Centre for Research Training

Science Foundation Ireland sfi For what's next

ReCCoVER: Detecting Causal Confusion for Explainable Reinforcement Learning

Jasmina Gajcin and Ivana Dusparic

HOST INSTITUTION

UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh

PARTNER INSTITUTIONS

University of Limerick Ollscoil Luimnigh

NUI Galway Ollscoil na Gaillimhe

DCU

Trinity College Dublin Coláiste na Tríonóide, Baile Átha Cliath The University of Dublin

CRT·AI SFI Centre for Research Training in Artificial Intelligence

Causal Confusion

- Phenomenon where RL agent relies on spurious correlations to make a decision.
- If spurious correlation is broken, agent is likely to make a wrong decision
- It is necessary to verify agent is not relying on spurious correlations before deployment.

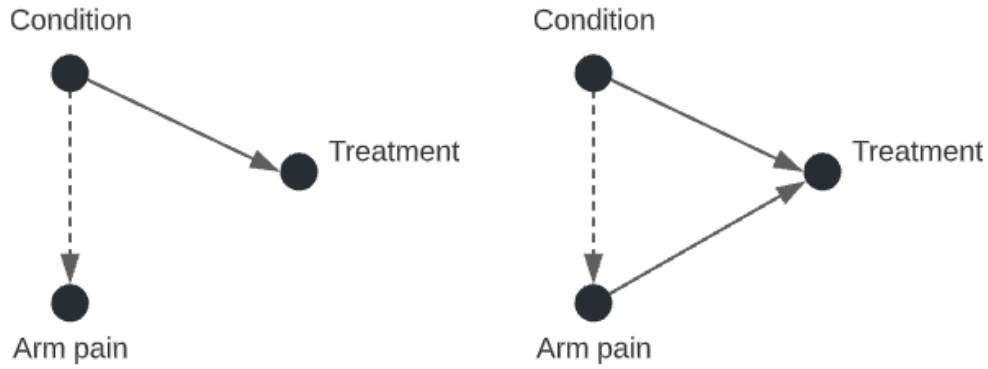


Figure 1: Different causal structures in different situations for a medical decision system. Left: if *Condition = HeartAttack*, arm pain should be disregarded. Right: If *Condition = ArmInjury*, arm pain is an important factor in determining treatment

[1] De Haan, P., Jayaraman, D., & Levine, S. (2019). Causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, *32*.

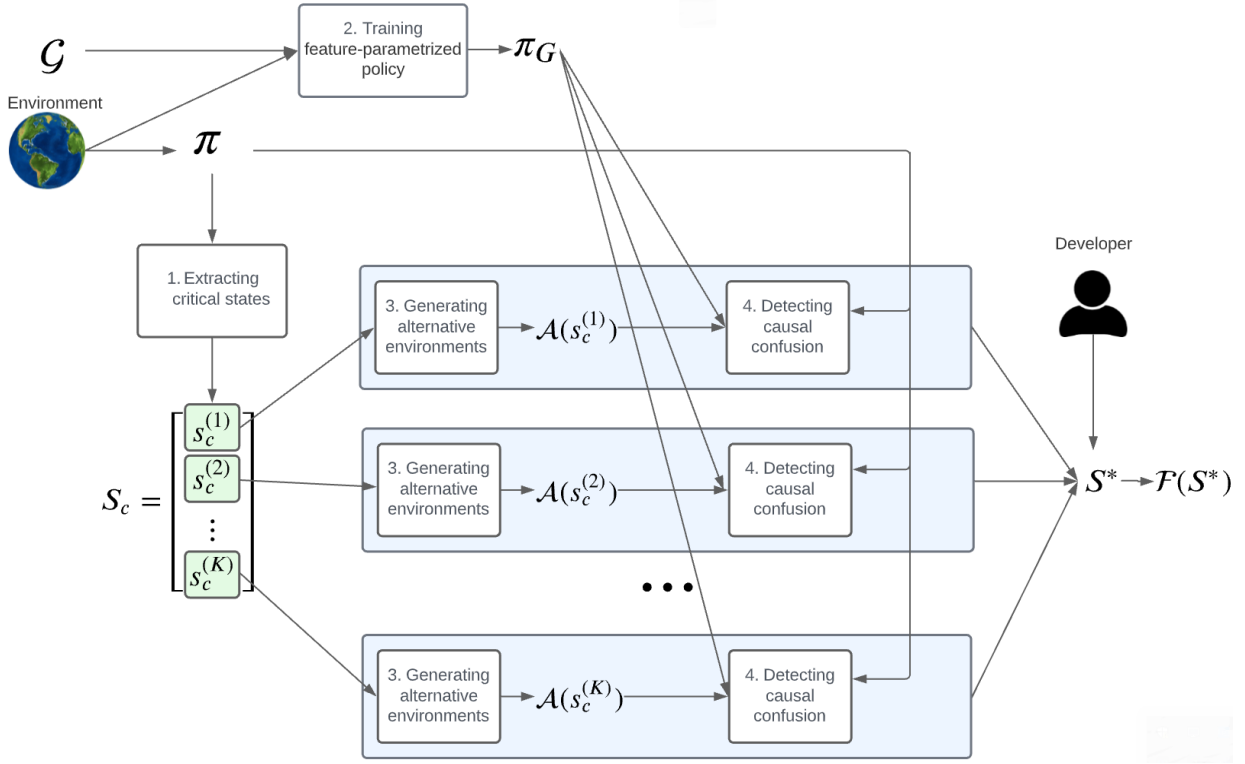
Research Goal

- **Research goal**: increase understanding of agent's behavior by detecting and correcting causal confusion in agent's behavior before deployment.
- Verify agent's reasoning in critical states by testing it in *alternative environments* where certain correlations between features are broken.

ReCCoVER

- Approach for detecting and correcting causal confusion in agent's reasoning

1. Extracting Critical States
2. Training Feature-Parametrized Policy
3. Generating Alternative Environments
4. Detecting Causal Confusion



Extracting Critical States

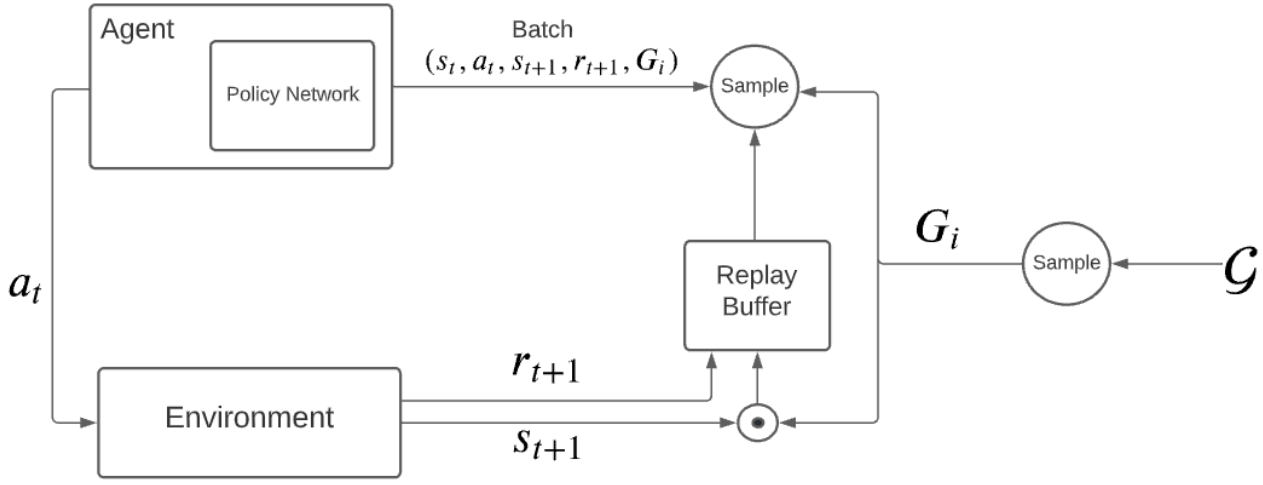
- Focus on verifying behavior in critical states.
- Consider states in which agent reaches a *local maximum* in terms of value function[2]

$$S_C = \{s \in S | v_\pi(s) \geq v_\pi(s'), \forall s' \in T_S\}$$

[2] Sequeira, P., Gervasio, M.: Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations. Artificial Intelligence 288, 103367 (2020)

Training Feature-parametrized Policy

- Simultaneously training a separate policy for each feature subset
- Evaluating this policy helps us understand how ignoring some features affects agent's reasoning



Generating Alternative Environments

- For each critical state s_c, a set of alternative environments $A(s_c)$ is generated in which certain correlations between feature do not hold
- Intervening on variables can break correlations between them
- We generate alternative environments by performing interventions that lead to novel states:

$$N(s) = \begin{cases} \frac{1}{\sqrt{n(s)}}, n \geq 1 \\ 1, \quad else \end{cases}$$

[3] Şimşek, Ö., & Barto, A. G. (2004, July). Using relative novelty to identify useful temporal abstractions in reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning* (p. 95).

Detecting Causal Confusion

- Evaluate both π and π_G in alternative environment $A(s_C)$.
- Causal confusion is detected in state s_C if policy π fails, but a policy $\pi_G(G')$ relying on a different subset of features G' performs well.
- We manually examine the results of causal confusion detection to extract a state subset $S^*$where alternative subset of features G' should be used.

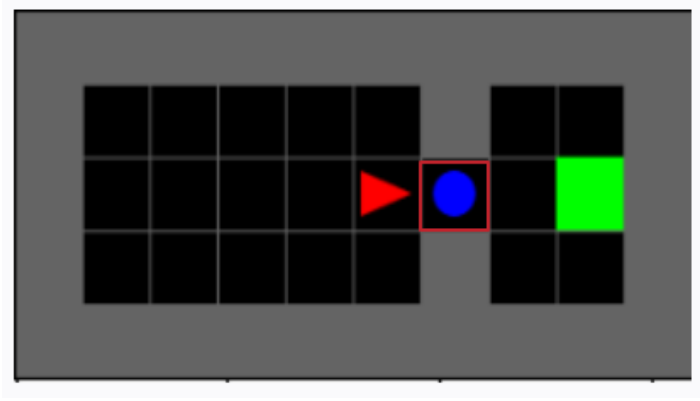
Evaluation Scenarios

Taxi Environment

- Added a spurious feature – passenger descriptor, which is highly correlated with destination choice.

MiniGridworld Traffic Environment

- Different situations require different features



Evaluation Goals

1. **Goal 1 (Critical states):** manual verification
2. **Goal 2 (Recognizing causal confusion):** train policy $\pi_{confused}$ by forcing it to rely on the spurious correlation and apply ReCCoVER to detect causal confusion
3. **Goal 3 (Proposing correct feature subset):** train policy $\pi_{correct}$ on feature subset extracted by ReCCoVER

Evaluation Results

Environment	Number of critical states	Number of states where causal confusion is detected	
		$\pi_{confused}$	$\pi_{correct}$
Taxi environment	16	4	0
Minigridworld traffic environment	2	1	0

Future Work

- Investigate the scalability of ReCCoVER
- Automate the manual parts of the algorithm
- Explore applications beyond explainability (e.g. transfer learning)

Centre for Research Training

Science Foundation Ireland sfi For what's next

Q/A?