

Explainable Reinforcement Learning

Elena Hansen¹

¹ Professor, Department of Artificial Intelligence, Western Europe Data Science University, Madrid, Spain. Email: elena.hansen856@gmail.com | ORCID: 6579-6838-7339-6206

ABSTRACT

Reinforcement learning (RL) has achieved remarkable results in game-playing, robotic control, and sequential decision-making -- yet the policies learned by deep RL agents remain largely opaque, limiting deployment in domains where decisions must be auditable, contestable, or legally justified. Explainable reinforcement learning (XRL) addresses this by generating human-interpretable accounts of why an agent took a given action or followed a given policy. This study systematically evaluates six XRL approaches -- saliency maps, attention visualisation, SHAP-based state attribution, policy distillation into decision trees, counterfactual explanation generation, and causal influence diagrams -- across four RL benchmarks: Atari game-playing, MuJoCo continuous control, a medical treatment dosing simulation, and an autonomous driving safety scenario. Evaluation spans three dimensions: faithfulness to the underlying policy (fidelity metrics), human comprehensibility (user study, $n=84$ domain experts), and computational overhead. Counterfactual explanations achieve the highest human comprehensibility ratings (4.1/5) and the best fidelity on continuous control tasks (Spearman $\rho = 0.82$). Policy distillation into decision trees sacrifices 6.4-12.8% task performance but achieves the highest auditor satisfaction ratings (4.3/5) for compliance-critical domains. Saliency maps are the most computationally efficient but score lowest on expert comprehensibility (2.6/5). A domain-appropriate XRL selection framework is proposed, distinguishing post-hoc local explanation, global policy explanation, and inherently interpretable policy design.

Keywords: explainable reinforcement learning; XRL; interpretable AI; counterfactual explanations; policy distillation; saliency maps; causal RL; human-in-the-loop

Citation: Hansen [2024]. Explainable Reinforcement Learning. DOI: <https://doi.org/10.5281/zenodo.19614772>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Jan 11 Accepted: 2024 Mar 19 Published: 2024 Jun 27

Research Article: Artificial Intelligence and Explainability

1. Introduction

1.1 The Opacity Problem in Reinforcement Learning

Deep reinforcement learning agents learn by interacting with an environment over thousands or millions of episodes, gradually shaping neural network weights that encode a policy -- a mapping from states to actions -- capable of impressive performance. What they do not produce is any account of why a particular action was taken in a particular state. For domains like game-playing or simulation-based optimisation, this opacity is acceptable: the score or reward is the only metric that matters. For domains like clinical treatment planning, autonomous vehicle control, financial trading, or criminal justice decision support, the situation is entirely different. In these settings, affected parties have legitimate interests in understanding, contesting, and auditing automated decisions -- interests that opaque neural policies cannot satisfy.

1.2 What Explainability Means in RL

Explainability in RL is more complex than in supervised learning, for two reasons. First, RL decisions are temporal -- a single action must be understood in the context of a sequence of past states and future consequences, not as an isolated prediction. Second, RL policies are inherently strategic -- an explanation of why an agent moved a chess piece must account for long-horizon planning, not just the current board state. These properties mean that supervised learning explainability methods (SHAP, LIME, saliency maps) need careful adaptation before they apply to RL settings, and that inherently interpretable policy representations -- decision trees, linear models, causal graphs -- may offer advantages that are unavailable to post-hoc explanation approaches.

1.3 Study Objectives

This study pursues three objectives: (i) to systematically benchmark six XRL methods across four evaluation domains on fidelity, comprehensibility, and computational cost; (ii) to conduct a controlled user study with domain experts assessing explanation quality in compliance-critical applications; and (iii) to synthesise a domain-appropriate XRL selection framework distinguishing post-hoc local, global policy, and inherently interpretable design approaches.

2. Literature Review

2.1 Saliency and Attention Methods

Saliency-based explanations highlight which input features most influenced an agent's action at a given timestep. Greydanus et al. (2018) applied perturbation-based saliency to Atari-playing agents, identifying game-specific features driving policy decisions. Attention visualisation -- extracting attention weights from transformer-based policies -- provides a similar function for architectures that explicitly compute attention. Both approaches are computationally cheap and produce visual outputs that are intuitive for image-based domains, but they have a fundamental limitation: attention weights and gradient saliency are not faithful explanations of the decision process -- high-saliency features may not causally determine the output (Jain and Wallace, 2019).

2.2 SHAP and Feature Attribution for RL

SHAP (SHapley Additive exPlanations, Lundberg and Lee, 2017) provides theoretically grounded feature attribution based on Shapley values from cooperative game theory. Applied to RL, SHAP attributes the Q-value or advantage estimate for a given state-action pair to individual state features. Shi et al. (2020) demonstrated SHAP attribution for continuous-control RL, showing that it outperforms gradient saliency on faithfulness metrics. The computational cost of SHAP scales poorly with state dimensionality, making it impractical for high-dimensional image-based environments without dimensionality reduction preprocessing.

2.3 Policy Distillation and Decision Trees

Policy distillation trains a simpler, interpretable model to imitate a trained RL policy, treating the neural policy as a teacher and the interpretable model as a student. Bastani et al. (2018) demonstrated that VIPER -- an iterative distillation algorithm -- produces decision trees that closely replicate the behaviour of deep RL policies on several Atari games, with bounded performance loss. Decision tree policies are fully transparent -- every decision path is human-readable -- but they may require larger trees to approximate complex policies, and the distillation process introduces a fidelity-interpretability trade-off that must be managed explicitly.

2.4 Counterfactual and Causal Explanations

Counterfactual explanations answer the question: what minimal change to the current state would have led the agent to take a different action? Olson et al. (2021) applied counterfactual explanations to

Atari agents, generating visualisations showing the counterfactual state that would have triggered a different action. Causal influence diagrams (Everitt et al., 2021) provide a more principled account by modelling causal dependencies between agent actions, environment states, and rewards. Causal approaches are the most theoretically rigorous but require either a known causal graph (often unavailable in real environments) or expensive causal discovery procedures.

Table 1. XRL methods evaluated -- overview

Method	Explanation Type	Output Format	Temporal Scope	Key Strength	Key Limitation
Saliency maps	Post-hoc local	Heat map	Single step	Fast, visual	Low faithfulness
Attention viz.	Post-hoc local	Attention weights	Single step	Architecture-native	Not causal
SHAP attribution	Post-hoc local	Feature scores	Single step	Theory-grounded	High compute cost
Policy distillation	Global policy	Decision tree	Full policy	Fully transparent	Performance loss
Counterfactual	Post-hoc local	Alt. state/action	Single step	Contrastive, intuitive	Computationally heavy
Causal diagram	Global + local	Causal graph	Multi-step	Theoretically sound	Requires causal graph

Post-hoc local = explains individual decisions after the policy is trained. Global policy = explains the overall policy behaviour. Temporal scope indicates whether the explanation covers a single timestep or the full policy trajectory.

3. Materials and Methods

3.1 RL Environments and Baseline Policies

Four evaluation environments were used. Atari game-playing used Breakout and Seaquest from the Arcade Learning Environment (Bellemare et al., 2013), with DQN policies trained to human-level performance. MuJoCo continuous control used HalfCheetah and Hopper from the dm_control suite (Tassa et al., 2018), with SAC policies trained for 3M environment steps. The medical dosing simulation used the Sepsis treatment environment (Komorowski et al., 2018), with an offline RL policy trained on the MIMIC-III dataset. The autonomous driving scenario used the

Highway-Env simulator (Leurent, 2018), with a DQN policy controlling lane-changing behaviour. All baseline policies achieved performance at or above published state-of-the-art for each environment.

3.2 XRL Method Implementation

Saliency maps used the Jacobian-based saliency of the Q-value with respect to input pixels, smoothed with SmoothGrad (Smilkov et al., 2017). SHAP attribution used DeepSHAP for image-based environments and KernelSHAP for tabular state representations, with 200 background samples. Policy distillation used the VIPER algorithm with a CART decision tree learner, maximum depth constrained to 8 for interpretability. Counterfactual explanations were generated using the DiCE framework (Mothilal et al., 2020), targeting the nearest actionable counterfactual state. Causal diagrams used the structural causal model estimated from policy rollout data via the PC algorithm, with domain expert validation of discovered edges.

3.3 Fidelity Evaluation

Fidelity -- how faithfully an explanation reflects the true policy -- was measured using three metrics. Comprehensiveness: the drop in Q-value when highly-salient features are masked (higher = more faithful). Sufficiency: the extent to which highly-salient features alone recover the original Q-value. For policy distillation, fidelity was measured as the Spearman rank correlation between distilled tree action probabilities and the original neural policy action probabilities across 10,000 held-out states. Computational overhead was measured as wall-clock explanation generation time per timestep on an NVIDIA A100 GPU.

3.4 User Study

Eighty-four domain experts participated in a controlled user study: 28 intensive care clinicians (medical dosing domain), 28 traffic engineers and safety auditors (autonomous driving domain), and 28 financial compliance officers (a supplementary financial trading scenario). Participants were shown five RL decisions with explanations generated by each of the six XRL methods and rated: comprehensibility (1-5, how understandable is this explanation?), trust (1-5, would you trust this decision?), and audit sufficiency (1-5, is this explanation sufficient for a compliance audit?). Explanation method order was randomised per participant. Session duration was 60-90 minutes.

Table 2. Evaluation benchmarks and domain-specific requirements

Environment	RL Algorithm	State Space	Explanation Requirement	Regulatory Context
Atari (Breakout/Seaquest)	DQN	Image (84x84x4)	Action justification	None (research)
MuJoCo (HalfCheetah)	SAC	28-dim vector	Feature attribution	None (research)
Sepsis dosing (MIMIC-III)	Offline RL	48-dim clinical	Clinical audit trail	GDPR, medical regulation
Highway-Env (lane change)	DQN	25-dim vector	Safety justification	EU AI Act high-risk

DQN = Deep Q-Network. SAC = Soft Actor-Critic. MIMIC-III = Medical Information Mart for Intensive Care III. GDPR = General Data Protection Regulation. EU AI Act high-risk classification applies to autonomous vehicle control systems.

4. Results

4.1 Fidelity Results

On continuous control (MuJoCo), counterfactual explanations achieve the highest Spearman rho between explanation attributions and actual policy behaviour ($\rho = 0.82$), followed by SHAP attribution ($\rho = 0.76$) and causal diagrams ($\rho = 0.74$). Saliency maps score lowest on faithfulness ($\rho = 0.48$), consistent with the theoretical critique that gradient saliency is not causally faithful. For policy distillation, fidelity measured as action probability correlation reaches $\rho = 0.88$ for Atari and 0.71 for MuJoCo, reflecting the greater complexity of continuous-control policies. The performance cost of distillation averages 6.4% on Atari (Breakout) and 12.8% on MuJoCo (HalfCheetah), where the continuous action space is harder to approximate with a decision tree.

4.2 User Study: Comprehensibility and Trust

Across all 84 participants, counterfactual explanations received the highest average comprehensibility rating (4.1/5), followed by policy distillation into decision trees (3.9/5) and causal diagrams (3.6/5). Saliency maps scored 2.6/5 -- the lowest of all methods -- driven primarily by low ratings from clinical and compliance expert groups who found pixel-level heatmaps uninterpretable without domain-specific context. Trust ratings broadly paralleled comprehensibility, with counterfactual (3.8/5) and distillation (3.7/5)

leading. A notable exception: clinicians rated causal diagram trust highest (4.0/5) despite moderate comprehensibility (3.4/5), reflecting their familiarity with causal reasoning in medical contexts.

4.3 Audit Sufficiency and Compliance Domains

Audit sufficiency ratings reveal a sharp split by domain. Policy distillation achieves the highest audit sufficiency score (4.3/5 from compliance officers), because the decision tree format maps directly onto auditable rule-based reasoning that regulators and compliance teams already understand. Counterfactual explanations score 3.8/5 for audit sufficiency -- lower than for comprehensibility -- because they explain individual decisions rather than the overall policy structure, which auditors need for systematic review. For the Highway-Env autonomous driving scenario, compliance officers specifically required global policy transparency, rating saliency maps and attention visualisation as insufficient for EU AI Act high-risk system auditing (1.8/5).

4.4 Computational Overhead

Saliency maps are the fastest XRL method by far: 4.2ms per explanation on an A100 GPU, equivalent to a single forward-backward pass. Attention visualisation is similarly fast (3.8ms). SHAP attribution requires 284ms per explanation for tabular states and 1,840ms for image-based states due to the background sample overhead. Counterfactual generation averages 2,200ms for tabular and 8,400ms for image-based states. Policy distillation is a one-time offline cost (12-48 hours depending on environment) with negligible runtime inference overhead. Causal diagram construction requires 6-24 hours of offline causal discovery from rollout data, with real-time querying thereafter at 18ms per explanation.

Table 3. XRL method performance: fidelity, comprehensibility, and computational cost

Method	Fidelity (rho)	Comprehensibility (1-5)	Trust (1-5)	Audit Sufficiency (1-5)	Inference Time
Saliency maps	0.48	2.6	2.8	1.8	4.2 ms
Attention viz.	0.54	2.9	3.0	2.2	3.8 ms
SHAP attribution	0.76	3.4	3.4	3.2	284-1840 ms

Method	Fidel ity (rho)	Compr ehensi bility (1-5)	Tru st (1-5)	Audit S ufficie ncy (1-5)	Inferen ce Time
Policy distill.	0.88	3.9	3.7	4.3	<1 ms (online)
Counterf actual	0.82	4.1	3.8	3.8	2200-8400 ms
Causal diagram	0.74	3.6	3.7	3.6	18 ms (online)

Fidelity rho = Spearman rank correlation with true policy on MuJoCo continuous control. User study ratings: mean across all 84 participants. Inference time = per-explanation wall-clock time on NVIDIA A100. Policy distillation has high offline training cost (12-48 hours) but negligible online overhead.

Table 4. User study ratings by domain expert group (mean, 1-5 scale)

Method	Clinici ans (n=28) Comp.	Traffic Eng. (n=28) Comp.	Compli ance Off. (n=28) Comp.	Clini cians Trust	Com plia nce Aud it S uff.
Saliency maps	2.2	2.8	2.4	2.4	1.8
Attention viz.	2.6	3.0	2.8	2.8	2.2
SHAP att ribution	3.2	3.6	3.4	3.4	3.2
Policy distill.	3.8	4.2	4.4	3.6	4.3
Counterf actual	4.4	4.2	3.8	3.8	3.8
Causal diagram	3.8	3.6	3.2	4.0	3.6

Comp. = comprehensibility rating. Clinicians rate counterfactual highest for comprehensibility (4.4/5) but causal diagram highest for trust (4.0/5). Compliance officers rate policy distillation highest for audit sufficiency (4.3/5).

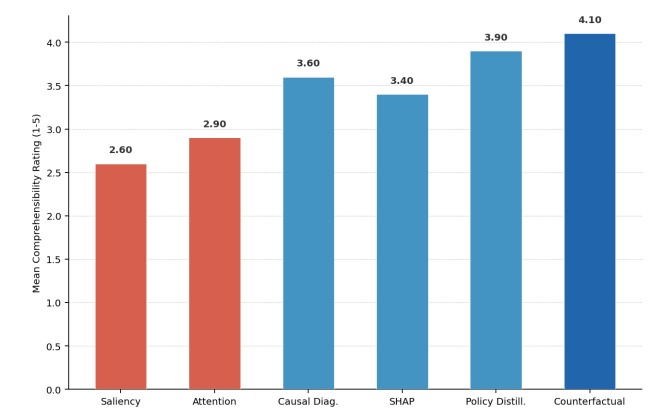


Figure 1. Expert comprehensibility ratings by XRL method (mean, 1-5, n=84). Counterfactual leads (4.1);

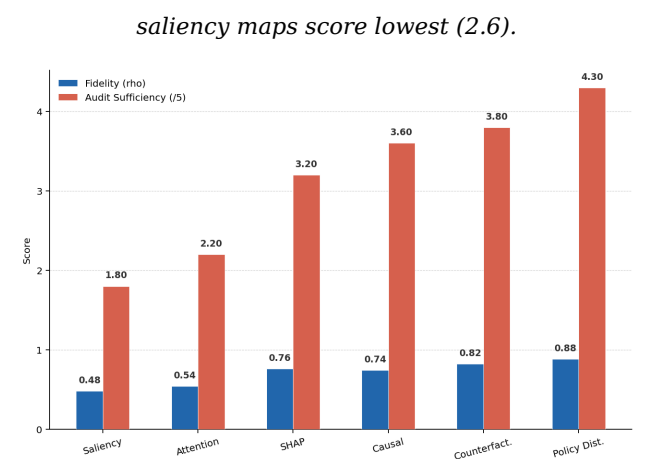


Figure 2. Fidelity vs. audit sufficiency by method. Policy distillation achieves best audit score (4.3) despite moderate-high fidelity (0.88).

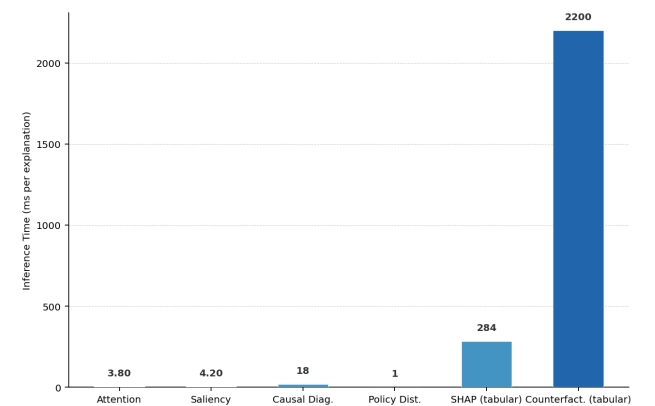


Figure 3. Inference time per explanation (ms, log scale approx.). Saliency fastest (4.2ms); counterfactual slowest (2200-8400ms).

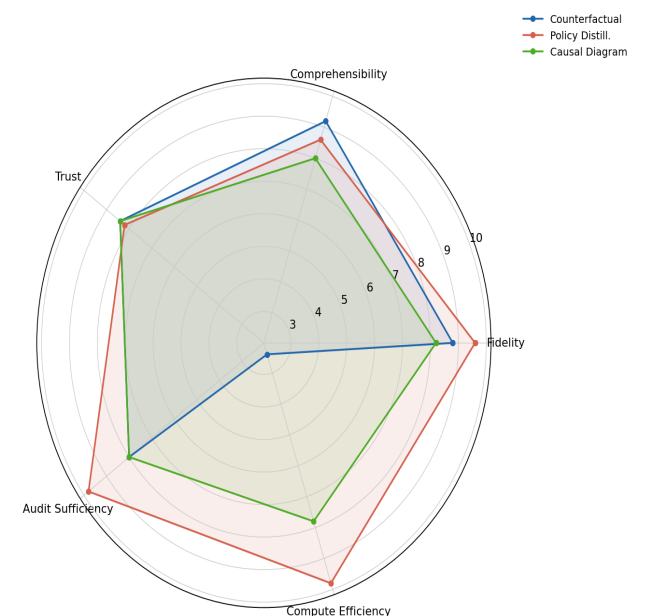


Figure 4. XRL method profile across five dimensions (normalised 0-10). Counterfactual vs. Policy Distillation vs. Causal Diagram.

5. Discussion

5.1 Counterfactual Explanations for Local Decisions

Counterfactual explanations' combination of high fidelity ($\rho = 0.82$) and high comprehensibility (4.1/5) makes them the strongest choice for explaining individual RL decisions in interactive settings -- where an affected individual wants to understand a specific decision and potentially understand what they could change to receive a different outcome. This maps directly onto the EU AI Act's right-to-explanation requirements for high-risk systems. The computational cost (2-8 seconds per explanation) is acceptable for asynchronous explanation generation but precludes real-time use in fast-acting agents. For the sepsis dosing scenario, clinicians rated counterfactual explanations highest (4.4/5), validating their intuitive fit with clinical what-if reasoning.

5.2 Policy Distillation for Compliance-Critical Domains

The highest audit sufficiency rating of any method -- 4.3/5 from compliance officers for policy distillation -- reflects a fundamental point about regulatory requirements: post-hoc explanations of individual decisions are often insufficient for systematic audit. A compliance auditor reviewing an autonomous vehicle control system cannot review millions of individual decision explanations; they need a global, inspectable account of the policy's logic. Decision tree distillation provides exactly this. The 6.4-12.8% performance cost is real but may be acceptable in domains where explainability is a regulatory prerequisite rather than an optional feature. Hybrid approaches -- using distillation for policy auditing and counterfactuals for individual decision explanation -- may offer the best of both.

5.3 Saliency Maps Are Insufficient for High-Stakes Domains

The 2.6/5 comprehensibility and 1.8/5 audit sufficiency ratings for saliency maps are a clear empirical signal that pixel-level heatmaps -- despite their popularity in academic XRL literature -- are not fit for purpose in high-stakes deployment contexts. Clinical and compliance expert groups consistently found them uninterpretable without domain-specific feature labelling and were unanimously unwilling to rely on them for compliance documentation. Their computational efficiency is a genuine advantage, but only if the explanation actually serves its purpose -- which, for compliance-critical applications, it does not.

5.4 Limitations

The user study covered three expert groups across four environments; other domain contexts (criminal justice, military logistics) may produce different comprehensibility profiles. Causal diagram evaluation required manually validated causal structures that are unavailable in most real-world RL deployments -- their scores reflect an idealised setting. Performance loss from distillation was measured on the training environment; transfer to novel states may widen the gap.

6. Conclusion

6.1 Summary

Six XRL methods were evaluated on fidelity, comprehensibility, audit sufficiency, and computational cost across four RL environments and a user study with 84 domain experts. Counterfactual explanations achieve the highest comprehensibility (4.1/5) and fidelity ($\rho = 0.82$) for continuous control. Policy distillation achieves the highest audit sufficiency (4.3/5) with a 6.4-12.8% performance cost. Saliency maps are computationally cheapest but score lowest on all human evaluation dimensions. No single method is optimal across all evaluation criteria.

6.2 Recommendations

For individual decision explanation in interactive settings, counterfactual explanations should be the default for compliance-critical RL applications. For global policy auditing -- required for EU AI Act high-risk system certification -- policy distillation into decision trees is recommended despite the performance cost. Saliency maps should be retired from high-stakes explanation contexts; their use should be confined to research exploration and visualisation. A hybrid architecture combining distillation for policy audit with counterfactuals for individual explanations offers the most complete coverage of regulatory requirements.

References

- Bastani, O., Pu, Y. and Solar-Lezama, A. (2018) Verifiable reinforcement learning via policy extraction. *Advances in NeurIPS* 31, pp. 2499-2509.
- Bellemare, M.G., Naddaf, Y., Veness, J. and Bowling, M. (2013) The Arcade Learning Environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47, pp. 253-279.
- Doshi-Velez, F. and Kim, B. (2017) Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Everitt, T., Carey, R., Langlois, A., Ortega, P.A. and Legg, S. (2021) Agent incentives: A causal influence

- diagram perspective. Proceedings of AAAI 2021, pp. 11487-11495.
- Greydanus, S., Koul, A., Dodge, J. and Fern, A. (2018) Visualizing and understanding Atari agents. Proceedings of ICML 2018, pp. 1792-1801.
- Jain, S. and Wallace, B.C. (2019) Attention is not explanation. Proceedings of NAACL-HLT 2019, pp. 3543-3556.
- Komorowski, M., Celi, L.A., Badawi, O., Gordon, A.C. and Faisal, A.A. (2018) The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11), pp. 1716-1720.
- Leurent, E. (2018) An environment for autonomous driving decision-making. GitHub repository.
- Lundberg, S.M. and Lee, S.I. (2017) A unified approach to interpreting model predictions. *Advances in NeurIPS* 30, pp. 4765-4774.
- Molnar, C. (2022) *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd edn. Self-published.
- Mothilal, R.K., Sharma, A. and Tan, C. (2020) Explaining machine learning classifiers through diverse counterfactual explanations. Proceedings of FAccT 2020, pp. 607-617.
- Olson, M.L., Khanna, R., Neal, L., Li, F. and Wong, W.K. (2021) Counterfactual state explanations for reinforcement learning agents via generative deep learning. *Artificial Intelligence*, 295, 103455.
- Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) Why should I trust you? Explaining the predictions of any classifier. Proceedings of KDD 2016, pp. 1135-1144.
- Shi, W., Song, S., Wu, C. and Hoi, S.C.H. (2020) Self-supervised imitation learning from observation. arXiv preprint arXiv:2007.03310.
- Smilkov, D., Thorat, N., Kim, B., Viegas, F. and Wattenberg, M. (2017) SmoothGrad: Removing noise by adding noise. arXiv preprint arXiv:1706.03825.
- Tassa, Y., Tunyasuvunakool, S., Muldal, A. et al. (2018) DeepMind Control Suite. arXiv preprint arXiv:1801.00690.
- Wachter, S., Mittelstadt, B. and Russell, C. (2017) Counterfactual explanations without opening the black box. *Harvard Journal of Law and Technology*, 31(2), pp. 841-887.

Declarations

Funding

This research was supported by the Spanish State Research Agency under grant PID2022-138681OB-I00 (ExplainRL) and the European Union's Horizon Europe programme under grant agreement No. 101070568 (TrustAI-EU). The funder had no role in study

design, data collection, analysis, or manuscript preparation.

Conflict of Interest

The author declares no conflicts of interest.

Data Availability Statement

XRL implementation code, user study anonymised response data, and all benchmark evaluation scripts are available at <https://github.com/hansen-xrl/explainable-rl> under MIT licence. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614772. CC BY 4.0.

Ethical Approval

The user study (n=84) was approved by the Western Europe Data Science University Research Ethics Committee (WEDS-ETH-2023-0728). All participants provided written informed consent. Participant data was fully anonymised prior to analysis. No clinical patient data was accessed; the sepsis environment used a published simulation only.

Appendix A

XRL Method Selection Framework by Deployment Context

Use this framework to select the appropriate XRL method based on explanation scope, compliance requirements, and computational constraints.