Incentivizing Exploration With Causal Curiosity as Intrinsic Motivation

Elias AOUN DURAND
ISIR / GATE
CNRS, Sorbonne Université
Paris, France
elias.aoundurand@gmail.com

Mateus JOFFILY
GATE UMR 5229
CNRS, Université de Lyon
69003 Lyon, France
joffily@gate.cnrs.fr

Mehdi KHAMASSI
ISIR UMR 7222
CNRS, Sorbonne Université
75005 Paris, France
mehdi.khamassi@upmc.fr

November 14, 2024

Abstract

Reinforcement learning (RL) has shown remarkable success in decision-making tasks but often lacks the ability to decipher and leverage causal relationships in complex environments. This paper introduces a novel "causal model-based reinforcement learning agent" that integrates causal inference with model-based RL to improve exploration and decision-making. Our approach incorporates an intrinsic motivation mechanism based on causal curiosity, quantified by the changes in the agent's internal causal model. We present an algorithm that maintains separate value functions for extrinsic rewards and intrinsic causal discovery, allowing for a balanced exploration of both task-oriented goals and causal structures. Theoretical analysis suggests convergence properties under certain conditions, while empirical results in a blackjack task and structural causal model environments demonstrate improved learning efficiency and strategic decision making compared to standard RL. This work contributes to bridging the gap between reinforcement learning and causal inference.

1 Introduction

Human beings are at the same time reinforcement learners [9] and prompt to causality judgments [5]. We use both evidential and causal knowledge to pursue our goals and infer properties of the world. Although evidential knowledge stems from statistical associations, causal knowledge allows us to understand the underlying mechanisms that generate these associations. Through a feedback signal originating from our surroundings, we learn and interact in a meaningful way. Yet, relations between the Reinforcement Learning (RL) framework and causal inference (CI) are not firmly established for now. Here, we seek to endow an artificial agent with causality-based motivation to explore and simulate it in order to more concretely appreciate the kind of behaviors that this would predict.

RL is an efficient way of representing decision-making problems and has proven crucial in solving concrete machine learning problems in the past [8]. CI is at the root of an ongoing revolution in statistics that provides a mathematical foundation for the notion of causality along with a set of well-developed tools to infer and build causal relations among statistical variables [17]. RL agents could benefit greatly from CI tools, evolving from 'evidential machines' that rely solely on statistical patterns to 'causal machines' capable of understanding and leveraging cause-effect relationships.

Accepted at the Intrinsically Motivated Open-ended Learning workshop at NeurIPS 2024

Moreover, CI seems suitable for a smooth integration within RL: the ladder of causation (observation, action, counterfactual reasoning) developed by Judea Pearl [10], resonates with RL considerations [1]. The leap taken by the CI revolution would bring RL agents to reason not only about statistical relationships like state transitions, but about causal relations, potentially greatly improving their capabilities.

2 Background and related work

The integration of RL and CI has received significant attention in recent years, with numerous approaches emerging to leverage causal understanding in decision-making processes [1, 5, 17, 2, 7, 4, 3]. One prominent direction is equipping RL agents with causal models for improved planning and counterfactual reasoning. Formally, a causal model comprises a graphical model, typically a directed acyclic graph (DAG), depicting causal relations between variables, and a set of structural equations describing the mechanisms underlying these relations. While traditional Model-based RL (MBRL) agents utilize forward, backward, or inverse models based on statistical correlations, these approaches can be misled by confounding variables - factors that influence both the apparent cause and effect. Causal models, on the contrary, can identify and account for these confounders, providing a more accurate representation of the underlying structure of the environment. Sontakke et al. [11] introduced a novel approach to causal curiosity in RL, proposing an intrinsic reward based on a distance function over causal factors - variables in the environment that have direct causal influences on other variables or outcomes. Their work demonstrates how incorporating causal understanding into the exploration process can lead to more efficient learning and better generalization. However, their method does not explicitly maintain or update a causal model of the environment, which can limit the agent's ability to reason about complex causal relationships. Recent work by Zeng et al. [17] provides a comprehensive survey of causal reinforcement learning, highlighting various techniques to integrate causal knowledge into RL frameworks. These include causal model learning, causal credit assignment, and causal transfer learning, all of which are aimed at enhancing the robustness and generalizability of RL agents. Gershman [4] adopted a cognitive science perspective to claim that while human learning often involves causal inferences, the RL framework implicitly assumes causality while only relying on (state,action)->state transition probability functions. Building on these foundations, our work introduces a novel model-based causal reinforcement learning agent that explicitly incorporates causal discovery and causal curiosity into the learning process. By maintaining separate value functions for extrinsic rewards and intrinsic causal discovery, our approach aims to balance task-oriented learning with exploration of the underlying causal structure of the environment.

3 Methods

3.1 Problem formulation

The environment is characterized by a Structural Causal Model (SCM), $M = \langle V, U, F, P(u) \rangle$, where V is the set of endogenous variables, U is the set of exogenous variables (noise terms), F is the set of structural equations, and $P(u)$ is the probability distribution over U. The observable state s_t at time t consists of values of a subset of V, denoted as $s_t = (v)_{v \in V_{\text{obs}}}$, where $V_{\text{obs}} \subseteq V$. If the agent takes action a_t at time t, it intervenes on the SCM, modifying the structural equations of the affected variables, noted M_{a_t}. The resulting state s_t consists of the values of the variables, denoted as $s_t = (v)_{v \in V}$. A special variable R, representing the reward, is considered external to the SCM, and the agent's objective is to maximize the cumulative value of R through its actions, which are conceptualized as interventions on environmental variables.

3.2 Intrinsic Motivation Through Causal Curiosity

Our model-based agent architecture integrates Q-learning with causal curiosity, incentivizing actions that modify the agent's internal causal model. The key components and their interactions are described in Algorithm 1. The algorithm maintains separate value functions for extrinsic rewards (Q) and intrinsic causal curiosity (I). The ϵ-greedy policy is applied to the combined value function $Q + \beta I$, where β balances between task-oriented exploitation and causal exploration. This ensures that the action selection process accounts for both extrinsic rewards and the potential for causal discovery. The intrinsic reward r_i is defined as the distance between the graphs $d(G, G')$, which quantifyes the causal discovery. Separate learning rates α and η allow for different learning dynamics in Q and I, respectively. The decay of the β parameter gradually shifts focus from exploration to exploitation. Our approach uses a causal discovery algorithm to generate a graphical causal model,

Algorithm 1 model-based RL with integrated causal curiosity

Input $\alpha > 0$, $\epsilon > 0$, $\gamma > 0$, $\beta > 0$, $\eta > 0$
Initialize $Q(s, a)$ arbitrarily for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
Initialize internal causal graph G
Initialize intrinsic value function $I(s, a) = 0$ for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
Initialize s to the initial state
repeat
    Choose a from s using policy derived from $Q + \beta I$ (e.g., ϵ-greedy)
    Take action a, observe reward r, and next state s'
    Update history dataset $D \leftarrow D \cup (a, s', r)$
    Compute G' from dataset D
    Compute intrinsic reward $r_i = d(G, G')$
    $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$
    $I(s, a) \leftarrow I(s, a) + \eta[r_i + \gamma \max_{a'} I(s', a') - I(s, a)]$
    $s \leftarrow s'$
    $G \leftarrow G'$
    Decrease β according to a decay schedule
until s is terminal

which represents qualitative causal relationships between variables. This could be augmented with structural equations to create a full quantitative SCM, allowing for more precise causal reasoning and potentially serving as a detailed transition model of the environment. We can use different causal discovery algorithms and graph metrics in our approach. For discovery algorithms, options include the Peter-Clark algorithm (PC) [13], the fast causal inference algorithm (FCI) [14], or simplified association and causation rules [10, 6]. The PC algorithm is widely used for its efficiency in sparse graphs, while the FCI algorithm is more suitable for scenarios with latent confounders.

3.3 Convergence Analysis

The convergence of our integrated model-based RL with the integrated causal curiosity algorithm depends on three key factors: the convergence of the PC algorithm for causal discovery, properties of the graph distance metric, and convergence of the intrinsic value function. We assume that the PC algorithm converges to the true causal graph G as $t \to \infty$, given the faithfulness of the probability distribution to G and the correctness of the conditional independence tests [12]. The graph distance metric $d(G_1, G_2)$ is assumed to be non-negative, symmetric, and satisfy the triangle inequality [15]. As $G_t \to G$, the intrinsic reward r_i approaches zero, leading I to converge to a fixed point corresponding to a policy that maintains the causal graph discovered. For a sketch of the proof, see the Appendix A.2.

4 Experimental Results

We evaluate our approach in two distinct environments with varying degrees of causal structure.

1. A blackjack task from OpenAI's gymnasium RL benchmark
2. A linear SCM task with three variables

We conducted experiments with 10 different random seeds to assess the robustness of our results. For the blackjack task, we trained our agents for 1000 episodes with 50000 trials per episode, while for the other environments, we used 1000 episodes with 1000 trials per episode. Statistical analysis across seeds showed that our agent consistently outperformed the baseline Q-learning agent on the blackjack task (mean improvement of 2.1%, $p < 0.001$). The causal discovery algorithm used in these experiments is the PC algorithm implemented by `causallearn`, as well as the Hamming distance for the computation of the causal graph distance [18] (see Annex A.3 for details).

4.1 Blackjack task

We evaluated the performance of three reinforcement learning agents in a blackjack simulation: a baseline random agent, a standard Q-learning model, and our causal model-based agent. Our results are summarized in Figure 1. The baseline agent consistently underperformed (Figure 1 left). In contrast, our agent not only began with higher initial rewards, but also showed a marked improvement

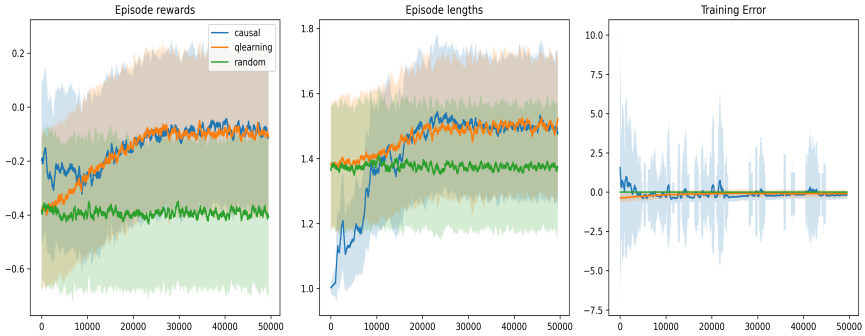

Figure 1: Performance comparison in the Blackjack task: cumulative rewards, episode lengths, and training errors for Baseline, Q-learning, and Causal Q-learning agents.

over time, indicating a progressive understanding of the environment's causal dynamics. This early advantage suggests that causal curiosity enables more efficient exploration of the state space from the beginning of training. Although both approaches eventually converge to similar performance levels after approximately 30,000 episodes, the causal agent maintains slightly higher variance in its rewards, indicating ongoing active exploration of the environment. The duration of the episodes increased for all agents, and our agent showed a significant increase before stabilizing, suggesting an effective strategy for prolonged gameplay (Figure 1 center).

The training error analysis (Figure 1, right) provides insight into learning dynamics. Although the causal agent exhibits greater variance in error signals, particularly during early training, this appears to be a constructive feature rather than a limitation. The periodic spikes in error coincide with periods of active causal model refinement, suggesting that the agent updates its understanding of the environment's causal structure. This more dynamic learning process, compared to the relatively stable error profile of standard Q-learning, aligns with our theoretical framework where causal curiosity drives ongoing exploration and model improvement.

These results validate our hypothesis that the incorporation of causal curiosity can enhance reinforcement learning performance, particularly in the critical early stages of learning. The maintained variance in both rewards and episodes lengths suggests that the agent continues to actively explore and refine its causal model even after achieving competitive performance, potentially leading to more robust and adaptable policies.

Performance Improvements	
Causal vs Random	+25.03%
Causal vs Q-Learning	+2.19%
Q-Learning vs Random	+22.84%
Statistical Significance	
Causal vs Random	$t = 134.19^{***}$
Causal vs Q-Learning	$t = 11.16^{***}$
Q-Learning vs Random	$t = 120.94^{***}$

$^{***}p < 0.001$

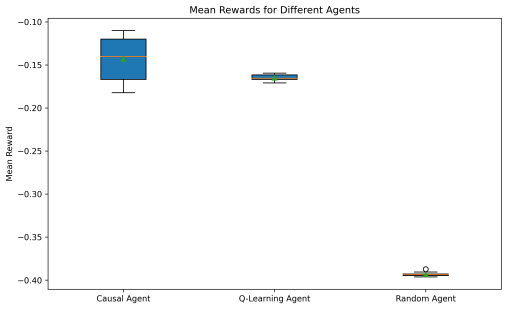

Figure 2: Comparative analysis of agent performance in the Blackjack task. Left: Statistical metrics showing performance improvements. Right: Distribution of mean rewards across different agent types.

Figure 2 presents a statistical analysis of agent performance in different learning approaches. The causal curiosity-driven agent demonstrates superior performance, achieving a 25.03% improvement over the random baseline and a notable 2.19% improvement over the standard Q-learning agent. This performance advantage is statistically significant and all comparisons yield $p < 0.001$. The box plots

reveal that the causal agent not only achieves better mean performance (approximately -0.14) but also exhibits greater variability in its rewards compared to the Q-learning agent (-0.16), suggesting more extensive exploration of the state space. Although both learning approaches significantly outperform the random baseline (-0.40), the higher variance of the causal agent and the superior mean performance indicate that causal curiosity effectively drives the exploration-exploitation trade-off. The relatively small t-statistic ($t = 11.16$) between the causal and Q-learning agents suggests that both approaches represent valid learning strategies.

4.2 Linear SCM task with three variables

We tested the algorithms in a linear Structural Causal Model (SCM) task with three variables X, Y, and Z, generated from a randomly determined causal graph. The structural equations were of the form: $X = U_X$, $Y = aX + U_Y$, and $Z = bX + cY + U_Z$, where a, b, and c are randomly initialized coefficients, and U_X, U_Y, U_Z are exogenous noise variables drawn from $\mathcal{N}(0, 100)$, ensuring a consistent starting point for each episode. The intervariate relationships within this SCM were defined as linear, providing a clear and quantifiable means of assessing the influence of one variable on another during the learning process. Actions were defined as interventions on the SCM replacing the original structural equations with $+1$ or -1 for each of the three variables.

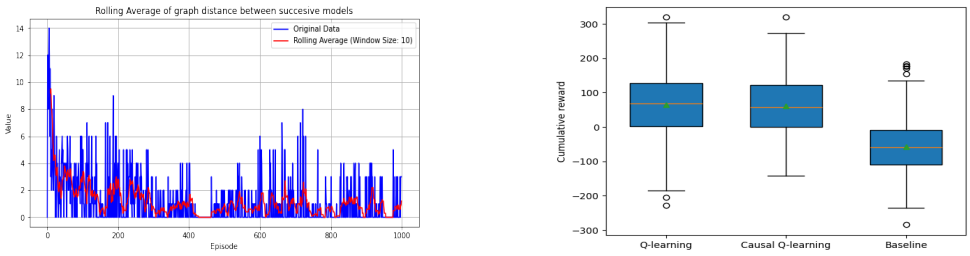

Figure 3: Performance in linear SCM task

The cumulative reward distribution, visualized in a boxplot (Figure 3 right), demonstrates that the Causal Q-learning approach has a more compact spread of the majority of its data points, as indicated by the shorter whiskers on the boxplot. This suggests that the Causal Q-learning are more densely packed around the median, with fewer extreme values. Figure 3 (left) illustrates the rolling average of the graph distance between successive causal models. The graph shows high initial variability followed by a general downward trend. As episodes progress, both the frequency and magnitude of fluctuations decrease, with the rolling average (red line) converging towards zero. This pattern indicates that the agent's causal model is stabilizing over time, suggesting successful learning of the environment's causal structure.

5 Discussion and Future work

This work introduced a causal exploration mechanism driven by an agent's intrinsic motivation to optimize its causal model. Our approach uses a causal discovery algorithm to generate a graph, which could be augmented with structural equations to create a complete structural causal model. This model could potentially serve as a transition model for the environment, allowing for more efficient state space factorization and planning. Our approach shows potential but has key limitations: high computational costs in causal discovery and graph comparison, plus the assumption that structural causal models can adequately represent the environment. Future work could explore alternative intrinsic reward formulations, such as the value of the information criterion proposed by Zemplenyi et al. [16], which estimates the value of interventions based on the expected reduction in posterior entropy. Or, we could explore using $d(G', \mathbb{E}[G''])$, where $\mathbb{E}[G'']$ is the expected future graph, to encourage a more forward looking causal exploration.

Acknowledgments

This work was supported by the French Agence Nationale de la Recherche (ANR-18-CE28-0016-03 CAUSAL Project and ANR-21-CE33-0019-01 ELSA Project).

References

[1] Elias Bareinboim, Andrew Forney, and Judea Pearl. "Bandits with Unobserved Confounders: A Causal Approach". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/795c7a7a5ec6b460ec00c5841019b9e9-Paper.pdf.

[2] K. J. Friston, J. Daunizeau, and S. J. Kiebel. "Reinforcement learning or active inference?" In: *PLoS One* 4.7 (2009), e6421. DOI: 10.1371/journal.pone.0006421. URL: https://doi.org/10.1371/journal.pone.0006421.

[3] Maxime Gasse et al. "Causal Reinforcement Learning using Observational and Interventional Data". en. In: *arXiv:2106.14421 [cs]* (June 2021). arXiv: 2106.14421. URL: http://arxiv.org/abs/2106.14421 (visited on 07/04/2021).

[4] Samuel J Gershman, Kenneth A Norman, and Yael Niv. "Discovering latent causes in reinforcement learning". In: *Current Opinion in Behavioral Sciences* 5 (2015). Neuroeconomics, pp. 43–50. ISSN: 2352-1546. DOI: https://doi.org/10.1016/j.cobeha.2015.07.007. URL: https://www.sciencedirect.com/science/article/pii/S2352154615001059.

[5] Samuel J. Gershman. *Reinforcement Learning and Causal Models*. en. Ed. by Michael R. Waldmann. Vol. 1. Oxford University Press, May 2017. DOI: 10.1093/oxfordhb/9780199399550.013.20. URL: http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780199399550.001.0001/oxfordhb-9780199399550-e-20 (visited on 04/14/2022).

[6] David Heckerman, Christopher Meek, and Gregory Cooper. "A Bayesian Approach to Causal Discovery". en. In: (), p. 28.

[7] Arquímides Méndez-Molina et al. "Causal Based Q-Learning". In: *Research in Computing Science* 149 (Mar. 2020), pp. 95–104.

[8] Volodymyr Mnih et al. "Human-level control through deep reinforcement learning". In: *Nature* 518 (7540 2015), pp. 529–533. ISSN: 1476-4687. DOI: 10.1038/nature14236. URL: https://doi.org/10.1038/nature14236.

[9] Stefano Palminteri et al. "Contextual modulation of value signals in reward and punishment learning". In: *Nature Communications* 6 (2015). URL: https://api.semanticscholar.org/CorpusID:3479783.

[10] Judea Pearl. "Causal inference in statistics: An overview". en. In: *Statistics Surveys* 3.none (Jan. 2009). ISSN: 1935-7516. DOI: 10.1214/09-SS057. URL: https://projecteuclid.org/journals/statistics-surveys/volume-3/issue-none/Causal-inference-in-statistics-An-overview/10.1214/09-SS057.full (visited on 09/06/2021).

[11] Sumedh A. Sontakke et al. "Causal Curiosity: RL Agents Discovering Self-supervised Experiments for Causal Representation Learning". en. In: *arXiv:2010.03110 [cs]* (Apr. 2021). arXiv: 2010.03110. URL: http://arxiv.org/abs/2010.03110 (visited on 04/27/2021).

[12] Peter Spirtes and Clark Glymour. "An Algorithm for Fast Recovery of Sparse Causal Graphs". en. In: *Social Science Computer Review* 9.1 (Apr. 1991), pp. 62–72. ISSN: 0894-4393, 1552-8286. DOI: 10.1177/089443939100900106. URL: http://journals.sagepub.com/doi/10.1177/089443939100900106 (visited on 10/08/2020).

[13] Peter Spirtes, Clark Glymour, and Richard Scheines. "Causation, prediction, and search". In: 1993. URL: https://api.semanticscholar.org/CorpusID:117765107.

[14] Peter L. Spirtes, Christopher Meek, and Thomas S. Richardson. *Causal Inference in the Presence of Latent Variables and Selection Bias*. 2013. arXiv: 1302.4983 [cs.AI]. URL: https://arxiv.org/abs/1302.4983.

[15] Peter Wills and François G. Meyer. "Metrics for graph comparison: A practitioner's guide". en. In: *PLOS ONE* 15.2 (Feb. 2020). Ed. by Pin-Yu Chen, e0228728. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0228728. URL: https://dx.plos.org/10.1371/journal.pone.0228728 (visited on 10/20/2020).

[16] Michele Zemplenyi and Jeffrey W. Miller. "Bayesian Optimal Experimental Design for Inferring Causal Structure". en. In: *arXiv:2103.15229 [stat]* (Mar. 2021). arXiv: 2103.15229. URL: http://arxiv.org/abs/2103.15229 (visited on 06/11/2021).

[17] Yan Zeng et al. *A Survey on Causal Reinforcement Learning*. en. arXiv:2302.05209 [cs]. Feb. 2023. URL: http://arxiv.org/abs/2302.05209 (visited on 02/13/2023).

[18] Yujia Zheng et al. "Causal-learn: Causal Discovery in Python". In: *arXiv preprint arXiv:2307.16405* (2023).

A Appendix

A.1 Structural Causal Models of Experimental Tasks

A.1.1 Blackjack Task SCM

For the blackjack task, we assumed a simplified SCM as follows:

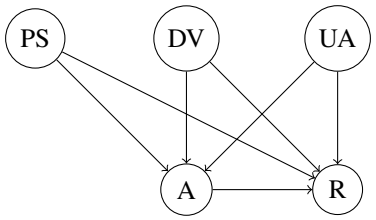

Figure 4: SCM for Blackjack Task. PS: Player's Sum, DV: Dealer's Visible Card, UA: Usable Ace, A: Action, R: Reward

A.1.2 Linear SCM Task

The linear SCM task with three variables is defined as:

$$
\begin{aligned}
X &= U_X \\
Y &= aX + U_Y \\
Z &= bX + cY + U_Z
\end{aligned}
$$

where a, b, and c are randomly initialized coefficients, and U_X, U_Y, U_Z are exogenous noise variables drawn from $\mathcal{N}(0, 100)$.

A.2 Proof Sketch for Convergence Analysis

To prove the convergence of our integrated model-based RL with causal curiosity, we need to establish:

1. Convergence of the PC algorithm
2. Properties of the graph distance metric
3. Convergence of the intrinsic value function

A.2.1 Convergence of PC Algorithm

Under the assumptions of causal sufficiency, faithfulness, and correct conditional independence tests, the PC algorithm converges to the true causal graph G as the sample size approaches infinity.

Proof Sketch:

1. As sample size increases, conditional independence tests become increasingly accurate.
2. The PC algorithm starts with a complete undirected graph and removes edges based on conditional independencies.
3. Given the correct tests, only the edges that correspond to true direct causal relationships remain.
4. The PC orientation rules correctly orient the remaining edges.

A.2.2 Properties of Graph Distance Metric

Let $d(G_1, G_2)$ be our graph distance metric. We assume:

1. Non-negativity: $d(G_1, G_2) \geq 0$
2. Symmetry: $d(G_1, G_2) = d(G_2, G_1)$
3. Triangle Inequality: $d(G_1, G_3) \leq d(G_1, G_2) + d(G_2, G_3)$

These properties ensure that our metric behaves like a proper distance function in the space of graphs.

A.2.3 Convergence of Intrinsic Value Function

As $G_t \to G$, the intrinsic reward $r_i \to 0$, and the intrinsic value function I converge to a fixed point.

Proof Sketch:

1. As the agent's internal causal model G_t approaches the true causal graph G, $d(G_t, G_{t+1}) \to 0$.
2. This implies $r_i \to 0$ as $t \to \infty$.
3. The update rule for I is: $I(s,a) \leftarrow I(s,a) + \eta[r_i + \gamma \max_{a'} I(s',a') - I(s,a)]$
4. As $r_i \to 0$, this update rule approaches: $I(s,a) \leftarrow I(s,a) + \eta[\gamma \max_{a'} I(s',a') - I(s,a)]$
5. This is equivalent to value iteration, which converges to a fixed point under standard assumptions.

While a complete proof requires more rigorous treatment, this sketch outlines the key steps towards establishing convergence of our algorithm.

A.3 Code Repository

The code implementation for this paper is available on GitHub. The repository contains the source code, experiments, and additional resources related to our Causal Curiosity approach.

Repository URL: `https://github.com/elichou/CausalCuriosity`