REFLECTIVE CAUSAL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We present the first empirical evaluation of Causal Reflection, a framework that equips agents with causal reasoning, structured self-correction, and explainable decision-making in dynamic environments. Standard reinforcement learning and large language model agents often fail under non-stationary conditions, relying on spurious correlations rather than robust causal models. Building on the theoretical foundations of Causal Reflection which formalizes causality as a temporal function and introduces a Reflect mechanism for hypothesis-driven model revision. We implement Reflective Causal Agents. Across a dynamic benchmark environment, these agents outperform ablated and associative baselines in adaptability, predictive accuracy, causal graph recovery, and hypothesis generation. Our results establish Causal Reflection as a practical approach toward robust, interpretable, and generalizable AI systems.

1 INTRODUCTION

The rapid advancement of artificial intelligence has intensified the demand for systems that can operate reliably in dynamic, uncertain environments and explain not just what happens, but why. This challenge exposes a fundamental weakness in two dominant AI paradigms. Traditional reinforcement learning (RL) agents, despite their success in maximizing reward signals, are often brittle and fail to generalize in the face of structural shifts or temporal delays. They optimize for outcomes without modeling the underlying temporal cause-effect relationships that govern the environment, rendering them opaque and poorly aligned with human expectations in evolving settings. This limitation is particularly pronounced in real-world applications where agents must adapt to changing dynamics (Kiciman et al., 2023; Seitzer et al., 2021).

Similarly, large language models (LLMs) excel at knowledge synthesis and reasoning over static information but lack an inherent understanding of causality in temporal contexts. Their reliance on memorized patterns and spurious correlations learned from vast datasets makes their causal reasoning shallow and unreliable when faced with novel scenarios (Jiao et al., 2024; Du et al., 2017). The failure of these correlation-based systems in non-stationary environments represents a critical barrier to deploying robust and trustworthy AI.

To address this, we turn to the Causal Reflection framework (Aryan & Liu, 2025), a theoretical architecture designed to shift an agent's objective from simple policy optimization to the construction and revision of an accurate, dynamic causal model of its environment.

Our paper provides the first empirical operationalization and validation of the framework. We implement Reflective Causal Agents and subject them to a series of rigorous experiments in simulated environments characterized by the exact challenges, such as structural breaks, time-delayed effects, and latent confounders that the framework was designed to overcome. We investigate whether agents equipped with Causal Reflection can adapt more effectively, achieve higher predictive accuracy, and correctly identify changes in the underlying causal graph compared to established baselines.

Our contributions are threefold: (1) we present the first functional implementation of Reflective Causal Agents based on the Causal Reflection framework, enabling agents to actively detect and revise failures in their internal causal models. (2) we introduce a dynamic benchmark environment with time-varying causal structures and sparse rewards, specifically designed to test agent adaptability under non-stationary conditions. (3) we provide a comprehensive empirical evaluation, comparing reflective agents to ablations and associative baselines across multiple metrics.

The results demonstrate that full Reflective Causal Agents achieve superior performance: they attain lower prediction error (MSE = 0.0004), higher cumulative rewards (0.72 ± 0.45), accurate causal graph recovery (F1 = 0.909), and maintain a high success rate (72%) with minimal reflection triggers. In contrast, ablated and associative variants exhibit slower adaptation, weaker causal modeling, and lower task performance.

These findings highlight that reflective agents not only adapt faster to environmental changes but also develop more interpretable and generalizable internal models, marking a step toward practical, causally grounded AI systems that understand, rather than merely optimize for, the evolving structure of their environment.

2 Related Work

Our work is situated at the intersection of causal reinforcement learning, temporal causal modeling, and self-reflective AI agents. This section reviews these areas to position our contribution.

2.1 Causal Reinforcement Learning

The integration of causal inference into reinforcement learning, known as Causal Reinforcement Learning (CRL), has emerged as a critical research direction to improve policy robustness, sample efficiency, and generalization (Deng et al., 2023). CRL approaches typically leverage formalisms like Structural Causal Models (SCMs) to represent the environment's underlying mechanics, moving beyond the spurious correlations that plague traditional RL. By embedding causal knowledge, agents can reason on higher rungs of Pearl's Causal Hierarchy, enabling interventional and counterfactual reasoning to improve decision-making (Bareinboim et al., 2021). Most existing CRL methods, however, assume a static, time-invariant causal graph, which limits their applicability in real-world scenarios where causal relationships can evolve over time.

2.2 Temporal Causality and Non-Stationarity

Modeling causal relationships in temporal data presents unique challenges, as cause-effect dynamics can change over time. Temporal Structural Causal Models (TSCMs) extend SCMs to capture such time-varying dependencies (Gkorgkolis et al., 2025). Other methods focus on learning from non-stationary data, where the data-generating process itself shifts (Zhang et al., 2017; Calderon & Berman, 2024). While these approaches can detect or represent non-stationarity, they typically rely on retraining and lack mechanisms for autonomous, online adaptation. However, the Causal Reflection framework addresses this limitation by equipping agents with an explicit, agent-centric mechanism to detect causal model failures in real-time and trigger hypothesis-driven revision. This enables continuous adaptation in dynamically changing environments.

2.3 Self-Reflection and Model Revision in AI Agents

Self-reflection has emerged as a strategy to improve AI agent performance. Reflexion (Shinn et al., 2023), for example, allows LLM agents to reflect verbally on past failures to refine future plans. Such approaches, however, treat reflection as a heuristic, linguistic process: the agent critiques past actions in unstructured natural language to generate new plans. Alternatively, Causal Reflection introduces a new approach where reflection is formal, structured, and model-based, triggered by quantitative signals indicating discrepancies between the agent's causal model and reality. This structured process allows the agent to generate falsifiable hypotheses about its internal model and revise it systematically. In contrast to heuristic reflection, this approach elevates self-correction to a principled, inference-driven mechanism.

2.4 Large Language Models for Causal Explanation

Large Language Models (LLMs) have shown strong capabilities in causal reasoning tasks, including counterfactual generation and pairwise causal discovery (Kiciman et al., 2023). Yet, their reasoning is often shallow and pattern-based, corresponding primarily to associational (Level-1) reasoning rather than mechanistic understanding. Consequently, LLMs can fail unpredictably in novel or

complex scenarios (Wang & Shen, 2024; Ashwani et al., 2024; Chi et al., 2024). On the other hand, Causal Reflection framework explicitly separates reasoning from explanation. The LLM functions as a constrained generative interpreter, translating the outputs of a formal causal model into natural language explanations or counterfactual. All core causal logic reside within the temporal causal function and Reflect mechanism, ensuring that explanations remain grounded in a verifiable causal structure while mitigating hallucination risks.

3 The Causal Reflection Framework

The Causal Reflection framework shifts an agent's objective from purely maximizing cumulative reward to building and maintaining an accurate, dynamic causal model of its environment.

The framework extends the state-action representation in RL to include temporal and perturbation factors, formalized as the causal tuple (S_t, A_t, T_t, δ) where:

- **State (S_t):** Complete configuration of the environment at time t.
- **Action (A_t):** Intervention performed by the agent.
- **Time (T_t):** Imposes temporal ordering to distinguish causation from correlation.
- **Perturbation Factor (δ):** Represents small, often unobserved influences that can produce nonlinear effects, capturing domain drift, hidden confounders, or unexpected events.

As described in Aryan & Liu (2025), this causal function

$$C(S_t, A_t, T_t, \delta) \rightarrow S_{t+k} \tag{1}$$

maps the current state and action to future states while accounting for temporal dynamics and perturbations. Its structure supports nonlinear propagation, delayed effects, and sensitivity to initial conditions. Self-correction is operationalized through the `Reflect` mechanism, which is triggered by prediction errors. This mechanism allows the agent to generate and evaluate hypotheses about its causal model and revise the model systematically. Structured reflection provides targeted adaptation in dynamic environments.

The Reflective Causal Agents follow a similar modular architecture, separating decision-making from causal reasoning:

- **Causal Inference Engine:** Handles predictive reasoning through the temporal causal function.
- **Reflect Mechanism:** Performs structured model revision when prediction errors exceed thresholds.
- **LLM-Based Interpreter:** Translates formal causal outputs and hypotheses into natural language for human interpretation.

This key inspiration for the modular design was to ensure that causal inference is maintained while leveraging LLMs for explanation without risking ungrounded reasoning.

4 Experimental Methodology

The goal of our experiments is to empirically evaluate the effectiveness of Reflective Causal Agents (RCA). We test whether agents equipped with the temporal causal function and the Reflect mechanism can outperform standard baselines in dynamic, perturbed environments, specifically in terms of prediction accuracy, adaptability, and explanatory depth.

4.1 Simulation Environment

We designed a custom environment, `TimeoutDoorKeyEnv`, adapted from Towers et al. (2024) to test temporal causality, perturbation sensitivity, and causal reasoning. The environment is an 8×8 grid world with an agent, a key, a locked door, and a goal.

The Causal Tuple $C(S_t, A_t, T_t, \delta)$ parameters are defined as-

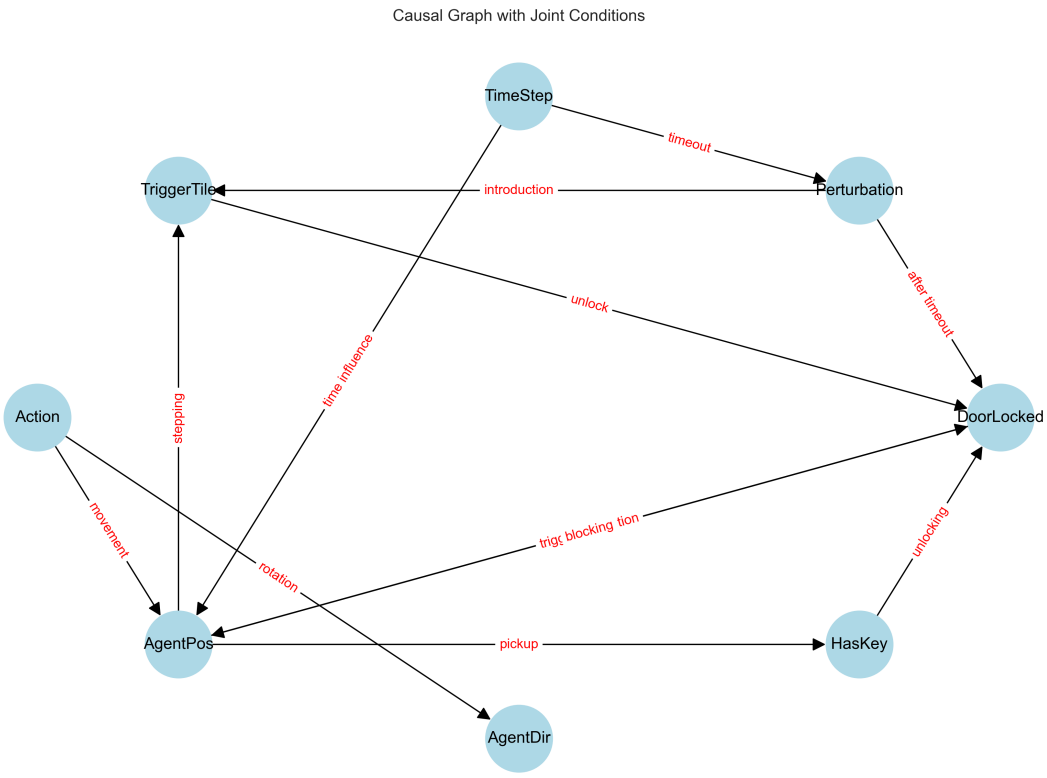

Figure 1: Causal graph

- **State Space (S_t):** This includes position, velocity, task indicators (e.g., switches, goals), and latent parameters governing dynamics.
- **Action Space (A_t):** This refers to the discrete actions (move, interact, delay) with deterministic low-level effects but nonlinear high-level consequences.
- **Perturbation Events (δ):** This indicates the structural breaks triggered after inactivity measured in time t (say, 30 steps) or at random intervals.

These perturbations define the ground-truth causal graph shifts against which we evaluate recovery. During each episode, perturbations alter the underlying causal graph via rule changes (e.g.a switch that previously opened a door now closes it), latent confounders (e.g., altered energy decay affecting motion), or adversarial noise (e.g., spurious influences on rewards)

We employ a three-stage evaluation:

1. **Training Phase:** 20 episodes with perturbations enabled and reflection active.
2. **Held-Out Evaluation:** 100 episodes with unseen perturbations. Reflection is disabled (frozen model) to test generalization.
3. **Ablation Studies:** 100 episodes per variant, selectively disabling reflection or perturbation modeling.

Episodes are capped at 100 steps during evaluation. For evaluation, we compare four agent variants:

1. Full Reflective Causal Agent (RCA), *our core contribution:* Temporal causal function + reflection + perturbation modeling.
2. No Reflection: Reflection mechanism disabled (error threshold $\to \infty$).
3. No Perturbation Modeling: Perturbation factor removed ($\delta = 0$).
4. Associative Baseline: Linear predictor without causal reasoning.

4.2 EVALUATION METRICS:

To measure the comparative performance of the four variants, we assess performance along five key dimensions:

Prediction Accuracy (MSE): Measures the error between predicted and observed states:

$$\text{PredictionError} = \frac{1}{T}\sum_{t=1}^{T} \|\hat{S}_{t+k} - S_{t+k}^{obs}\|^2 \tag{2}$$

Lower MSE indicates more accurate causal understanding. Spikes in MSE signal perturbations or distribution shifts.

Causal Graph Recovery (F1-Score): Accuracy of inferred causal edges compared to ground-truth:

$$\text{GraphF1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Higher F1 reflects better causal structure learning. High precision but low recall indicates conservative edge detection whereas high recall but low precision indicates aggressive detection.

Reward Achievement: Total accumulated reward per episode:

$$\text{TotalReward} = \sum_{t=1}^{T} R_t \tag{4}$$

Higher rewards indicate more efficient task completion. Positive rewards correspond to successful episodes, while negative rewards indicate failure.

Reflection Frequency: Number of reflection triggers per episode, when prediction error exceeds a threshold:

$$\text{ReflectionFrequency} = \frac{\#\text{steps with reflection triggered}}{\text{episode length}} \tag{5}$$

Excessive reflections indicate over-sensitivity, while zero reflections indicate missed learning opportunities. Therefore, moderate frequency signals balanced learning sensitivity.

Success Rate: Fraction of episodes where the agent successfully completes the task (reaches the goal with positive total reward):

$$\text{SuccessRate} = \frac{\#\text{episodes with TotalReward} > 0}{\text{Total number of episodes}} \tag{6}$$

High success rate implies effective goal-directed behavior whereas low success rate indicates difficulties in task completion.

All metrics are reported as mean $\pm$ standard deviation across 100 held-out episodes.

4.3 IMPLEMENTATION

The *Reflective Causal Agent (RCA)* is implemented in Python and includes four core components. The Causal Encoder maps raw environmental states to a structured set of causal variables. The Causal Graph maintains a dynamic directed acyclic graph representing the agent's current understanding of causal dependencies. The Reflection Engine monitors prediction errors and triggers updates to the causal model when discrepancies exceed a threshold. Finally, the Causal Planner leverages the learned causal graph to perform counterfactual action selection, operating with a planning horizon of 1 and a beam width of 2.

4.4 CONFIGURATION

- **Environment:** 8×8 grid, timeout threshold = 30, perturbations enabled.
- **Training:** error threshold = 0.007, reflection enabled, $\epsilon = 0.1$ with decay.
- **Evaluation:** reflection disabled, maximum steps = 100.
- **Logging:** All state transitions, reflections, predictions, and causal graphs are serialized to JSON.

5 RESULTS

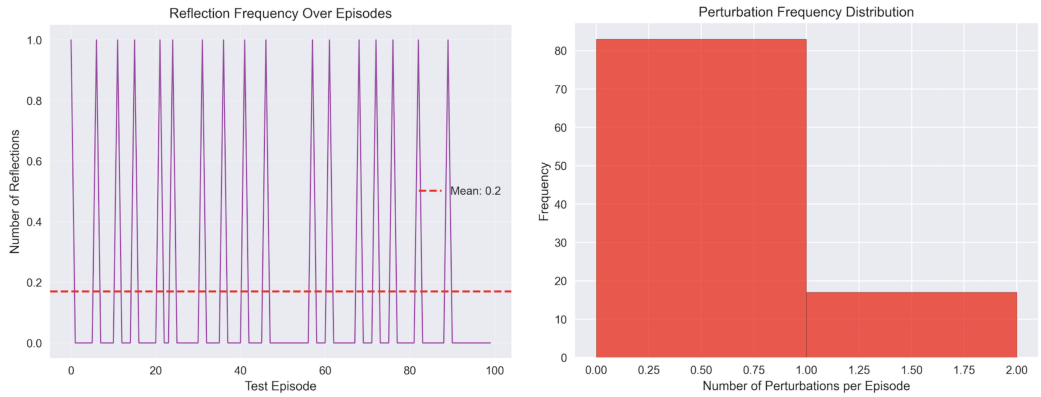

Figure 2: Reflection frequency over episodes, perturbation frequency distribution.

We evaluated the performance of four agent variants across five key metrics in our dynamic simulation environment, as summarized in Table 1. Each metric provides insight into different aspects of the agents' causal reasoning, adaptability, and task efficiency.

Prediction Accuracy (MSE): The Full Reflective Causal Agent achieved the lowest mean squared error (0.0004), indicating highly accurate internal modeling of the environment. By contrast, agents without the Reflect mechanism or perturbation modeling exhibited higher MSE (0.0009), reflecting reduced prediction fidelity, while the purely associative baseline performed worst (0.0487), highlighting the necessity of structured causal reasoning in dynamic, perturbed environments. These results demonstrate that the combination of reflection and perturbation-aware causal modeling enables precise state predictions even under distribution shifts.

Reward Achievement: Reflective agents consistently achieved higher cumulative rewards, with the full system averaging 0.72 ± 0.45 per episode, outperforming the No Reflect variant (0.62 ± 0.46), the No Perturbation Modeling agent (0.56 ± 0.48), and the associative baseline (0.13 ± 0.30). These differences underscore that accurate causal modeling and adaptive reflection directly translate to more effective task completion in non-stationary environments.

Causal Graph Recovery (F1-Score): Only agents employing the Reflect mechanism were able to reconstruct the underlying causal structure reliably. The Full Reflective Causal Agent achieved near-perfect graph F1 (0.909), whereas disabling reflection completely prevented causal structure recovery (F1 = 0.0), and ignoring perturbations yielded intermediate results (0.460). The associative baseline lacked any causal representation (F1 = 0.333), confirming that causal inference is essential for capturing dynamic dependencies.

Reflection Frequency: The Full Reflective Causal Agent triggered reflections sparingly (0.2 per episode), reflecting a balanced sensitivity to prediction errors. Agents without reflection or perturbation modeling either triggered no reflections or excessively frequent reflections (up to 46.8 in the associative baseline), indicating either missed learning opportunities or overreactive behavior. This highlights the importance of a calibrated Reflect mechanism for efficient, targeted adaptation.

Success Rate: The probability of task completion followed the same trend. The Full Reflective Causal Agent achieved a 72% success rate, compared to 64% for No Reflection, 58% for No Perturbation Modeling, and only 10% for the associative baseline. These results emphasize that reflection combined with perturbation-aware causal reasoning substantially improves goal-directed performance.

Overall, these results demonstrate that the Full Reflective Causal Agent not only achieves superior predictive accuracy and causal understanding but also adapts efficiently to environmental changes while maintaining high task success. Ablation studies further confirm that both the Reflect mechanism and perturbation modeling are critical contributors to these gains, validating the design choices of the Causal Reflection framework.

Table 1: Summary of agent performance across evaluation metrics

Variant	MSE	Reward (mean±std)	Graph F1 (mean±std)	Mean Refl.	Success
Full Reflective Causal Agent	0.0004	0.72 ± 0.45	0.909 ± 0.000	0.2	72.0%
Causal Reflection (No Reflect)	0.0009	0.62 ± 0.46	0.000 ± 0.000	0.0	64.0%
No Perturbation Modeling	0.0009	0.56 ± 0.48	0.460 ± 0.013	1.1	58.0%
Associative Baseline	0.0487	0.13 ± 0.30	0.333 ± 0.003	46.8	10.0%

6 Conclusion

In this work, we presented the first empirical implementation and validation of Reflective Causal Agents, building on the Causal Reflection framework (Aryan & Liu, 2025). Our experiments demonstrate that prioritizing the construction and revision of a dynamic causal model—rather than simple reward maximization yields agents that are significantly more robust, adaptive, and interpretable in non-stationary environments. The full reflective agent not only recovers more quickly from structural changes but also develops a more accurate internal representation of its world, resulting in superior predictive and decision-making performance. These findings establish a concrete path toward AI systems that do not merely act, but understand and reason about the evolving causal structure of their environment.

7 Limitations and Future Works

While our framework demonstrates a promising new direction, several challenges and opportunities for future research remain.

Scalability: Modeling complex, high-dimensional systems is computationally intensive. As noted in the foundational framework, inferring the causal function C in a large state space could be a significant challenge, echoing broader issues in high-dimensional causal inference. Future work should explore factorization and representation learning techniques, such as Variational Autoencoders (VAEs), to learn lower-dimensional, causally sufficient state spaces that make the problem more tractable.

LLM Fidelity and Controllability: The framework relies on the LLM to be a faithful interpreter of the formal model's output. However, LLMs can "hallucinate" or misrepresent information, creating a potential mismatch between the agent's causal inference and its explanation. Developing methods to quantify and mitigate these "translation errors" is crucial for ensuring that the natural language outputs remain rigorously grounded in the underlying causal model.

Hypothesis Quality and Search: In this work, we did not deeply evaluate the quality of the generated causal hypotheses beyond their impact on predictive accuracy. Future studies should assess hypotheses based on their plausibility, actionability, and fidelity to the true structural shift. Furthermore, the search space for potential hypotheses can be vast; developing more efficient search and testing strategies is a key avenue for improving the Reflect mechanism's efficiency.

Future Work: A compelling direction is to extend the Causal Reflection framework to more environments including Multi-Agent Reinforcement Learning (MARL). In MARL settings, an agent must model not only the causal impact of its own actions but also the causal influence of other agents' actions on the environment and on each other. Applying our framework to such complex social and strategic interactions is a nascent but critical research area for understanding and building cooperative and competitive AI systems (Briglia et al., 2025).

References

Abi Aryan and Zac Liu. Causal reflection with language models, 2025. URL https://arxiv.org/abs/2508.04495.

Shubham Ashwani, Kavya Hegde, Narendra Reddy Mannuru, Divyansh Singh Sengar, Mayank Jindal, K. C. R. Kathala, and Aseem Chadha. Cause and effect: can large language models truly understand causality? In *Proceedings of the AAAI Symposium Series*, volume 4, pp. 2–9, 2024.

Elias Bareinboim, Jiji Zhang, and Sang-Hyeun Lee. An introduction to causal reinforcement learning. *arXiv preprint arXiv:2101.06498*, 2021.

Giovanni Briglia, Stefano Mariani, and Franco Zambonelli. A roadmap towards improving multi-agent reinforcement learning with causal discovery and inference. *arXiv preprint arXiv:2503.17803*, 2025.

Jonathan Calderon and Gordon J. Berman. Inferring the time-varying coupling of dynamical systems with temporal convolutional autoencoders. *arXiv preprint arXiv:2406.03212*, 2024.

Heng Chi, Haotian Li, Wen Yang, Feiyi Liu, Lihui Lan, Xiang Ren, and B. Han. Unveiling causal reasoning in large language models: Reality or mirage? In *Advances in Neural Information Processing Systems*, volume 37, pp. 96640–96670, 2024.

Zheyuan Deng, Jun Jiang, Guodong Long, and Chengqi Zhang. Causal reinforcement learning: A survey. *arXiv preprint arXiv:2307.01452*, 2023.

Siqi Du, Guang Song, Lixin Han, and Han Hong. Temporal causal inference with time lag. *Neural computation*, 30(1):271–291, 2017.

Nikolaos Gkorgkolis, Nikolaos Kougioulis, Minlan Wang, Baris Caglayan, Alberto Tonon, Diego Simionato, and Ioannis Tsamardinos. Temporal causal-based simulation for realistic time-series generation. *arXiv preprint arXiv:2506.02084*, 2025.

Lijing Jiao, Ya-nan Wang, Xuan Liu, Lisha Li, Fang Liu, Wenping Ma, and Boyu Hou. Causal inference meets deep learning: A comprehensive survey. *Research*, 7:0467, 2024.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2023.

Maximilian Seitzer, Bernhard Schölkopf, and Georg Martius. Causal influence detection for improving efficiency in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22905–22918, 2021.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments, 2024. URL `https://arxiv.org/abs/2407.17032`.

Linyi Wang and Yuxi Shen. Evaluating causal reasoning capabilities of large language models: A systematic analysis across three scenarios. *Electronics*, 13(23):4584, 2024.

Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from nonstationary/heterogeneous data: Principle and method. *arXiv preprint arXiv:1708.07171*, 2017.