$. This model learns to predict the expected action probabilities for a state s where the values of cells not included in the coalition $\mathcal{C}$ are masked by -1. This specific masking value is chosen because it falls strictly outside the valid state space [44].

Next, we train a parametric explainer network $\hat{\phi}_\theta : \mathbb{R}^{|\mathcal{F}|} \to \mathbb{R}^{|\mathcal{F}| \times |A|}$ to directly predict the per-feature Shapley values for all cells simultaneously, where $|A|$ is the size of the action space. For each training state s, we sample coalitions $\mathcal{C}$ from the following distribution [44]:

$$p(\mathcal{C}) \propto \frac{n-1}{\binom{n}{|\mathcal{C}|}\,|\mathcal{C}|\,(n - |\mathcal{C}|)}, \qquad n = |\mathcal{F}|, \quad 0 < |\mathcal{C}| < n.$$

We train $\hat{\phi}_\theta$ using a weighted least-squares objective, minimising the mean squared error between the output of the characteristic model and the sum of the predicted Shapley values for the cells included in the sampled coalition:

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,\mathcal{C})}\left[\left\| \hat{\pi}(s|\mathcal{C}) - \hat{\pi}(s|\emptyset) - \sum_{i \in \mathcal{C}} \hat{\phi}_\theta(s)_i \right\|^2\right]$$

At inference, the fully unmasked state s is fed to the explainer network. To ensure the approximations satisfy the Shapley efficiency axiom, an additive correction is applied to the raw outputs so that the sum of the feature attributions exactly equals the total change in action probability from knowing no features to knowing all features. Denoting these corrected approximations by $\hat{\phi}_i(s)$, we have:

$$\sum_{i \in \mathcal{F}} \hat{\phi}_i(s) = \pi(s) - \hat{\pi}(s|\emptyset)$$

The final approximate Shapley values $\{\hat{\phi}_i(s)\}$ are rendered either as a spatial heatmap over the grid (*SVERL Map*) or as a bar chart of the top-10 contributors ranked by absolute attribution magnitude (*SVERL Bar*).

B.3 AAA

We implemented the AAA allocation rule in Equation 2 using the trained PPO critic and sampled transitions. The target quantity is the action advantage $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$, which compares taking action a in state s with following policy π at that state.

For each candidate action, a sampled transition (s, a, r, s') supplies the next state s' and the observed immediate reward $r = \sum_c r_c$, where r_c is the observed reward for component $c \in \{\text{fuel}, \text{iceberg}, \text{island}\}$. The transition includes the automatic current movement that follows the action. As in the Background, $\widehat{V}^\pi(s)$ is the PPO critic's learned estimate of the state-value function $V^\pi(s)$, the expected return from state s under policy π, and γ is the discount factor. We compute the one-step temporal-difference estimate of action advantage as

$$\widehat{A}^\pi(s, a) = r + \gamma \widehat{V}^\pi(s') - \widehat{V}^\pi(s),$$

and apply the same allocation rule to obtain

$$\widehat{A}^\pi_c(s, a) = r_c + \frac{|r_c|}{\sum_{c'} |r_{c'}|} \big(\gamma \widehat{V}^\pi(s') - \widehat{V}^\pi(s)\big). \tag{B.1}$$

Here, $\widehat{A}^\pi_c(s, a)$ is the contribution allocated to reward component c. The component contributions sum exactly to the computed advantage estimate: $\sum_c \widehat{A}^\pi_c(s, a) = \widehat{A}^\pi(s, a)$. Absolute reward magnitudes prevent cancellation between positive and negative components when computing the allocation weights. In this task, the fuel penalty is always non-zero, so the denominator is defined.

On non-terminal steps, only the fuel component has a non-zero immediate reward, giving $\widehat{A}^\pi_{\text{fuel}}(s, a) = \widehat{A}^\pi(s, a)$. Differences between actions therefore remain visible in the fuel bars. Here, the fuel contribution is the share of the overall advantage allocated to fuel; it also incorporates information about delayed goal and collision rewards through the critic's value estimates.

The three component contributions together with their sum are displayed in a bar chart for each candidate action (*AAA Bar*).

B.4 Misaligned explanations

To isolate the effect of explanation content from mere visual presence, we constructed placebo explanations that preserve the statistical appearance of authentic explanations while disrupting their feature attributions or reward-component composition. For AAA Bar, only the component composition changes; the action totals and their ranking remain informative and unchanged.

SARFA and SVERL placebos are constructed by first compiling the empirical distribution of attribution values for each mechanism from states sampled during policy training. For a placebo trial we draw values from this distribution and randomly assign them to the grid positions of the current state. The resulting heatmap or bar chart retains the magnitude distribution and sparsity patterns of authentic explanations, yet the highlighted locations are unrelated to the agent's actual rationale.

AAA Bar placebos preserve the action ranking by retaining the total one-step temporal-difference estimate of each action while randomly permuting the three component contributions within that action. Consequently, the chart preserves the authentic ranking of action totals, but makes the component composition spurious, removing its semantic link to the agent's evaluation.

C Example explanations

The ship navigation state is shown in Figure 4. The four explanation conditions for this state are illustrated in Figure 5. In this state, the agent recommends optimal action up.

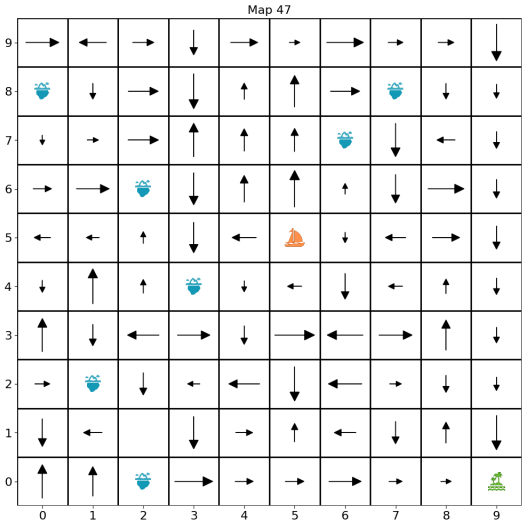


Figure 4: Example state. The grid shows the ship at the centre, icebergs (blue), the island (green), and ocean currents (arrows).

D User Experiment Design

D.1 Experiment procedure

Each trial followed the structure below:

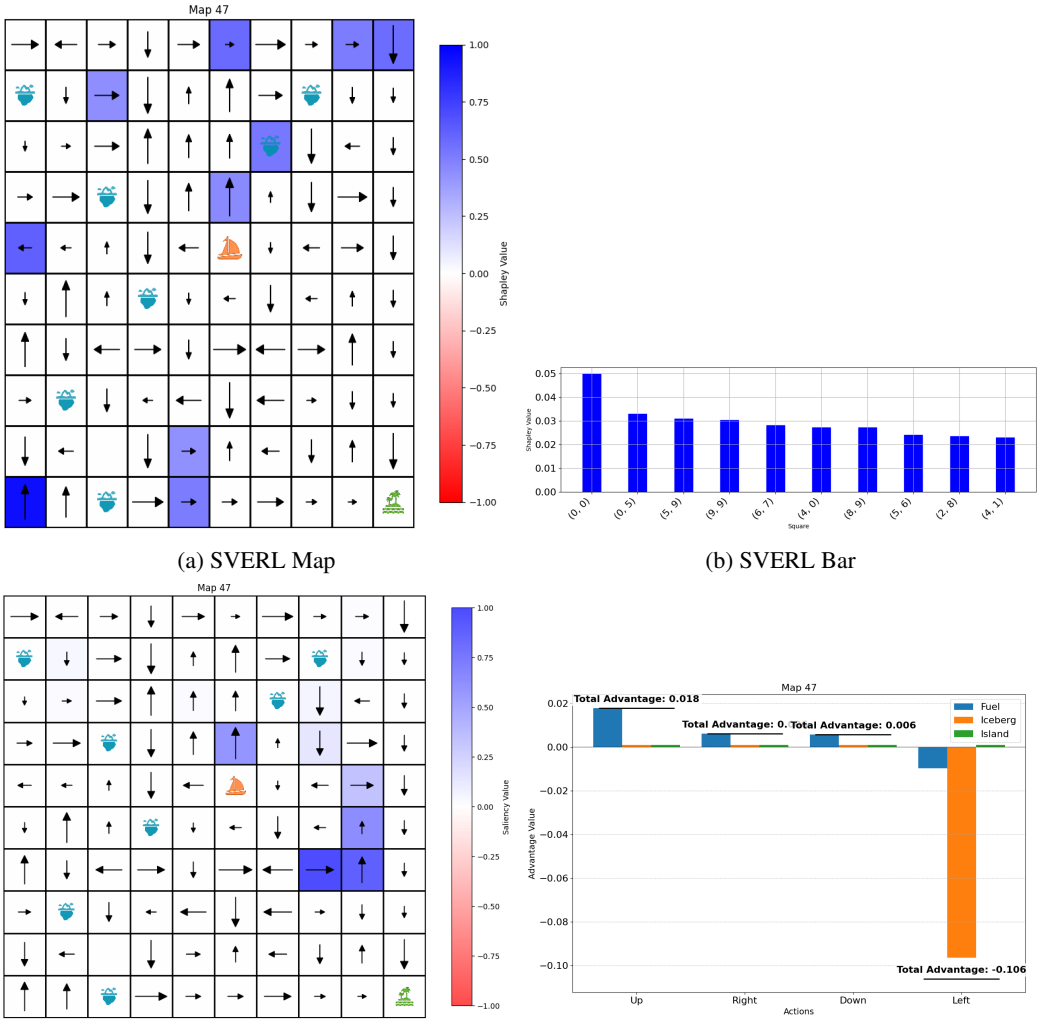
1. The initial decision could be made one second after the state appeared. The grid showed the ship, icebergs, island, and currents. Participants studied the state and answered the prompt "Which action do you choose?" by selecting one of four options: up, right, down, or left.
2. The final decision could be made one second after the agent's recommendation and explanation appeared. The interface showed the recommended action and the explanation assigned to that trial. For NEB, only the action was shown. If the initial action differed from the recommendation, the options were "I switch to [agent]'s recommended action [action]" and "I stick to my original action [action]". If the initial action matched the recommendation, the options read "You agreed with agent X" and "I change my mind". Response keys were counterbalanced to avoid order effects.

D.2 Agent names across explanation conditions

Each explanation condition presented the same trained PPO agent under a different name to reduce the influence of participants' impressions from earlier conditions. Names were assigned in fixed order (Aurora, Breeze, Comet, Drift) and counterbalanced with explanation conditions using Latin square design. All conditions used identical policy parameters and behaviour; only the displayed name varied.

D.3 Trial composition

The four trial types in each condition have fixed composition as shown in Table D.1; trial counts are deliberately unbalanced, with AE trials dominating to reflect ideal decision-support scenarios where the agent mostly recommends optimal actions, maintaining reliance incentives. Suboptimal trials used states in which the trained policy itself recommended a suboptimal action. Authentic explanations were generated from the same policy for the recommended action.



(a) SVERL Map

(b) SVERL Bar

(c) SARFA Map

(d) AAA Bar

Figure 5: Example explanation for the same state.

Table D.1: Summary of action optimality, explanation veracity, and trial counts across conditions

	AE	$\bar{A}E$	$A\bar{E}$	$\bar{A}\bar{E}$
Action	Optimal	Suboptimal	Optimal	Suboptimal
Explanation	Authentic	Authentic	Misaligned	Misaligned
Number of trials	22	3	3	2

E Questionnaires

All questionnaires were administered in English. Exact items are listed below.

Pre-experimental phase

Demographics. Gender (Male, Female, Non-binary, Other, Prefer not to say), age (Free text between 18 and 99), ethnicity (“Asian”, “Black, Caribbean or African”,“Mixed or multiple ethnic groups”, White, Other ethnic group).

Gaming experience. Single item: “Please rate the amount of experience you have with computer games.” Response scale 0 (No experience) – 10 (Expert).

Perceived task competence. Adapted from the Intrinsic Motivation Inventory [76]. Six items, 7-point Likert (Not at all true, Somewhat true, Very true):

1. I think I am pretty good at this ship navigation game.
2. I think I did pretty well at this ship navigation game, compared to other participants.
3. After playing this ship navigation game for a while, I felt pretty competent.
4. I am satisfied with my performance at this ship navigation game.
5. I was pretty skilled at this ship navigation game.
6. This was a game that I couldn't play very well. (reversed)

AI literacy. PAILQ-6 [77]. Six items, 7-point Likert (Strongly Disagree, Disagree, Somewhat Disagree, Neutral, Somewhat Agree, Agree, Strongly Agree):

1. I understand the basic concepts of artificial intelligence (AI).
2. I believe I can contribute to AI projects.
3. I can judge the pros and cons of AI.
4. I keep up with the latest AI trends.
5. I'm comfortable talking about AI with others.
6. I can think of new ways to use existing AI tools.

The same six items were repeated for explainable AI (XAI), reinforcement learning (RL), and explainable RL (XRL).

Attitudes towards AI. Five items [78], 11-point Likert (0 Strongly disagree – 10 Strongly agree):

1. I fear artificial intelligence.
2. I trust artificial intelligence.
3. Artificial intelligence will destroy humankind.
4. Artificial intelligence will benefit humankind.
5. Artificial intelligence will cause many job losses.

Baseline trust in the ship navigation agent. Human–Computer Trust Scale [79], 12 items, 5-point Likert (1 Strongly disagree – 5 Strongly agree), covering perceived risk, competence, benevolence, and reciprocity:

- I believe that using RL agents could lead to negative outcomes, such as choosing an unsafe or inefficient route.
- I feel I need to be cautious when following reinforcement learning agent's navigation decisions.
- Relying on the ship navigation reinforcement learning agent feels risky.
- I believe that reinforcement learning agent is designed to prioritise ship navigation safety and efficiency.
- I believe that the ship navigation reinforcement learning agent will try its best to navigate the ship safely and efficiently if I rely on it.
- I believe that the ship navigation reinforcement learning agent considers factors that matter to me, such as avoiding icebergs and reaching the destination efficiently.
- I think that the ship navigation reinforcement learning agent is capable and effective at finding the safest and most efficient route.
- I think that the ship navigation reinforcement learning agent performs its role of navigating ships through safe and efficient routes very well.
- I believe that the ship navigation reinforcement learning agent has all the necessary functionalities to navigate a ship safely and efficiently.

- If I were using the ship navigation reinforcement learning agent in a real situation, I think I would be able to rely on it completely.
- I feel confident that I can always rely on the ship navigation reinforcement learning agent to navigate the ship safely and efficiently.
- I trust the decisions provided by the ship navigation reinforcement learning agent.

Attention check. A single embedded item asking participants to select a specific response.

Main experiment phase

After each explanation condition the following were administered:

Trust in the agent. The same 12-item Human–Computer Trust Scale, with the agent name adjusted to the current condition.

Explanation satisfaction. Explanation Satisfaction Scale [80] with agent name adjusted to the current condition, 9 items, 5-point Likert (I Disagree Strongly – I Agree Strongly). Example items:

- From the explanation, I understand how the [agent] works.
- This explanation of how the [agent] works is satisfying.
- This explanation of how the [agent] works has sufficient detail.
- This explanation seems complete.
- This explanation shows me how accurate the [agent] is.
- This explanation shows me how reliable the [agent] is.
- This explanation tells me how to use the [agent]
- This explanation is useful to my goals.
- This explanation helps me know when I should trust and not trust the [agent].

Explanation comprehension check. Three multiple-choice questions assessing explanation understanding for the condition just viewed (e.g., what the colours in the heatmap represent, what the bar heights encode). Exact questions are available upon request.

Free-text feedback. "Please give us your thoughts on this explanation, including both positive and negative considerations."

Post-experimental phase

Reinforcement learning comprehension check. Three multiple-choice questions on basic RL concepts (e.g., definition of a policy, what a reward signal is). Exact questions available upon request.

Placebo awareness. "During the experiment, did you feel some explanations were unrelated to the ship's actual decision-making process?" (Yes / No).

Free-text feedback. "Please describe your thoughts on the explanations. Did any seem unhelpful or unrelated? If so, why?"

Personality. 10-item Big Five Inventory (BFI-10) [81], 5-point Likert (Completely Disagree – Completely Agree). I see myself as someone who...

1. Who is talkative, communicative.
2. Likes to cooperate with others.
3. Is original, always having new ideas.
4. Is inventive, creative.
5. Is helpful and assists other people.
6. Does things in an efficient way.
7. Is sociable, extroverted.
8. Is a reliable worker.
9. Gets tense often.
10. Gets nervous easily.

F Effective sample sizes for the adoption rate

The adoption rate is defined only on trials where the participant initially disagreed with the agent. Within each explanation condition, an adoption rate is undefined for a participant whose initial choices match the agent's recommendations on all trials used to calculate that rate, because there are no initially disagreeing trials. The number and percentage of participants with at least one non-agreement trial for each condition and trial type are reported in Table F.1. The majority of participants contributed adoption rates for every explanation condition and trial category. For the primary outcome, appropriate adoption, all participants contributed at least one non-agreement trial among the 22 AE trials in each condition. Even for $\bar{A}\bar{E}$ trials in Study 2, which had the fewest contributing participants, over 71% contributed an adoption rate.

Table F.1: Number (percentage) of participants contributing an adoption rate per condition and trial type. Participants contributed if they had at least one non-agreement trial. Sample sizes were $n = 100$ in Study 1 and $n = 70$ in Study 2. For NEB, entries under AE and $\bar{A}E$ correspond to A and $\bar{A}$ trials, respectively.

Condition	AE	$A\bar{E}$	$\bar{A}E$	$\bar{A}\bar{E}$
SVERL Map	100 (100%)	75 (75%)	94 (94%)	96 (96%)
SVERL Bar	100 (100%)	99 (99%)	77 (77%)	92 (92%)
SARFA Map	100 (100%)	91 (91%)	81 (81%)	86 (86%)
AAA Bar	100 (100%)	85 (85%)	87 (87%)	93 (93%)
AAA Bar (S2)	70 (100%)	62 (89%)	66 (94%)	50 (71%)
NEB	70 (100%)	—	68 (97%)	—

Table F.2: Statistical results by condition, trial type and metric.

Condition	Trial Type	Metric	Mean_0	Mean_+	Diff	p
SVERL Map	$\bar{A}E$	Adoption(AE)	0.520	0.470	-0.050	0.6736
	$\bar{A}E$	Independent Accuracy	0.478	0.489	0.012	0.5451
	$A\bar{E}$	Adoption(AE)	0.417	0.492	0.075	0.2155
	$A\bar{E}$	Independent Accuracy	0.551	0.468	-0.083	0.0015
	$\bar{A}\bar{E}$	Adoption(AE)	0.532	0.470	-0.062	0.4925
	$\bar{A}\bar{E}$	Independent Accuracy	0.358	0.494	0.136	0.0155
SVERL Bar	$\bar{A}E$	Adoption(AE)	0.518	0.513	-0.006	0.9477
	$\bar{A}E$	Independent Accuracy	0.432	0.447	0.015	0.4127
	$A\bar{E}$	Adoption(AE)	0.800	0.511	-0.289	0.1765
	$A\bar{E}$	Independent Accuracy	0.633	0.441	-0.192	0.1054
	$\bar{A}\bar{E}$	Adoption(AE)	0.460	0.519	0.058	0.4965
	$\bar{A}\bar{E}$	Independent Accuracy	0.362	0.450	0.088	0.0125
SARFA Map	$\bar{A}E$	Adoption(AE)	0.419	0.551	0.131	0.0302
	$\bar{A}E$	Independent Accuracy	0.402	0.433	0.032	0.2312
	$A\bar{E}$	Adoption(AE)	0.417	0.536	0.119	0.1431
	$A\bar{E}$	Independent Accuracy	0.485	0.422	-0.064	0.0345
	$\bar{A}\bar{E}$	Adoption(AE)	0.419	0.543	0.124	0.0485
	$\bar{A}\bar{E}$	Independent Accuracy	0.436	0.426	-0.010	0.9046
AAA Bar	$\bar{A}E$	Adoption(AE)	0.693	0.672	-0.021	0.9795
	$\bar{A}E$	Independent Accuracy	0.451	0.472	0.020	0.7853
	$A\bar{E}$	Adoption(AE)	0.655	0.678	0.022	0.5984
	$A\bar{E}$	Independent Accuracy	0.636	0.440	-0.196	< 0.0001
	$\bar{A}\bar{E}$	Adoption(AE)	0.683	0.674	-0.009	0.6456
	$\bar{A}\bar{E}$	Independent Accuracy	0.371	0.476	0.105	0.0527

We compared excluded and contributing participants on independent accuracy and appropriate adoption using Mann–Whitney tests. The comparisons are reported in Table F.2. Bonferroni correction was applied across 24 comparisons.

For AE trials, all participants contributed at least one non-agreement trial in every explanation condition, so no participants were excluded from the appropriate adoption calculation.

For $A\bar{E}$ trials, excluded participants had higher independent accuracy than contributing participants in three conditions at the uncorrected $\alpha = .05$ threshold: SVERL Map ($\mathrm{Mean}_0 = 0.551$ vs. $\mathrm{Mean}_+ = 0.468$, $p = .002$), SARFA Map ($\mathrm{Mean}_0 = 0.485$ vs. $\mathrm{Mean}_+ = 0.422$, $p = .035$), and AAA Bar ($\mathrm{Mean}_0 = 0.636$ vs. $\mathrm{Mean}_+ = 0.440$, $p < .001$). The difference was not significant for SVERL Bar ($\mathrm{Mean}_0 = 0.633$ vs. $\mathrm{Mean}_+ = 0.441$, $p = .105$). After correction, the differences for SVERL Map and AAA Bar remained significant. Appropriate adoption, calculated from AE trials, did not differ significantly between excluded and contributing participants in any condition (all uncorrected $p > .14$).

For trials with suboptimal recommendations ($\bar{A}E$ and $\bar{A}\bar{E}$), none of the comparisons remained significant after correction. Before correction, contributing participants had higher appropriate adoption than excluded participants under SARFA Map for $\bar{A}E$ trials ($p = .030$) and $\bar{A}\bar{E}$ trials ($p = .049$). Contributing participants also had higher independent accuracy for $\bar{A}\bar{E}$ trials under SVERL Map ($p = .0155$) and SVERL Bar ($p = .0125$).

We compute all derived metrics at the participant level. For each participant and condition, we first calculate the adoption rates needed for each metric from that participant's eligible trials and then form selective reliance, explanation sensitivity, and decision efficiency from those participant-level quantities. The reported means and standard deviations summarise the resulting participant-level metrics across participants. Consequently, the reported mean of a difference or ratio need not equal the difference or ratio formed from separately reported group means, particularly when these adoption rates are based on different sets of participants.

G Outcomes

Results for the outcome measures are provided in Tables G.1 and G.2.

Table G.1: Summary of primary outcomes across all conditions, reported as means (SD) where available. Appropriate adoption is Adoption(AE) in explanation conditions and Adoption(A) in NEB. SR = selective reliance; ES = explanation sensitivity.

Condition	Adoption	Appropriate adoption	Adopt($\bar{A}$)	SR	ES
SVERL Map	0.448 (0.189)	0.473 (0.214)	0.287 (0.288)	0.174 (0.233)	-0.138 (0.424)
SVERL Bar	0.475 (0.201)	0.514 (0.227)	0.414 (0.363)	0.112 (0.217)	0.147 (0.365)
SARFA Map	0.528 (0.212)	0.526 (0.228)	0.639 (0.341)	-0.114 (0.216)	0.159 (0.447)
AAA Bar	0.631 (0.210)	0.674 (0.243)	0.494 (0.361)	0.201 (0.228)	0.074 (0.360)
AAA Bar (S2)	0.610 (0.217)	0.647	0.495	0.346	0.082
NEB	0.337 (0.130)	0.343 (0.162)	0.341	-0.004	—

Table G.2: Trust and Satisfaction scores across all conditions (means, SD).

Condition	Trust	Satisfaction
SVERL Map	0.485 (0.233)	0.509 (0.186)
SVERL Bar	0.510 (0.215)	0.492 (0.242)
SARFA Map	0.520 (0.204)	0.545 (0.193)
AAA Bar	0.578 (0.237)	0.662 (0.189)
AAA Bar (S2)	0.595	0.656
NEB	0.361	—

H Between-participant comparisons with NEB

SARFA Map, SVERL Map, and SVERL Bar in Study 1 were compared with NEB in Study 2 on appropriate adoption, over-reliance, and selective reliance using two-sided Welch independent-samples t-tests. Full results are reported in Table H.1. The uncorrected p-values and Bonferroni-adjusted p_{corr} values are reported separately, with correction across all nine comparisons. Positive t and Hedges' g indicate higher values under the Study 1 explanation condition than under NEB.

Table H.1: Between-participant comparisons of Study 1 explanation conditions with Study 2 NEB (Welch independent-samples t-tests).

Comparison	t	df	p	p_{corr}	Hedges' g
Appropriate adoption					
SVERL Map vs. NEB	4.52	166.96	$< .001$	$< .001$	0.67
SVERL Bar vs. NEB	5.75	167.92	$< .001$	$< .001$	0.84
SARFA Map vs. NEB	6.13	167.95	$< .001$	$< .001$	0.90
Over-reliance					
SVERL Map vs. NEB	-1.30	157.12	.194	1.000	-0.20
SVERL Bar vs. NEB	1.53	162.91	.127	1.000	0.23
SARFA Map vs. NEB	6.49	161.88	$< .001$	$< .001$	0.97
Selective reliance					
SVERL Map vs. NEB	3.81	162.07	$< .001$	.002	0.57
SVERL Bar vs. NEB	1.92	162.99	.057	.510	0.28
SARFA Map vs. NEB	-2.55	159.76	.012	.105	-0.39

I Nonparametric tests

This section reports nonparametric pairwise comparisons for the key outcome variables in Study 1. Study 2 trust-subscale comparisons are reported separately below. Study 1 comparisons use the Wilcoxon signed-rank test with Bonferroni correction across the six condition pairs. Statistical significance is determined using Bonferroni-adjusted p-values, with $p_{\text{corr}} < 0.05$, correcting for the six pairwise comparisons. Effect size r is reported as $Z/\sqrt{N}$. In the tables below, p is uncorrected and p_{corr} is Bonferroni-adjusted. Significance symbols refer to p_{corr}: * $p_{\text{corr}} < .05$, ** $p_{\text{corr}} < .01$, and *** $p_{\text{corr}} < .001$. The symbol † denotes an uncorrected $p < .05$ that does not remain significant after correction; ns denotes a nonsignificant comparison.

I.1 Appropriate adoption rate

Table I.1: Post-hoc pairwise comparisons of appropriate adoption rate (Wilcoxon signed-rank test).

Comparison	W	p	p_{corr}	Sig.	r
SVERL Map vs. SVERL Bar	2002.0	0.0988	0.5925	ns	0.166
SVERL Map vs. SARFA Map	1245.5	<0.0001	0.0002	***	0.422
SVERL Map vs. AAA Bar	589.5	<0.0001	<0.0001	***	0.657
SVERL Bar vs. SARFA Map	1694.0	0.0141	0.0843	†	0.249
SVERL Bar vs. AAA Bar	750.0	<0.0001	<0.0001	***	0.610
SARFA Map vs. AAA Bar	1121.0	<0.0001	<0.0001	***	0.459

I.2 Overall trust

Table I.2: Post-hoc pairwise comparisons of overall trust (Wilcoxon signed-rank test).

Comparison	W	p	p_{corr}	Sig.	r
SVERL Map vs. SVERL Bar	1585.0	0.0214	0.1282	†	0.239
SVERL Map vs. SARFA Map	1185.5	0.0003	0.0020	**	0.377
SVERL Map vs. AAA Bar	337.0	<0.0001	<0.0001	***	0.726
SVERL Bar vs. SARFA Map	1852.0	0.3397	1.0000	ns	0.100
SVERL Bar vs. AAA Bar	736.0	<0.0001	<0.0001	***	0.563
SARFA Map vs. AAA Bar	844.5	<0.0001	<0.0001	***	0.518

I.3 Overall explanation satisfaction

Table I.3: Post-hoc pairwise comparisons of overall explanation satisfaction (Wilcoxon signed-rank test).

Comparison	W	p	p_{corr}	Sig.	r
SVERL Map vs. SVERL Bar	2016.5	0.4152	1.0000	ns	0.084
SVERL Map vs. SARFA Map	1586.0	0.0312	0.1873	†	0.225
SVERL Map vs. AAA Bar	740.0	<0.0001	<0.0001	***	0.592
SVERL Bar vs. SARFA Map	1583.0	0.0303	0.1818	†	0.226
SVERL Bar vs. AAA Bar	556.5	<0.0001	<0.0001	***	0.661
SARFA Map vs. AAA Bar	770.0	<0.0001	<0.0001	***	0.563

I.4 Study 2: Trust dimensions

Descriptive statistics and significance results for the four trust subscales (benevolence, competence, perceived risk, and reciprocity), comparing AAA Bar with NEB in Study 2, are provided in Table I.4. All comparisons are within-participant Wilcoxon signed-rank tests, with Bonferroni correction across the four subscales ($\alpha = 0.05/4 = 0.0125$).

Table I.4: Study 2: Trust dimension comparisons between AAA Bar and NEB (Wilcoxon signed-rank).

	AAA Bar		NEB		
Dimension	M	SD	M	SD	p
Benevolence	0.621	0.188	0.444	0.178	$< .001$
Competence	0.699	0.169	0.415	0.187	$< .001$
Perceived Risk	0.462	0.247	0.670	0.197	$< .001$
Reciprocity	0.520	0.254	0.257	0.167	$< .001$

Note. M = mean; SD = standard deviation; p denotes the uncorrected Wilcoxon p-value. All four comparisons had $p < .001$ and remained significant at the Bonferroni-corrected threshold. AAA Bar elicited significantly higher benevolence, competence, and reciprocity, and significantly lower perceived risk, relative to the no-explanation condition.

J Mixed-effects models

This appendix provides full details of the linear mixed-effects models referenced in Section 6. All models were fitted using restricted maximum likelihood (REML) with a random intercept per participant (Group). The dependent variable is the appropriate adoption rate (Adoption(AE) in explanation conditions and Adoption(A) in NEB) measured as the proportion of trials on which a participant switched from an initial disagreement to adopt the agent's recommendation—computed after excluding trials where the initial choice already matched the recommendation, as defined in Section 5. Categorical predictors are dummy-coded with AAA Bar as the reference level for both Study 1 and Study 2.

J.1 Study 1

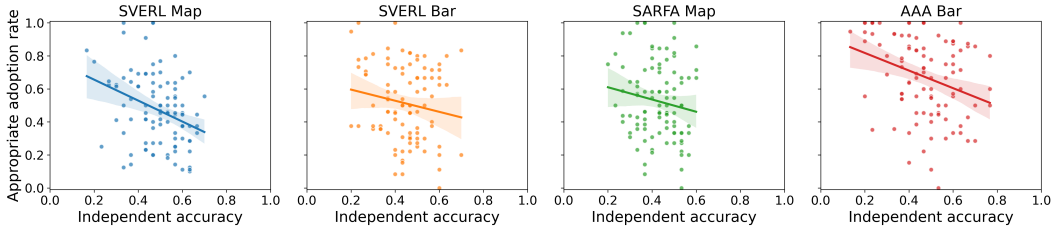


Figure 6: Relationship between independent accuracy and appropriate adoption, shown separately for each of the four explanation conditions. Each subplot displays individual participant data points with the corresponding linear regression fit and 95% confidence interval. Appropriate adoption tends to decrease slightly as independent accuracy increases, but adoption rates vary widely among participants with similar independent accuracy.

The full covariate model adjusts for participant characteristics.

To reduce multicollinearity, responses to the AI, explainable AI, reinforcement learning, and explainable reinforcement learning literacy items were combined into a single *Technology Literacy* factor via principal component analysis (PCA). The first component explained 81.3% of the variance; loadings were AI = 0.492, explainable = 0.495, reinforcement learning = 0.511, explainable reinforcement learning = 0.502. The resulting factor scores had a mean of 0 and an SD of 1.813.

The model includes Explanation Condition along with Technology Literacy, Independent Accuracy (proportion of initial actions matching the optimal action), Gaming Experience, Reinforcement Learning Understanding (proportion correct on the final reinforcement learning comprehension check), AI Attitude, and Perceived Task Competence.

Table J.1: Study 1: full covariate model predicting appropriate adoption.

Predictor	Coef.	Std.Err.	z	$P > \|z\|$	95% CI
Intercept (AAA Bar)	1.124	0.135	8.325	<0.001	[0.859, 1.388]
SARFA Map	-0.149	0.025	-6.019	<0.001	[-0.197, -0.100]
SVERL Bar	-0.160	0.025	-6.500	<0.001	[-0.209, -0.112]
SVERL Map	-0.201	0.025	-8.158	<0.001	[-0.250, -0.153]
Technology Literacy	0.025	0.011	2.235	0.025	[0.003, 0.047]
Independent Accuracy	-0.721	0.192	-3.758	<0.001	[-1.096, -0.345]
Gaming Experience	0.025	0.067	0.376	0.707	[-0.106, 0.156]
RL Understanding	-0.029	0.042	-0.678	0.498	[-0.112, 0.054]
AI Attitude	0.025	0.112	0.220	0.826	[-0.195, 0.245]
Perceived Competence	-0.189	0.113	-1.671	0.095	[-0.412, 0.033]

Random effects: `Group Var = 0.016, Scale = 0.0305`
Model fit: `Log-Likelihood = 56.775, No. Obs = 400 (100 groups)`

Mechanism effects are robust after controlling for all covariates. Higher independent accuracy predicts lower appropriate adoption, consistent with the pattern reported in the main text.

The explanation-understanding interaction model tests whether differences in adoption across explanation conditions are moderated by Explanation Understanding (the proportion of correct responses on the comprehension check administered after each condition).

Table J.2: Study 1: interaction between explanation condition and explanation understanding.

Predictor	Coef.	Std.Err.	z	$P > \|z\|$	95% CI
Intercept (AAA Bar)	0.626	0.031	20.102	<0.001	[0.565, 0.687]
SARFA Map	-0.102	0.033	-3.096	0.002	[-0.167, -0.038]
SVERL Bar	-0.116	0.040	-2.883	0.004	[-0.195, -0.037]
SVERL Map	-0.098	0.042	-2.360	0.018	[-0.179, -0.017]
Explanation Understanding	0.088	0.039	2.269	0.023	[0.012, 0.164]
SARFA Map × Expl. Understanding	-0.074	0.065	-1.140	0.254	[-0.200, 0.053]
SVERL Bar × Expl. Understanding	-0.081	0.054	-1.502	0.133	[-0.187, 0.025]
SVERL Map × Expl. Understanding	-0.170	0.054	-3.121	0.002	[-0.277, -0.063]

Random effects: `Group Var = 0.022, Scale = 0.0298`
Model fit: `Log-Likelihood = 49.677, No. Obs = 400 (100 groups)`

The significant negative interaction for SVERL Map indicates that the association between explanation understanding and appropriate adoption was less positive under SVERL Map than under the AAA Bar reference condition.

Mechanism and presentation format were examined in the same model. For this analysis, the four conditions were recoded using two binary dummy variables: `Method_Saliency` (1 for SARFA) and `Method_Shapley` (1 for SVERL), with AAA as the baseline method. A further dummy `Visualisation_Map` (1 for SARFA Map and SVERL Map, 0 for SVERL Bar and AAA Bar) captures the effect of spatial heatmap versus bar chart.

Table J.3: Study 1: recoding of explanation conditions by method family and visual format.

Predictor	Coef.	Std.Err.	z	$P > \|z\|$	95% CI
Intercept (AAA, bar format)	0.674	0.023	29.565	<0.001	[0.630, 0.719]
Method_Saliency	-0.108	0.035	-3.084	0.002	[-0.176, -0.039]
Method_Shapley	-0.160	0.025	-6.500	<0.001	[-0.209, -0.112]
Visualisation_Map	-0.041	0.025	-1.658	0.097	[-0.089, 0.007]

Random effects: `Group Var = 0.022, Scale = 0.0305`
Model fit: `Log-Likelihood = 53.676, No. Obs = 400 (100 groups)`

The coefficient for `Visualisation_Map` was not statistically significant ($p = .097$), providing no evidence of a map-versus-bar difference in this recoding.

Big Five personality traits were tested as moderators, each in a separate model with Explanation Condition, the trait, and their interaction.

Table J.4: Study 1: appropriate adoption predicted by explanation condition and extraversion.

Predictor	Coef.	Std.Err.	z	$P > \|z\|$	95% CI
Intercept (AAA Bar)	0.717	0.054	13.219	<0.001	[0.610, 0.823]
SARFA Map	-0.257	0.058	-4.395	<0.001	[-0.371, -0.142]
SVERL Bar	-0.208	0.058	-3.559	<0.001	[-0.323, -0.093]
SVERL Map	-0.276	0.058	-4.727	<0.001	[-0.391, -0.162]
Extraversion	-0.074	0.086	-0.861	0.390	[-0.243, 0.095]
SARFA Map × Extraversion	0.190	0.093	2.043	0.041	[0.008, 0.372]
SVERL Bar × Extraversion	0.084	0.093	0.898	0.369	[-0.099, 0.266]
SVERL Map × Extraversion	0.131	0.093	1.414	0.157	[-0.051, 0.314]

Random effects: `Group Var = 0.022, Scale = 0.0303`
Model fit: `Log-Likelihood = 49.464, No. Obs = 400`

Table J.5: Study 1: appropriate adoption predicted by explanation condition and neuroticism.

Predictor	Coef.	Std.Err.	z	$P > \|z\|$	95% CI
Intercept (AAA Bar)	0.637	0.062	10.350	<0.001	[0.517, 0.758]
SARFA Map	-0.138	0.067	-2.073	0.038	[-0.269, -0.008]
SVERL Bar	-0.152	0.067	-2.278	0.023	[-0.283, -0.021]
SVERL Map	-0.173	0.067	-2.596	0.009	[-0.304, -0.042]
Neuroticism	0.061	0.094	0.648	0.517	[-0.123, 0.244]
SARFA Map × Neuroticism	-0.017	0.101	-0.167	0.867	[-0.215, 0.182]
SVERL Bar × Neuroticism	-0.014	0.101	-0.138	0.890	[-0.212, 0.185]
SVERL Map × Neuroticism	-0.046	0.101	-0.457	0.648	[-0.245, 0.152]

Random effects: `Group Var = 0.022, Scale = 0.0308`
Model fit: `Log-Likelihood = 47.753, No. Obs = 400`

Table J.6: Study 1: appropriate adoption predicted by explanation condition and openness.

Predictor	Coef.	Std.Err.	z	$P > \|z\|$	95% CI
Intercept (AAA Bar)	0.478	0.069	6.958	<0.001	[0.344, 0.613]
SARFA Map	-0.013	0.076	-0.168	0.867	[-0.161, 0.136]
SVERL Bar	-0.101	0.076	-1.332	0.183	[-0.249, 0.048]
SVERL Map	-0.146	0.076	-1.932	0.053	[-0.295, 0.002]
Openness	0.322	0.107	3.012	0.003	[0.113, 0.532]
SARFA Map × Openness	-0.224	0.118	-1.898	0.058	[-0.455, 0.007]
SVERL Bar × Openness	-0.098	0.118	-0.831	0.406	[-0.329, 0.133]
SVERL Map × Openness	-0.091	0.118	-0.769	0.442	[-0.322, 0.140]

Random effects: `Group Var = 0.020, Scale = 0.0304`
Model fit: `Log-Likelihood = 53.620, No. Obs = 400`

Table J.7: Study 1: appropriate adoption predicted by explanation condition and conscientiousness.

Predictor	Coef.	Std.Err.	z	$P > \|z\|$	95% CI
Intercept (AAA Bar)	0.760	0.086	8.802	<0.001	[0.591, 0.929]
SARFA Map	-0.148	0.093	-1.583	0.113	[-0.330, 0.035]
SVERL Bar	-0.275	0.093	-2.951	0.003	[-0.458, -0.092]
SVERL Map	-0.296	0.093	-3.173	0.002	[-0.479, -0.113]
Conscientiousness	-0.125	0.122	-1.026	0.305	[-0.364, 0.114]
SARFA Map × Conscientiousness	-0.001	0.132	-0.011	0.991	[-0.260, 0.257]
SVERL Bar × Conscientiousness	0.168	0.132	1.276	0.202	[-0.090, 0.426]
SVERL Map × Conscientiousness	0.139	0.132	1.051	0.293	[-0.120, 0.397]

Random effects: `Group Var = 0.022, Scale = 0.0305`
Model fit: `Log-Likelihood = 50.067, No. Obs = 400`

Table J.8: Study 1: appropriate adoption predicted by explanation condition and agreeableness.

Predictor	Coef.	Std.Err.	z	$P > \|z\|$	95% CI
Intercept (AAA Bar)	0.765	0.112	6.821	<0.001	[0.545, 0.985]
SARFA Map	-0.374	0.121	-3.091	0.002	[-0.611, -0.137]
SVERL Bar	-0.331	0.121	-2.733	0.006	[-0.568, -0.094]
SVERL Map	-0.406	0.121	-3.353	0.001	[-0.643, -0.169]
Agreeableness	-0.122	0.148	-0.824	0.410	[-0.413, 0.169]
SARFA Map × Agreeableness	0.305	0.160	1.903	0.057	[-0.009, 0.618]
SVERL Bar × Agreeableness	0.230	0.160	1.437	0.151	[-0.084, 0.544]
SVERL Map × Agreeableness	0.276	0.160	1.725	0.085	[-0.038, 0.590]

Random effects: `Group Var = 0.022, Scale = 0.0303`
Model fit: `Log-Likelihood = 51.822, No. Obs = 400`

J.2 Study 2

Study 2 compared AAA Bar with a no-explanation baseline (NEB) using an otherwise identical within-participant design. The models below include Explanation Condition (AAA Bar = reference) and, where interaction effects are examined, the Big Five trait.

The full covariate model adjusts the AAA Bar–NEB comparison for participant characteristics.

Table J.9: Study 2: full covariate model predicting appropriate adoption.

Predictor	Coef.	Std.Err.	z	$P > \|z\|$	95% CI
Intercept (AAA Bar)	0.676	0.161	4.201	<0.001	[0.361, 0.991]
NEB	-0.304	0.034	-8.850	<0.001	[-0.372, -0.237]
Technology Literacy	0.040	0.014	2.903	0.004	[0.013, 0.067]
Independent Accuracy	-0.146	0.242	-0.606	0.545	[-0.620, 0.327]
Gaming Experience	-0.222	0.089	-2.499	0.012	[-0.396, -0.048]
RL Understanding	0.081	0.052	1.566	0.117	[-0.020, 0.182]
AI Attitude	0.179	0.180	0.996	0.319	[-0.173, 0.532]

Random effects: `Group Var = 0.009, Scale = 0.0414`
Model fit: `Log-Likelihood = 1.6142, No. Obs = 140 (70 groups)`
Convergence: `Yes`

The significant negative coefficient for NEB shows that appropriate adoption was lower without an explanation after controlling for the covariates listed in Table J.9.

Big Five personality traits were tested as moderators, as in Study 1, each in a separate model with Explanation Condition, the trait, and their interaction.

Table J.10: Study 2: appropriate adoption predicted by condition and extraversion.

Predictor	Coef.	Std.Err.	z	$P > \|z\|$	95% CI
Intercept (AAA Bar)	0.975	0.068	14.410	<0.001	[0.843, 1.108]
NEB	-0.531	0.086	-6.203	<0.001	[-0.698, -0.363]
Extraversion	-0.577	0.110	-5.252	<0.001	[-0.792, -0.361]
NEB × Extraversion	0.397	0.139	2.862	0.004	[0.125, 0.669]

Random effects: `Group Var = 0.009, Scale = 0.0375`
Model fit: `Log-Likelihood = 10.811, No. Obs = 140`

Table J.11: Study 2: appropriate adoption predicted by condition and neuroticism.

Predictor	Coef.	Std.Err.	z	$P > \|z\|$	95% CI
Intercept (AAA Bar)	0.487	0.080	6.093	<0.001	[0.330, 0.644]
NEB	-0.216	0.098	-2.206	0.027	[-0.408, -0.024]
Neuroticism	0.263	0.123	2.139	0.032	[0.022, 0.504]
NEB × Neuroticism	-0.145	0.151	-0.962	0.336	[-0.440, 0.150]

Random effects: `Group Var = 0.014, Scale = 0.0415`
Model fit: `Log-Likelihood = 0.671, No. Obs = 140`

Table J.12: Study 2: appropriate adoption predicted by condition and openness.

Predictor	Coef.	Std.Err.	z	$P > \|z\|$	95% CI
Intercept (AAA Bar)	0.194	0.056	3.441	0.001	[0.083, 0.304]
NEB	0.060	0.071	0.842	0.400	[-0.079, 0.199]
Openness	0.755	0.086	8.799	<0.001	[0.587, 0.924]
NEB × Openness	-0.607	0.108	-5.600	<0.001	[-0.820, -0.395]

Random effects: `Group Var = 0.007, Scale = 0.0288`
Model fit: `Log-Likelihood = 28.531, No. Obs = 140`

Table J.13: Study 2: appropriate adoption predicted by condition and conscientiousness.

Predictor	Coef.	Std.Err.	z	$P > \|z\|$	95% CI
Intercept (AAA Bar)	0.207	0.132	1.571	0.116	[-0.051, 0.465]
NEB	0.053	0.160	0.331	0.741	[-0.261, 0.367]
Conscientiousness	0.642	0.188	3.413	0.001	[0.273, 1.010]
NEB × Conscientiousness	-0.521	0.229	-2.279	0.023	[-0.970, -0.073]

Random effects: `Group Var = 0.014, Scale = 0.0390`
Model fit: `Log-Likelihood = 4.853, No. Obs = 140`

Table J.14: Study 2: appropriate adoption predicted by condition and agreeableness.

Predictor	Coef.	Std.Err.	z	$P > \|z\|$	95% CI
Intercept (AAA Bar)	0.814	0.081	9.995	<0.001	[0.655, 0.974]
NEB	-0.419	0.099	-4.215	<0.001	[-0.614, -0.224]
Agreeableness	-0.240	0.110	-2.190	0.029	[-0.456, -0.025]
NEB × Agreeableness	0.165	0.134	1.228	0.219	[-0.098, 0.427]

Random effects: `Group Var = 0.014, Scale = 0.0411`
Model fit: `Log-Likelihood = 0.476, No. Obs = 140`

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's contributions—including the comparative findings for three explanation mechanisms, the novel experimental framework, and the introduction of AAA—and accurately summarise the key results and scope of the two user studies without overclaiming.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated "Limitations and future work" paragraph in the Discussion that explicitly addresses the simplicity of the task environment, the demographic homogeneity of the participant sample, and a technical limitation of the AAA method in sparse-reward domains.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is an empirical user study evaluating explanation mechanisms through controlled behavioural experiments and does not contain formal theoretical results, theorems, or proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides detailed specifications of the task environment, agent training, explanation mechanisms, experimental design, metrics, and statistical models; all essential information to replicate the behavioural study is included.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The supplementary material includes code, data, and instructions to reproduce the main experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides reinforcement learning agent training hyperparameters, and environment details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports statistical significance tests, p-values, and confidence intervals for the mixed-effects models supporting the main claims.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies compute resources for training the reinforcement learning agent.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper states that ethics approval was obtained, participants provided informed consent, were debriefed on key manipulations, and an exclusion criterion (color-vision deficiency) was applied, consistent with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper's introduction and discussion explicitly address the positive societal impact of enabling calibrated human–AI collaboration in safety-critical decision-support settings, and the negative impact of mechanism-induced miscalibration leading to over-reliance and potential downstream safety failures.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release any pre-trained language models, image generators, or scraped datasets; the model is a frozen reinforcement learning policy for a simple grid navigation task that poses no high risk for misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper does not use existing datasets, pre-trained models; it re-implements the explanation methods and uses custom-created environments. Original method sources are cited appropriately. The experiments were implemented in PsychoPy v2024.2.4, distributed under the GNU General Public License v3 (GPLv3). PsychoPy is cited in the paper, and its license is available at https://github.com/psychopy/psychopy/blob/2024.2.4/LICENSE.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper fully describes the AAA method (including the decomposition equation in Section 5 and detailed construction in Appendix B.3), and the accompanying supplementary material zip file contains anonymised code, data, and instructions that serve as documentation for the new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: The paper includes screenshots of the task interface and participant instructions in the supplementary materials, and explicitly states that compensation was provided in the form of course credit via the university's Research Participation Scheme.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The paper explicitly states that the study involved no more than minimal risk (a low-stakes computer task) and that this was disclosed to participants during the consent process, alongside the obtained ethics approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The paper's core methodology involves reinforcement learning, explanation mechanisms, and user studies, and does not use LLMs as an important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.