Confounding Robust Deep Reinforcement Learning: A Causal Approach

Mingxuan Li

Junzhe Zhang

Elias Bareinboim

Speaker: Mingxuan Li

CausalAI Lab @ Columbia University

December 2025

Motivations

Motivations

Off-policy learning with unknown behavioral policies may introduce biases.

Motivations

Off-policy learning with unknown behavioral policies may introduce biases.

Q1: Can standard algorithm like DQN still learn optimal policies?

Motivations

Off-policy learning with unknown behavioral policies may introduce biases.

Q1: Can standard algorithm like DQN still learn optimal policies?

Such biases are often due to observation/action space mismatches.

Motivations

Off-policy learning with unknown behavioral policies may introduce biases.

Q1: Can standard algorithm like DQN still learn optimal policies?

Such biases are often due to observation/action space mismatches.

Q2: How do we model it formally with causal inference tools?

Motivations

Off-policy learning with unknown behavioral policies may introduce biases.

Q1: Can standard algorithm like DQN still learn optimal policies?

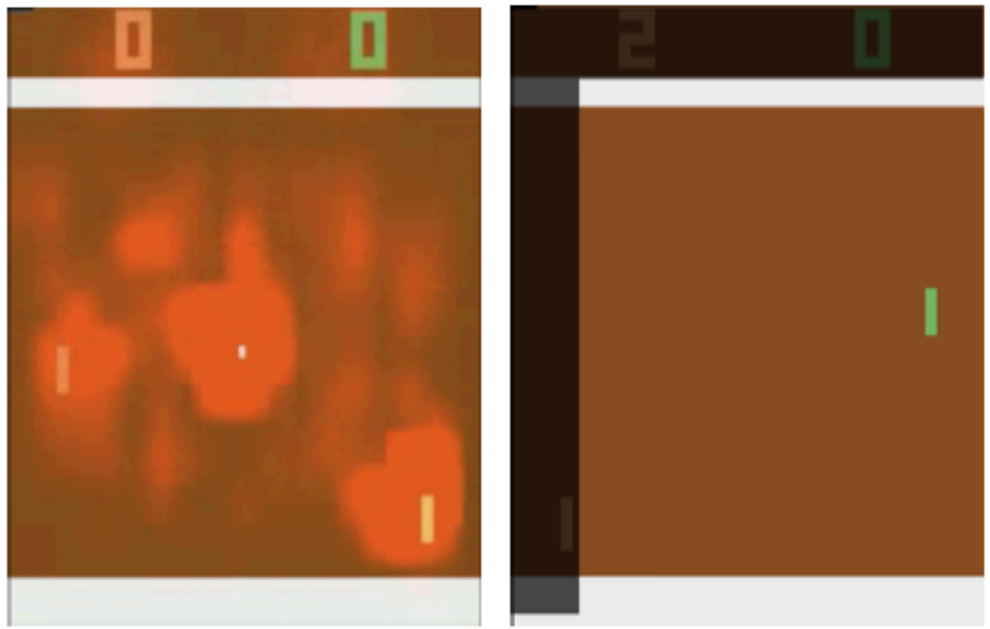
Such biases are often due to observation/action space mismatches.

Q2: How do we model it formally with causal inference tools?

Q3: How can we utilize such data for off-policy learning?

A Motivating Example - Confounded Pong

In the Pong game, score and opponent's paddle location shouldn't be a determining factor for a good human player,

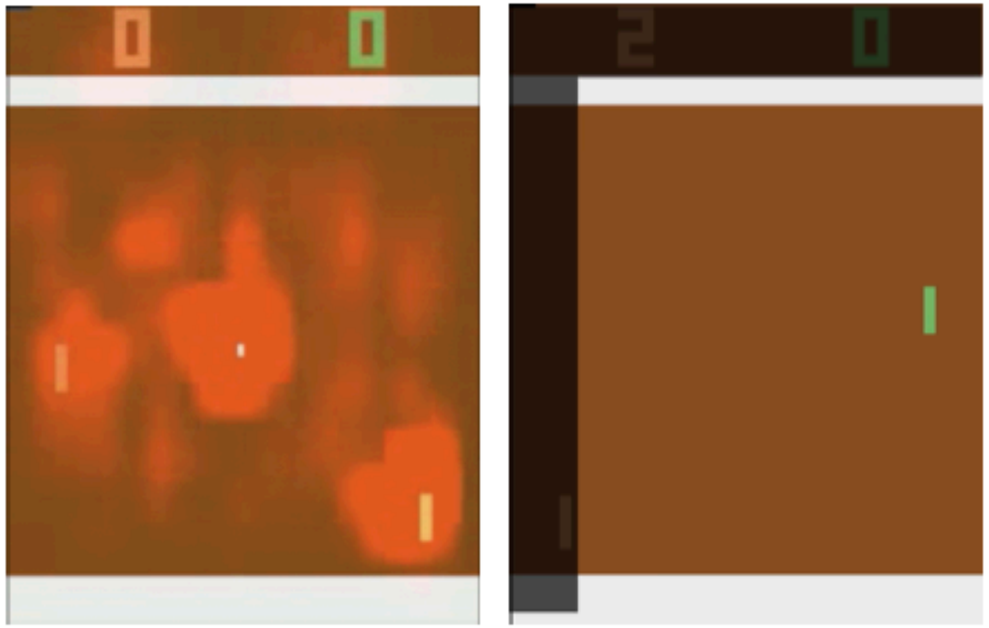


(a) (b)

(a) Saliency maps of a behavioral policy in Pong; (b) A confounded Pong game where score board and opponent's paddle location is masked.

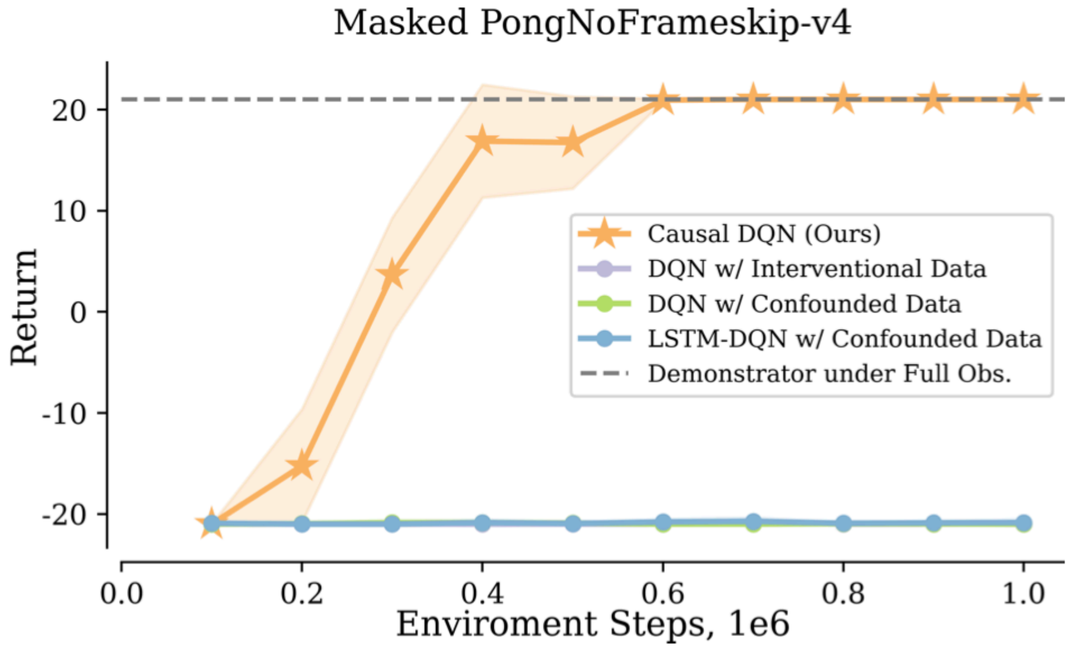
A Motivating Example - Confounded Pong

Surprisingly, all standard DQN variants failed to learn the correct policy in this confounded environment,



(a)

(b)

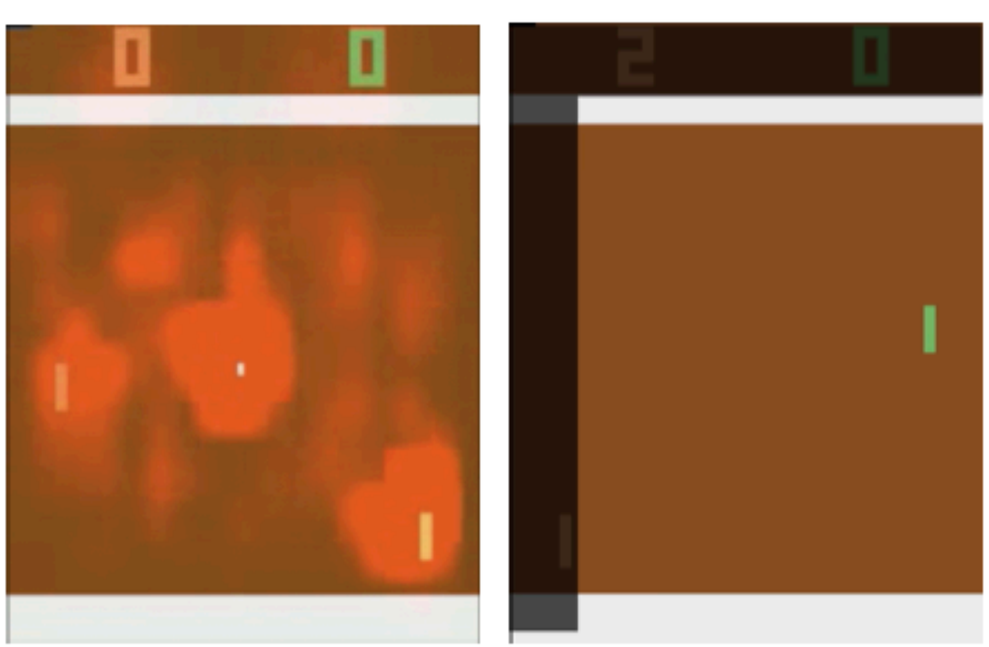


(c)

(a) Saliency maps of a behavioral policy in Pong; (b) A confounded Pong game where score board and opponent's paddle location is masked; (c) Evaluation performance of different DQN variants.

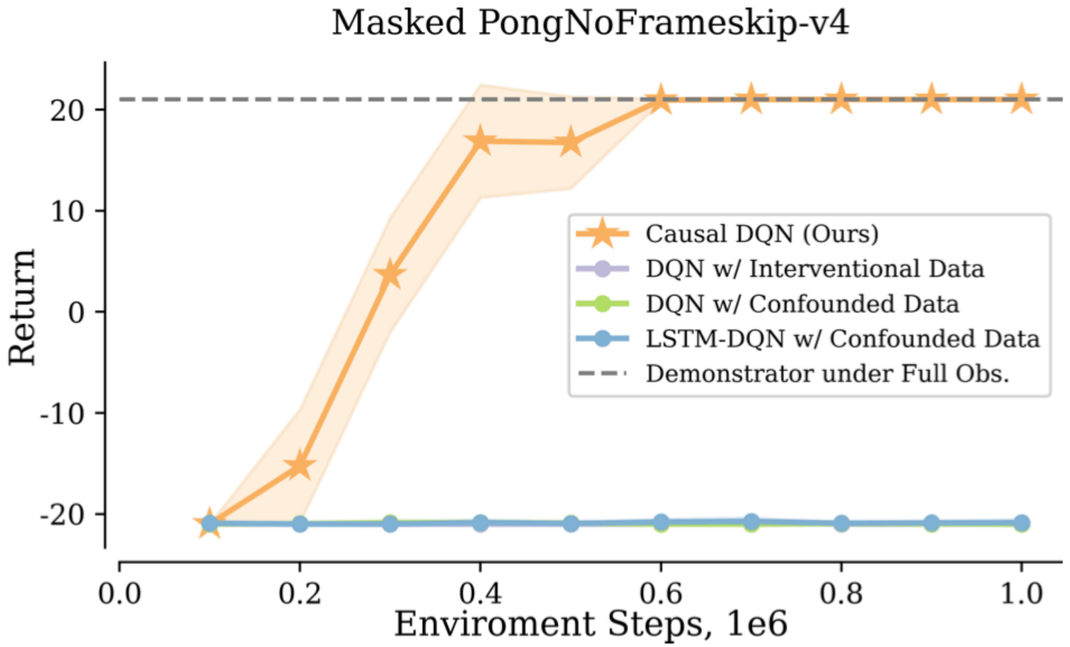
A Motivating Example - Confounded Pong

Except for `Causal DQN`, our solution to **off-policy DRL under unobserved confounders**!



(a)

(b)

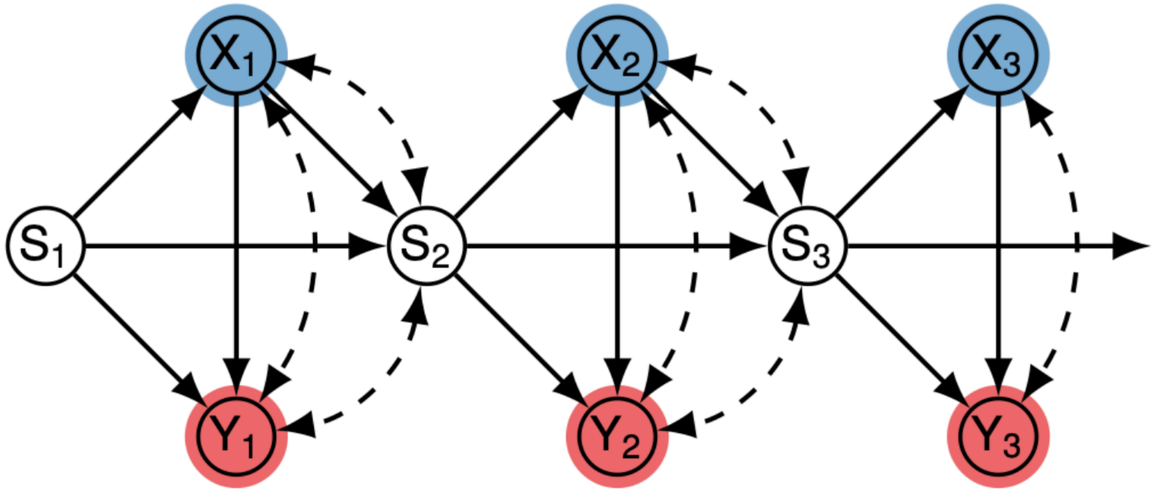


(c)

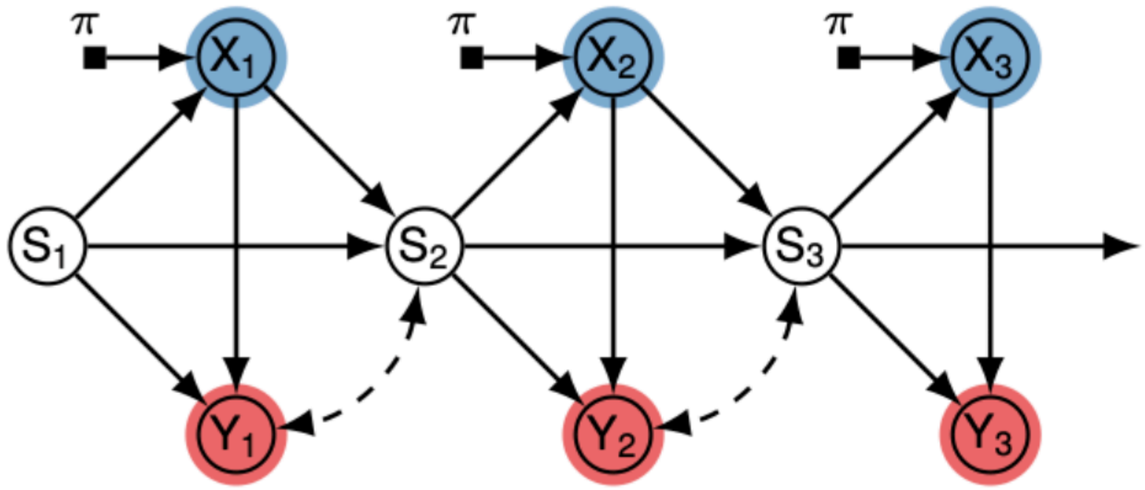
(a) Saliency maps of a behavioral policy in Pong; (b) A confounded Pong game where score board and opponent's paddle location is masked; (c) Evaluation performance of different DQN variants.

Confounded MDP — A Causal Diagram View

The off-policy dataset is confounded while the online agent is interventional, i.e., loses access to those unobserved features.



(a) CMDP



(b) CMDP-Online

Conservative Off-policy Learning via Causal Bellman Optimality Equation

Instead of directly learning unreliable Q-values from off-policy data, we use it only to lower bound the optimal Q-values,

Proposition (Causal Bellman Optimality Equation). For a CMDP environment $\mathscr{M}$ with reward $Y_t \in [a, b] \subset \mathbb{R}$, its optimal Q-value function satisfies, $Q^*(s, x) \geq \underline{Q_*}(s, x), \forall (s, x) \in \mathcal{S} \times \mathcal{X}$, where the lower bound $\underline{Q_*}(s, x)$ is given by,

$$\underline{Q_*}(s,x) = P(x \mid s)\left(\widetilde{\mathscr{R}}(s,x) + \mathbb{E}_{\widetilde{\mathscr{T}}}[\max_{x'} \underline{Q_*}(s',x')]\right) + P(\neg x \mid s)\left(a + \min_{s'} \max_{x'} \underline{Q_*}(s',x')\right)$$

where $P(x \mid s) = P(X_t = x \mid S_t = s)$ and $P(\neg x \mid s) = 1 - P(x \mid s)$; $\tilde{T}, \tilde{R}$ are estimated transition distribution and rewards from off-policy dataset, respectively.

Causal Deep Q-Learning

Algorithm 1 Causal Deep Q-Learning (`Causal-DQN`)

Initialize replay memory $\mathcal{D}$
Initialize action-value function $\underline{Q_*}(\cdot;\theta)$ with random weights θ
for episodes $= 1, \ldots, M$ **do**
    Sample initial state s_1
    for $t = 1, \ldots, T$ **do**
        Observe an action x_t taken by the demonstrator and subsequent reward y_t and state s_{t+1}
        Store transition (s_t, x_t, y_t, s_{t+1}) in $\mathcal{D}$
        Sample a minibatch of transitions $\{(s_i, x_i, y_i, s_{i+1})\}_{i=1}^{B}$ from $\mathcal{D}$
        Set value target $w_i(x)$ for every action $x \in \mathcal{X}$ w.r.t sample (s_i, x_i, y_i, s_{i+1}),

$$w_i(x) = \begin{cases} y_i + \gamma \max_{x'} \underline{Q_*}(s_{i+1}, x'; \theta) & \text{if } x = x_i \\ a + \gamma \min_{s'} \max_{x'} \underline{Q_*}(s', x'; \theta) & \text{if } x \neq x_i \end{cases} \tag{11}$$

        Perform a gradient descent step on $\sum_x \left(w_i(x) - \underline{Q_*}(s_i, x; \theta)\right)^2$ according to Eq. (10)
    end for
end for

Experiment Results

Our `Causal DQN` outperforms all baselines on 12 confounded Atari games even surpassing the demonstrator performance slightly.

Game	Demonstrator	Random	Interv. DQN	Conf. DQN	Conf. LSTM-DQN	`Causal-DQN` **(ours)**
Amidar	232.4	5.8	44.0	37.8	59.0	**282.6**
Asterix	3080.6	210.0	650.0	429.0	479.0	**2587.0**
Boxing	89.0	0.1	-0.62	-9.8	-6.9	**71.5**
Breakout	219.2	1.7	2.2	1.2	4.9	**131.2**
ChopperCommand	1280.0	811.0	1192.0	1076.0	1116.0	**1658.0**
Gopher	5480.6	257.6	288.8	752.0	485.6	**7327.2**
KungFuMaster	35400.0	258.5	12416.0	13674.0	6526.0	**44196.0**
MsPacman	2316.8	307.3	1191.6	881.8	787.4	**1747.6**
Pong	20.8	-20.7	-20.8	-20.8	-20.4	**21.0**
Qbert	4420.6	163.9	322.5	208.5	253.5	**4458.5**
RoadRunner	16560.6	11.5	1154.0	1168.0	484.0	**27414.0**
Seaquest	1412.4	68.4	237.2	281.6	164.8	**980.0**
Normalized Mean (↑)	1.00	0.00	0.13	0.10	0.09	**1.04**
Normalized Median (↑)	1.00	0.03	0.13	0.14	0.10	**1.01**
Normalized IQM (↑)	1.00	0.03	0.13	0.13	0.11	**1.02**

Summary

1. **Causal Foundations for Off-Policy RL:** Introduces the Causal Bellman Equation, extending traditional RL theory to handle confounded observational data and enabling more reliable policy learning.

2. **Causal-DQN:** A novel algorithm that learns effective policies even under unobserved confounding, outperforming standard DQN across twelve confounded Atari games.

3. **Beyond Benchmarks:** Unobserved confounding pervades in real-world RL, spanning robotics, RLHF for LLMs, and decision-making in critical domains like healthcare & self-driving, where scaling data without causal reasoning can amplify bias and misalignment.

4. **Vision for Causal RL:** We envision a future of confounding-robust, causally grounded agents capable of reasoning about interventions and consequences, learning not just what works, but why it works.