International Joint Conference on
Neural Networks (IJCNN) 2024
Yokohama, Japan

Towards Causal Reinforcement Learning

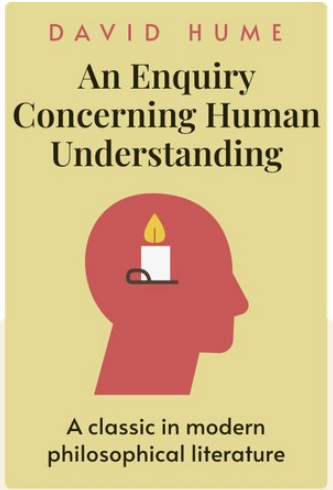
Empowering Agents with Causality

Zhihong Deng
Zhi-Hong.Deng@student.uts.edu.au

Jing Jiang
Jing.Jiang@uts.edu.au

Chengqi Zhang
Chengqi.Zhang@uts.edu.au

University of Technology Sydney (UTS)
Australian Artificial Intelligence Institute (AAII)



”All reasonings concerning matter of fact seem to be founded on the relation of cause and effect. By means of that relation alone we can go beyond the evidence of our memory and senses.“

-- David Hume, 1748

Captain Robert Falcon Scott

Imagine you went back to a century ago and joined Captain Scott's Antarctic expedition ...

Captain Robert Falcon Scott

Consuming rotten meat causes scurvy

Captain Robert Falcon Scott

Consuming rotten meat causes scurvy

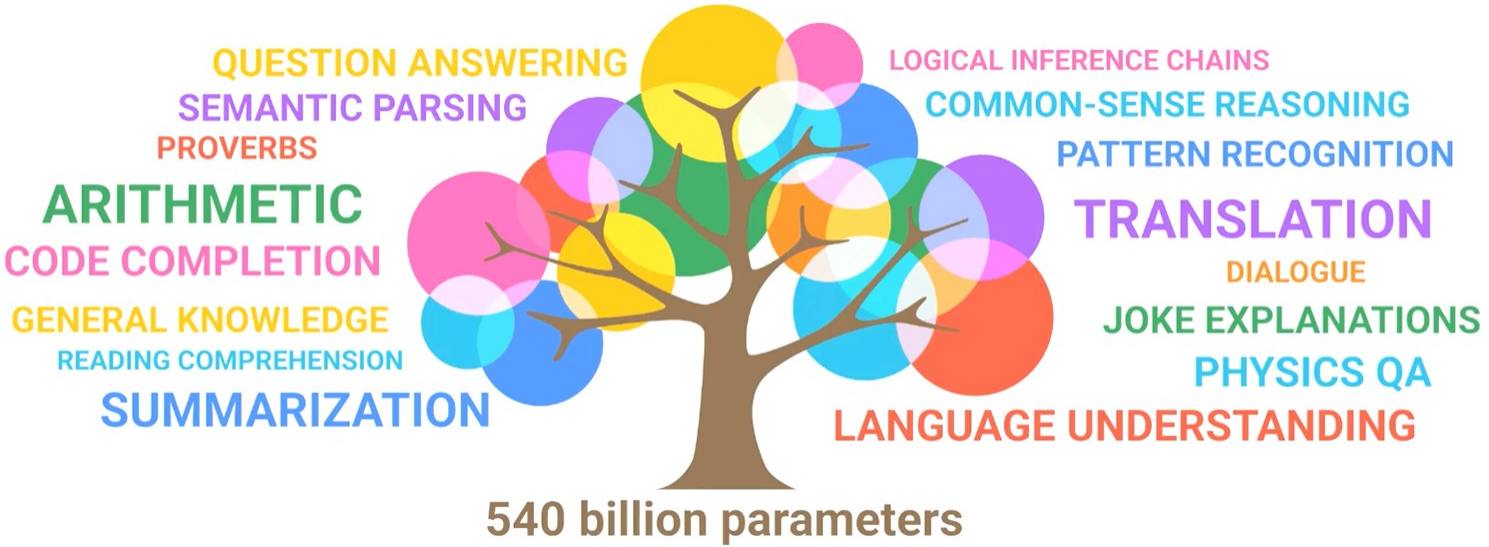
Captain Robert Falcon Scott

Consuming rotten meat scurvy

Lack of Vitamin C scurvy

The AI Scaling Law

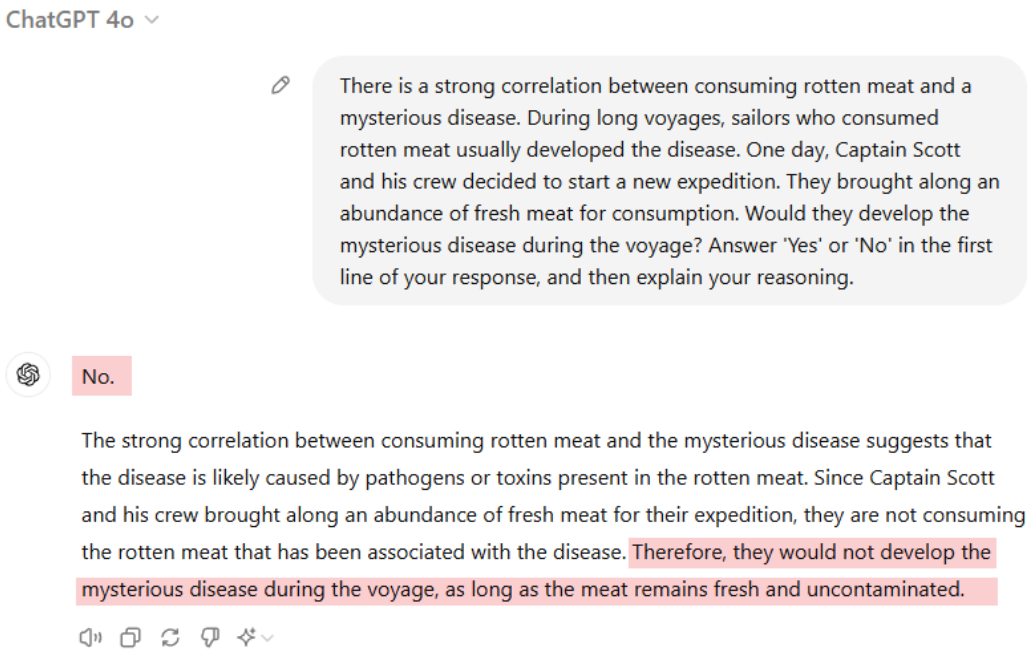
For more information, check "Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance", 2022



https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

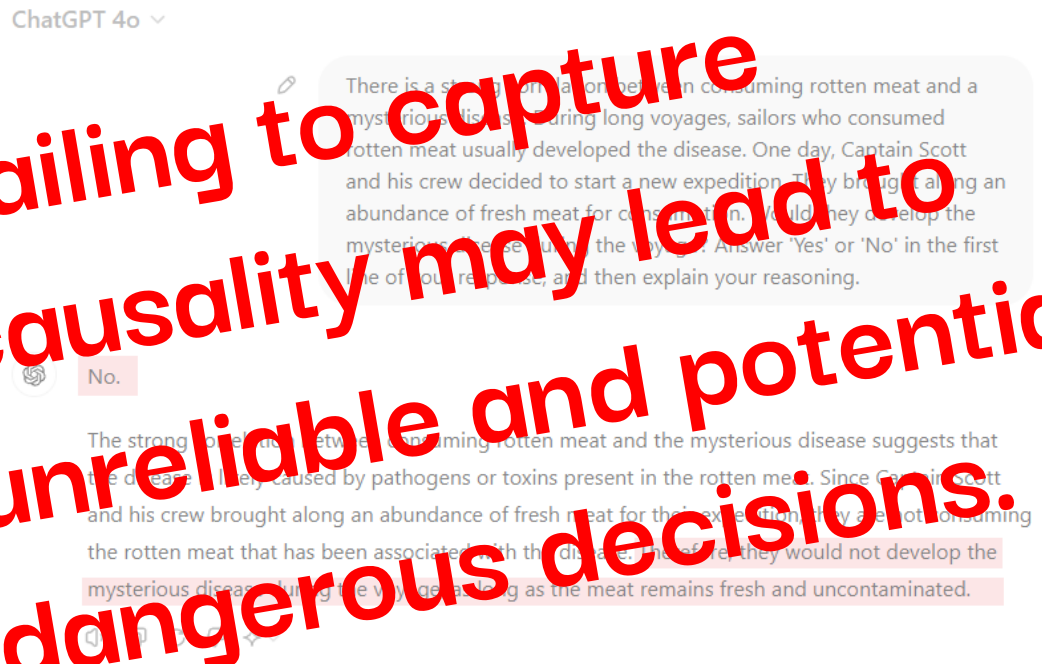
The PaLM pretraining dataset consists of a high-quality corpus of 780 billion tokens.

What Can Go Wrong?



A failure case of ChatGPT: ChatGPT, one of the best AI-powered chatbots trained using Reinforcement Learning from Human Feedback (RLHF) to align with human values.

What Can Go Wrong?



A failure case of ChatGPT: ChatGPT, one of the best AI-powered chatbots trained using Reinforcement Learning from Human Feedback (RLHF) to align with human values.

"This idea that we're going to just scale up the current large language models and eventually human-level AI will emerge–I don't believe this at all, not for one second."

Yann Lecun

Yann Lecun

"This idea that we're going to just scale up the current large language models and eventually human-level AI will emerge–I don't believe this at all, not for one second."

Judea Pearl

"Much of this data-centric history still haunts us today. We live in an era that presumes Big Data to be the solution to all our problems. Courses in 'data science' are proliferating in our universities, and jobs for 'data scientists' are lucrative in the companies that participate in the 'data economy.' But I hope with this book to convince you that data are profoundly dumb."

Yann Lecun

"This idea that we're going to just scale up the current large language models and eventually human-level AI will emerge–I don't believe this at all, not for one second."

Judea Pearl

"Much of this data-centric history still haunts us today. We live in an era that presumes Big Data to be the solution to all our problems. Courses in 'data science' are proliferating in our universities, and jobs for 'data scientists' are lucrative in the companies that participate in the 'data economy.' But I hope with this book to convince you that data are profoundly dumb. "

Yoshua Bengio

"One of the big debates these days is: What are the elements of higher-level cognition? Causality is one element of it. "

Goals of This Tutorial

1. Introduce basic concepts of causality and reinforcement learning.

Goals of This Tutorial

1. Introduce basic concepts of causality and reinforcement learning.

2. Discuss what causal reinforcement learning is and how it is different from traditional reinforcement learning.

Goals of This Tutorial

1. Introduce basic concepts of causality and reinforcement learning.

2. Discuss what causal reinforcement learning is and how it is different from traditional reinforcement learning.

3. Explore recent advances in causal reinforcement learning.

Tutorial Outline

Part 1

- [x] Introduction
- [ ] Causality Background
- [ ] Reinforcement Learning Background

Part 2

- [ ] Sample Efficiency
- [ ] Generalization Ability
- [ ] Reliability

Correlation vs. Causation

Consuming rotten meat ⟷ (correlates with) scurvy

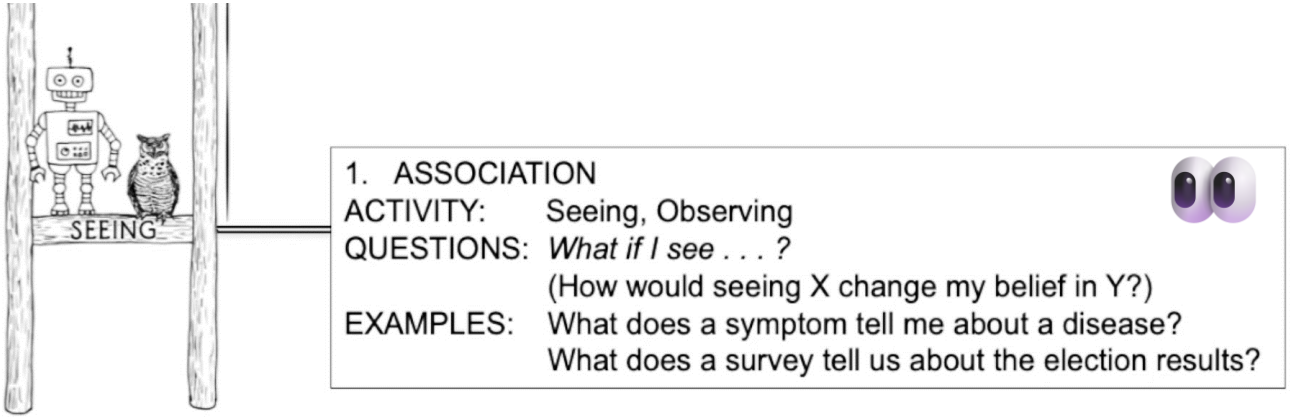
Lack of Vitamin C → (causes) scurvy

There is a strong correlation between rotten meat and scurvy in historical data, but eating rotten meat does not cause scurvy. The lack of vitamin C does.

Ladder of Causation

Judea Pearl

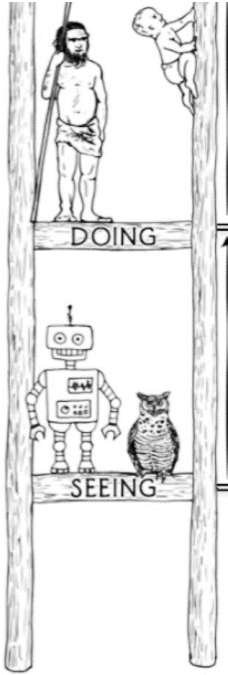
“ The Book of Why” , 2018



Ladder of Causation

Judea Pearl

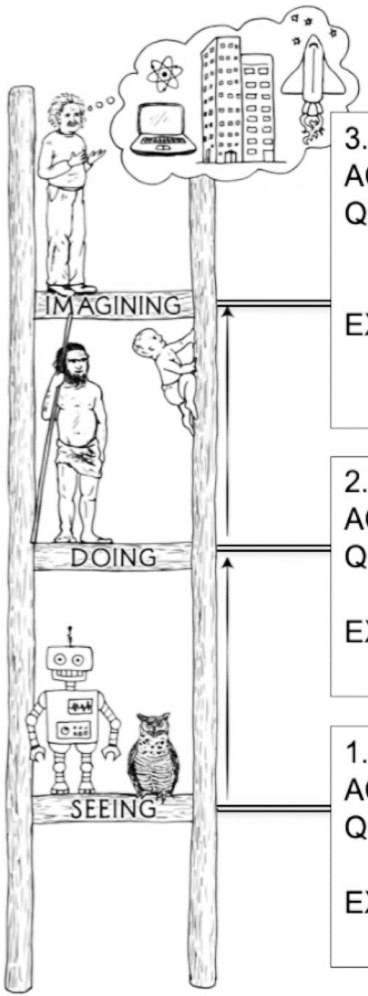
“ The Book of Why” , 2018



2. INTERVENTION
ACTIVITY: Doing, Intervening
QUESTIONS: *What if I do . . . ? How?*
(What would Y be if I do X?)
EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

1. ASSOCIATION
ACTIVITY: Seeing, Observing
QUESTIONS: *What if I see . . . ?*
(How would seeing X change my belief in Y?)
EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?

Ladder of Causation



3. COUNTERFACTUALS
ACTIVITY: Imagining, Retrospection, Understanding
QUESTIONS: *What if I had done . . . ? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)
EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked the last 2 years?

2. INTERVENTION
ACTIVITY: Doing, Intervening
QUESTIONS: *What if I do . . . ? How?*
(What would Y be if I do X?)
EXAMPLES: If I take aspirin, will my headache be cured? What if we ban cigarettes?

1. ASSOCIATION
ACTIVITY: Seeing, Observing
QUESTIONS: *What if I see . . . ?*
(How would seeing X change my belief in Y?)
EXAMPLES: What does a symptom tell me about a disease? What does a survey tell us about the election results?

Judea Pearl

" The Book of Why" , 2018

Structural Causal Model (SCM)

Definition. An SCM is represented by a quadruple $(\mathbf{V}, \mathbf{U}, \mathbf{F}, P(\mathbf{U}))$, where

- $\mathbf{V} = \{V_1, V_2, \cdots, V_m\}$ is a set of endogenous variables that are of interest in a research problem,
- $\mathbf{U} = \{U_1, U_2, \cdots, U_n\}$ is a set of exogenous variables that represent the source of stochasticity in the model and are determined by external factors that are generally unobservable,
- $\mathbf{F} = \{f_1, f_2, \cdots, f_m\}$ is a set of structural equations that assign values to each of the variables in $\mathbf{V}$ such that f_i maps $\mathbf{PA}(V_i) \cup U_i$ to V_i, where $\mathbf{PA}(V_i) \subseteq \mathbf{V} \setminus V_i$ and $U_i \subseteq \mathbf{U}$,
- $P(\mathbf{U})$ is the joint probability distribution of the exogenous variables in $\mathbf{U}$.

$$U_1 \perp\!\!\!\perp U_2 \perp\!\!\!\perp \cdots \perp\!\!\!\perp U_n$$

Structural Causal Model (SCM)

Structural Equations

$$f_X: X = U_X$$
$$f_Z: Z = a \cdot X + U_Z$$
$$f_Y: Y = b \cdot X + c \cdot Z + U_Y$$

An example of SCM with structural equations

$\mathrm{F} = \{f_X, f_Z, f_Y\}$

X: food consumption,
Z: intake of vitamin C,
Y: occurrence of scurvy

Structural Causal Model (SCM)

Endogenous Variables

$$f_X: X = U_X$$
$$f_Z: Z = a \cdot X + U_Z$$
$$f_Y: Y = b \cdot X + c \cdot Z + U_Y$$

An example of SCM with structural equations

$$\mathrm{F} = \{f_X, f_Z, f_Y\}$$

X: food consumption,
Z: intake of vitamin C,
Y: occurrence of scurvy

Structural Causal Model (SCM)

Exogenous Variables

$$f_X: X = U_X$$

$$f_Z: Z = a \cdot X + U_Z$$

$$f_Y: Y = b \cdot X + c \cdot Z + U_Y$$

An example of SCM with structural equations

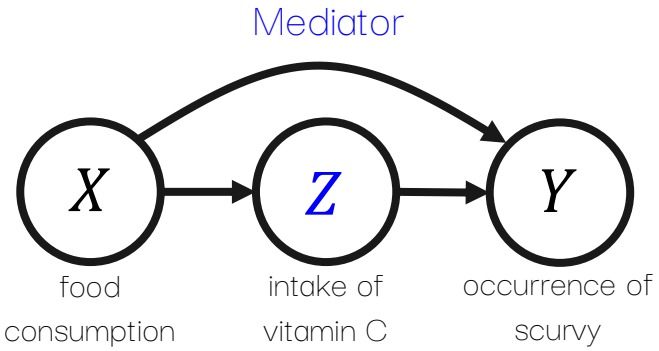
$\mathrm{F} = \{f_X, f_Z, f_Y\}$

X: food consumption,

Z: intake of vitamin C,

Y: occurrence of scurvy

Causal Graph



Causal Graph

$$f_X\colon\ X = U_X$$

$$f_Z\colon\ Z = a \cdot X + U_Z$$

$$f_Y\colon\ Y = b \cdot X + c \cdot Z + U_Y$$

An example of SCM with structural equations

$\mathrm{F} = \{f_X, f_Z, f_Y\}$

X: food consumption,
Z: intake of vitamin C,
Y: occurrence of scurvy

Observations vs. Interventions

Observations

Passively observe people with different food consumption.

Observations vs. Interventions

Observations

Q: What does consuming citrus fruits tell me about the possibility of getting scurvy?

$P(Y \mid X=$  $)$

Observations vs. Interventions

Observations

Q: What does consuming citrus fruits tell me about the possibility of getting scurvy?

$P(Y \mid X=$  $)$

Interventions

Actively force all sailors to consume fresh citrus fruits.

Observations vs. Interventions

Observations

Q: What does consuming citrus fruits tell me about the possibility of getting scurvy?

$P(Y \mid X=$  $)$

Interventions

Q: What if all sailors consume fresh citrus fruits, will they get scurvy?

$P(Y \mid \text{do}\ (X=$  $)\ )$

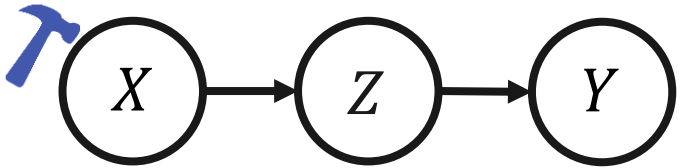
Interventions

New SCM

f_X: $X =$ 

f_Z: $Z = a \cdot X + U_Z$

f_Y: $Y = c \cdot Z + U_Y$



New Causal Graph

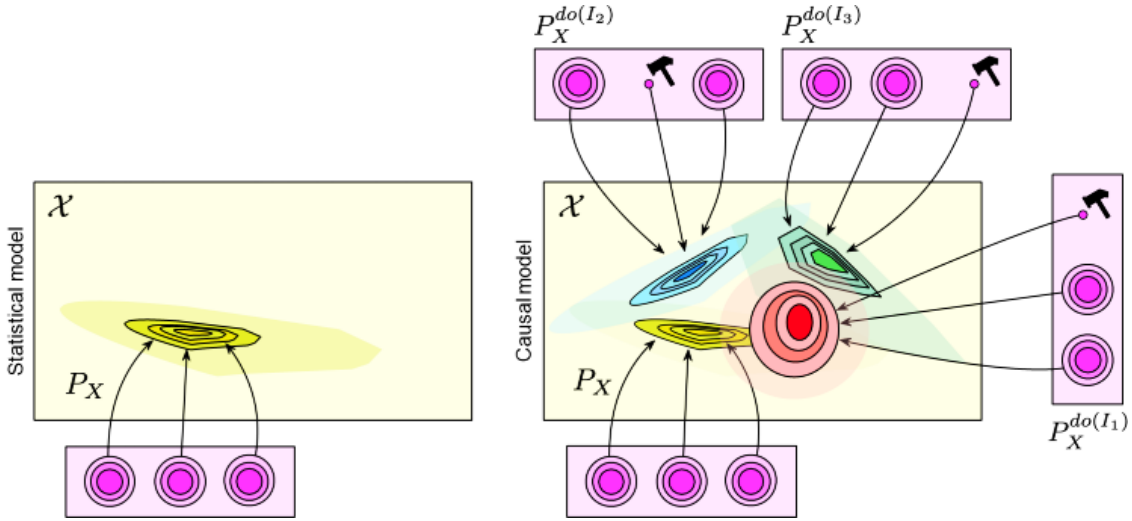
Interventions

Q: What if all sailors consume fresh citrus fruits, will they get scurvy?

$P(Y \mid \text{do}\ (X=$  $))$

For more information, check "Towards Causal Representation Learning", Proceedings of the IEEE 2021

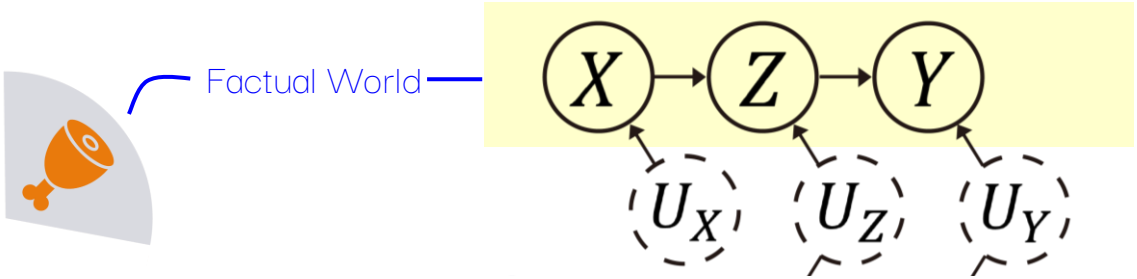
Statistical vs. Causal Model



Causal models are inherently more powerful than statistical model!

Difference between statistical (left) and causal models (right) on a given set of three variables. While a statistical model specifies a single probability distribution, a causal model represents a set of distributions, one for each possible intervention (indicated with a in the figure)

Counterfactuals



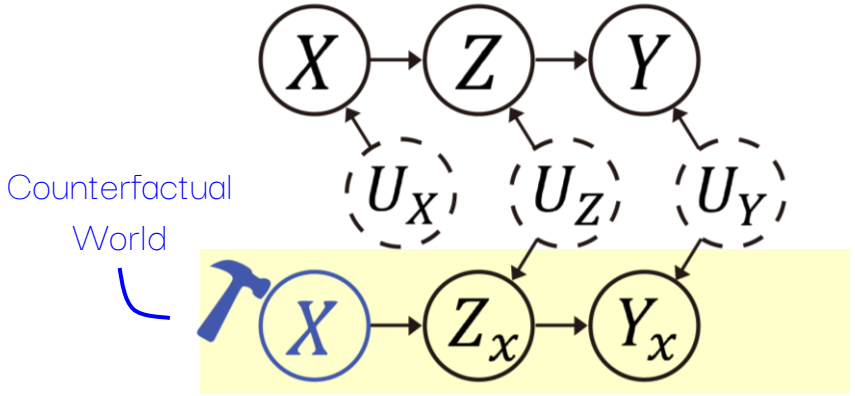
Observations

Passively observe people who consume rotten meat can also get scurvy.

Counterfactuals

Observations

Passively observe people who consume rotten meat can also get scurvy.



Counterfactuals

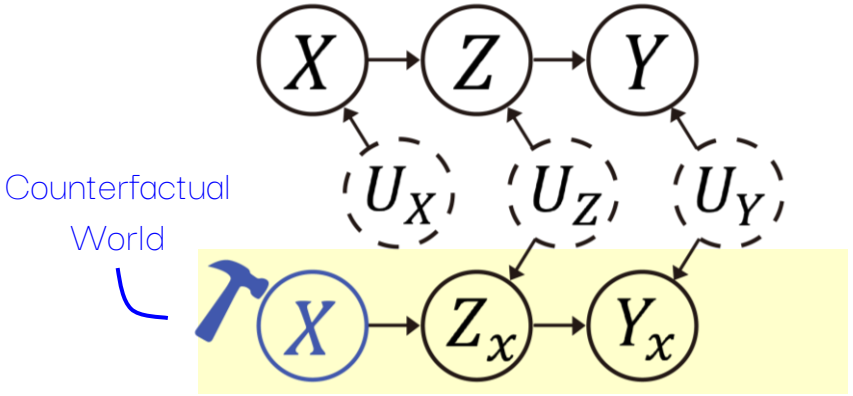
Imagine people who would have consumed rotten meat choosing to consume fresh citrus fruits instead.

Counterfactuals

Observations

Passively observe people who consume rotten meat can also get scurvy.

$P(Y \mid X=$ $)$

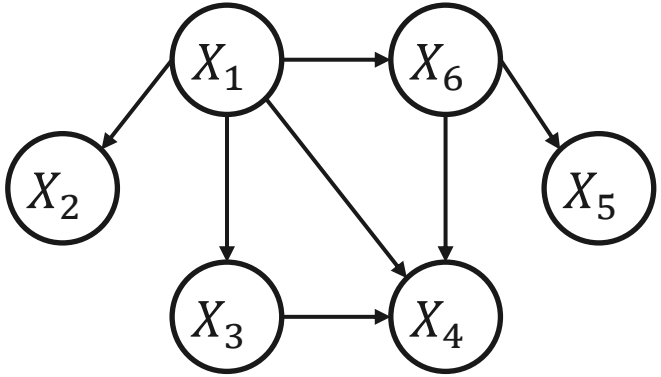


Counterfactuals

Q: Considering that they consumed rotten meat in reality, would sailors have been protected from scurvy if they had consumed enough citrus fruit?

$P(Y_{X=}$  $|X =$  $, Y =$  $)$

Non-Causal vs. Causal Factorization



Non-Causal Factorization

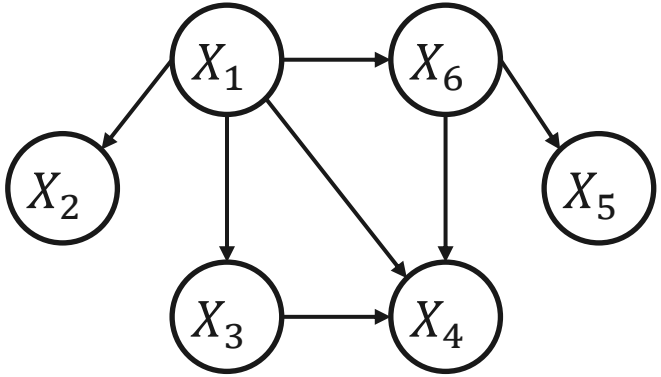
e.g.,

$$P(\boldsymbol{X}) = \prod_{i=1}^{N} P(X_i|X_1, \dots, X_{i-1})$$

$$= P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1, X_2) \cdot P(X_4|X_1, X_2, X_3)$$

$$\cdot P(X_5|X_1, X_2, X_3, X_4) \cdot P(X_6|X_1, X_2, X_3, X_4, X_5)$$

Non-Causal vs. Causal Factorization



Non-Causal Factorization

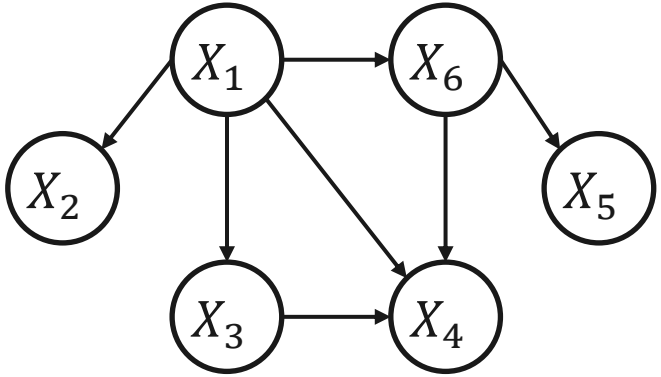
e.g.,

$$P(\boldsymbol{X}) = \prod_{i=1}^{N} P(X_i|X_1, \dots, X_{i-1})$$
$$= P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1, X_2) \cdot P(X_4|X_1, X_2, X_3)$$
$$\cdot P(X_5|X_1, X_2, X_3, X_4) \cdot P(X_6|X_1, X_2, X_3, X_4, X_5)$$

Causal Factorization

$$P(\boldsymbol{X}) = \prod_{i=1}^{N} P(X_i|\mathrm{PA}(X_i))$$
$$= P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1) \cdot P(X_4|X_1, X_3, X_6)$$
$$\cdot P(X_5|X_6) \cdot P(X_6|X_1)$$

Non-Causal vs. Causal Factorization



Non-Causal Factorization

e.g.,

$$P(\boldsymbol{X}) = \prod_{i=1}^{N} P(X_i|X_1, \dots, X_{i-1})$$

$$= P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1, X_2) \cdot P(X_4|X_1, X_2, X_3)$$

$$\cdot P(X_5|X_1, X_2, X_3, X_4) \cdot P(X_6|X_1, X_2, X_3, X_4, X_5)$$

Causal Factorization

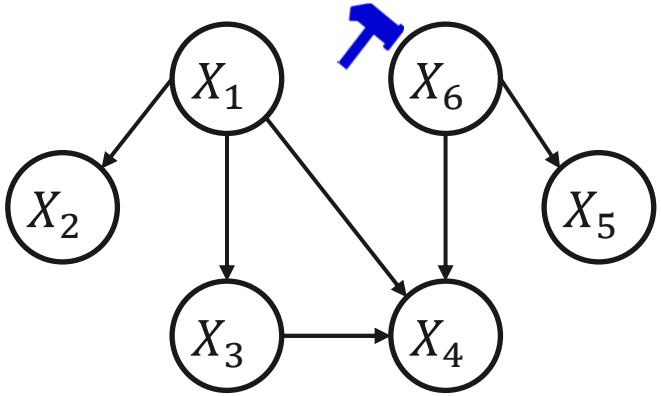
$$P(\boldsymbol{X}) = \prod_{i=1}^{N} P(X_i|\mathrm{PA}(X_i))$$

$$= P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1) \cdot P(X_4|X_1, X_3, X_6)$$

$$\cdot P(X_5|X_6) \cdot P(X_6|X_1)$$

Causal factorization yields practical computational advantages during inference.

Non-Causal vs. Causal Factorization



Non-Causal Factorization

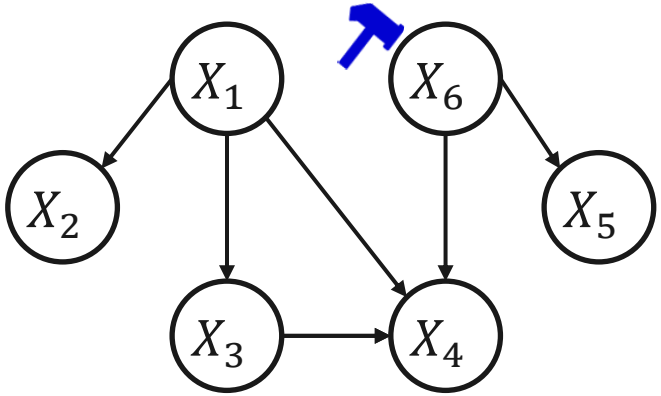
e.g.,

$$P(\boldsymbol{X}) = \prod_{i=1}^{N} P(X_i|X_1, \dots, X_{i-1})$$
$$= P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1, X_2) \cdot P(X_4|X_1, X_2, X_3)$$
$$\cdot P(X_5|X_1, X_2, X_3, X_4) \cdot P(X_6|X_1, X_2, X_3, X_4, X_5)$$

Causal Factorization

$$P(\boldsymbol{X}) = \prod_{i=1}^{N} P(X_i|\mathrm{PA}(X_i))$$
$$= P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1) \cdot P(X_4|X_1, X_3, X_6)$$
$$\cdot P(X_5|X_6) \cdot P(X_6|X_1)$$

Non-Causal vs. Causal Factorization



Non-Causal Factorization

e.g.,

$$P(\boldsymbol{X}) = \prod_{i=1}^{N} P(X_i|X_1, \dots, X_{i-1})$$

$$= P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1, X_2) \cdot P(X_4|X_1, X_2, X_3)$$

$$\cdot P(X_5|X_1, X_2, X_3, X_4) \cdot P(X_6|X_1, X_2, X_3, X_4, X_5)$$

Causal Factorization

$$P(\boldsymbol{X}) = \prod_{i=1}^{N} P(X_i|\mathrm{PA}(X_i))$$

$$= P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1) \cdot P(X_4|X_1, X_3, X_6)$$

$$\cdot P(X_5|X_6) \cdot P(X_6|X_1)$$

Causal factorization is more robust to variations.

For more information, check "Towards Causal Representation Learning", Proceedings of the IEEE 2021

Independent Causal Mechanism

Independent Causal Mechanisms (ICM) Principle. *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.*

Applied to the causal factorization, this principle tells us that the factors should be independent in the sense that

(a) changing (or performing an intervention upon) one mechanism $P(X_i|\mathrm{PA}(X_i))$ does not change any of the other mechanisms $P\left(X_j\middle|\mathrm{PA}\left(X_j\right)\right)$ $(i \neq j)$.

(b) Knowing some other mechanisms $P\left(X_j\middle|\mathrm{PA}\left(X_j\right)\right)$ $(i \neq j)$ does not give us information about a mechanism $P(X_i|\mathrm{PA}(X_i))$.

Summary (so far)

- Pearl's Ladder of Causation is a conceptual framework that categorizes levels of causal relationships, spanning from association, intervention, and counterfactuals.
- Structural Causal Model (SCM) provides a powerful framework for representing and analyzing causal relationships, offering a systematic approach to climb the Ladder of Causation.
- Interventions refer to actively manipulating variables, leading to a set of possible joint distributions, but a statistical model typically captures only one of them.
- Counterfactuals allow us to envision the outcomes of different decisions through the lens of imagination and retrospection.
- Causal Factorization decompose a joint distribution into independent causal mechanisms, yielding practical computational advantages and is robust to variations.

Summary (so far)

- Pearl's Ladder of Causation is a conceptual framework that categorizes levels of causal relationships, spanning from association, intervention, and counterfactuals.
- Structural Causal Model (SCM) provides a powerful framework for representing and analyzing causal relationships, offering a systematic approach to climb the Ladder of Causation.
- Interventions refer to actively manipulating variables, leading to a set of possible joint distributions, but a statistical model typically captures only one of them.
- Counterfactuals allow us to envision the outcomes of different decisions through the lens of imagination and retrospection.
- Causal Factorization decompose a joint distribution into independent causal mechanisms, yielding practical computational advantages and is robust to variations.

Summary (so far)

- Pearl's Ladder of Causation is a conceptual framework that categorizes levels of causal relationships, spanning from association, intervention, and counterfactuals.
- Structural Causal Model (SCM) provides a powerful framework for representing and analyzing causal relationships, offering a systematic approach to climb the Ladder of Causation.
- Interventions refer to actively manipulating variables, leading to a set of possible joint distributions, but a statistical model typically captures only one of them.
- Counterfactuals allow us to envision the outcomes of different decisions through the lens of imagination and retrospection.
- Causal Factorization decompose a joint distribution into independent causal mechanisms, yielding practical computational advantages and is robust to variations.

Summary (so far)

- Pearl's Ladder of Causation is a conceptual framework that categorizes levels of causal relationships, spanning from association, intervention, and counterfactuals.
- Structural Causal Model (SCM) provides a powerful framework for representing and analyzing causal relationships, offering a systematic approach to climb the Ladder of Causation.
- Interventions refer to actively manipulating variables, leading to a set of possible joint distributions, but a statistical model typically captures only one of them.
- Counterfactuals allow us to envision the outcomes of different decisions through the lens of imagination and retrospection.
- Causal Factorization decompose a joint distribution into independent causal mechanisms, yielding practical computational advantages and is robust to variations.

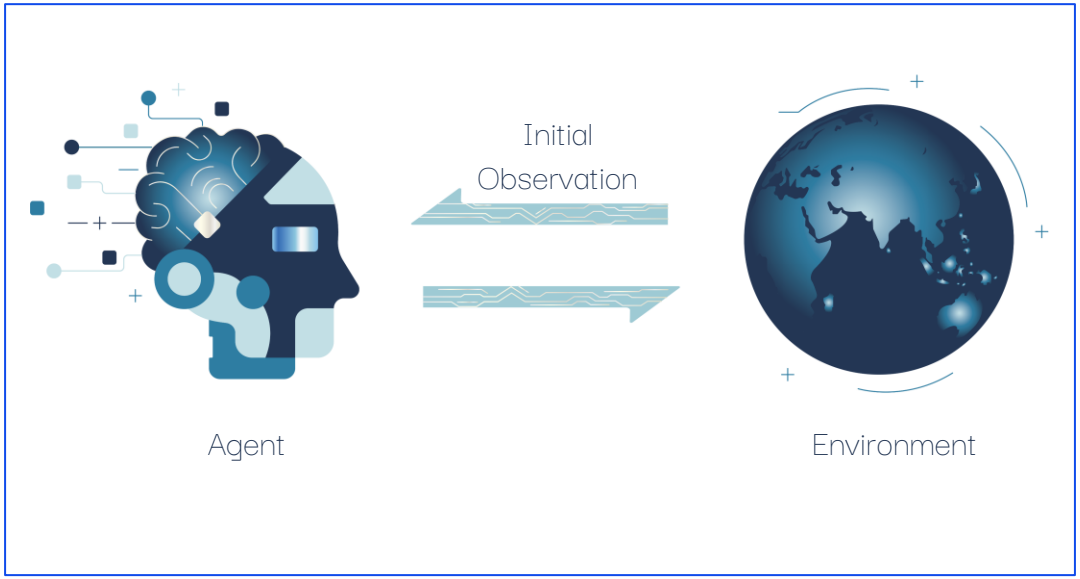
Summary (so far)

- Pearl's Ladder of Causation is a conceptual framework that categorizes levels of causal relationships, spanning from association, intervention, and counterfactuals.
- Structural Causal Model (SCM) provides a powerful framework for representing and analyzing causal relationships, offering a systematic approach to climb the Ladder of Causation.
- Interventions refer to actively manipulating variables, leading to a set of possible joint distributions, but a statistical model typically captures only one of them.
- Counterfactuals allow us to envision the outcomes of different decisions through the lens of imagination and retrospection.
- Causal Factorization decompose a joint distribution into independent causal mechanisms, yielding practical computational advantages and is robust to variations.

Summary (so far)

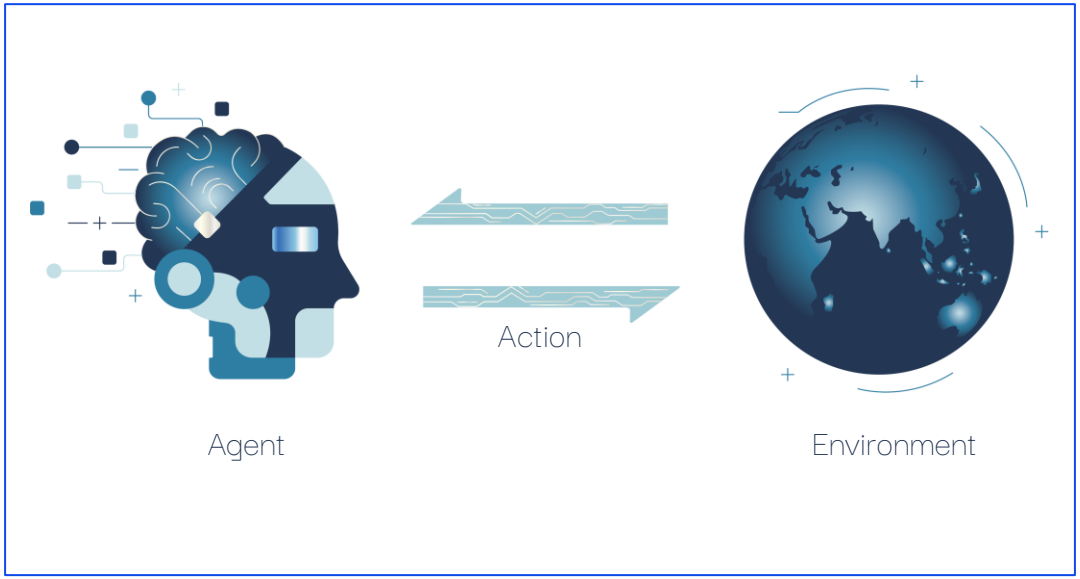
- Pearl's Ladder of Causation is a conceptual framework that categorizes levels of causal relationships, spanning from association, intervention, and counterfactuals.
- Structural Causal Model (SCM) provides a powerful framework for representing and analyzing causal relationships, offering a systematic approach to climb the Ladder of Causation.
- Interventions refer to actively manipulating variables, leading to a set of possible joint distributions, but a statistical model typically captures only one of them.
- Counterfactuals allow us to envision the outcomes of different decisions through the lens of imagination and retrospection.
- Causal Factorization decompose a joint distribution into independent causal mechanisms, yielding practical computational advantages and is robust to variations.

Reinforcement Learning (RL)



Receive an initial observation

Reinforcement Learning (RL)



Make a decision

Reinforcement Learning (RL)

Receive and learning from feedback

Reinforcement Learning (RL)

The agent-environment feedback loop

Markov Decision Process (MDP)

Definition (Markov decision process). An MDP $\mathcal{M}$ is specified by a tuple $\{\mathcal{S}, \mathcal{A}, P, R, \mu_0, \gamma\}$, where

- $\mathcal{S}$ denotes the state space and $\mathcal{A}$ denotes the action space,
- $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is the transition probability function that yields the probability of transitioning into the next states s_{t+1} after taking an action a_t at the current state s_t,
- $R : \mathcal{S} \times \mathcal{A} \rightarrow$ R is the reward function that assigns the immediate reward for taking an action a_t at state s_t,
- $\mu_0 : \mathcal{S} \rightarrow [0, 1]$ is the probability distribution that specifies the generation of the initial state, and
- $\gamma \in [0, 1]$ denotes the discount factor that accounts for how much future events lose their value as time passes.

RL aims to maximize the expected cumulative reward rather than the immediate one.

$$G_t = R_t + \gamma R_{t+1} + \ldots + \gamma^T R_{t+T}$$

R_t

Markov Decision Process (MDP)

Definition (Markov decision process). An MDP $\mathcal{M}$ is specified by a tuple $\{\mathcal{S}, \mathcal{A}, P, R, \mu_0, \gamma\}$

In essence, we want the agent to maximize

$$\mathrm{E}[G_t \mid S_t, \mathrm{do}(A_t = a_t)],$$

the expected cumulative reward across a sequence of interventions.

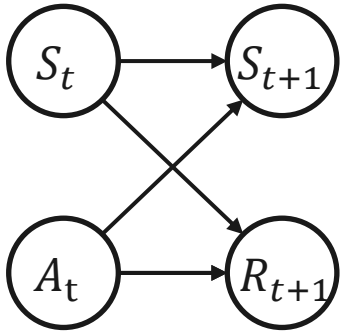
RL aims to maximize the expected cumulative reward rather than the immediate one.

Markov Decision Process (MDP)

Definition (Markov decision process). An MDP $\mathcal{M}$ is specified by a tuple $\{\mathcal{S}, \mathcal{A}, P, R, \mu_0, \gamma\}$

We can always cast an MDP into an SCM without imposing extra constraints.

- The state, action, and reward at each step correspond to endogenous variables.
- The state transition and reward functions are casted into structural equations $\mathcal{F}$ in the SCM, represented by deterministic functions with independent exogenous variables.



Causal Graph

Policy

Definition (Policy)**.** A policy is defined as the probability distribution of actions at a give state:

$$\pi(A_t = a | S_t = s), \forall\, S_t \in \mathcal{S},$$

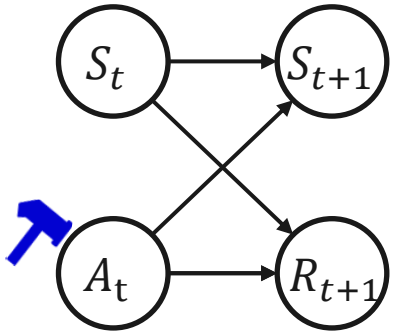
where $A_t \in \mathcal{A}(s)$ is the state specific action space.

Policy

Definition (Policy)**.** A policy is defined as the probability distribution of actions at a give state:

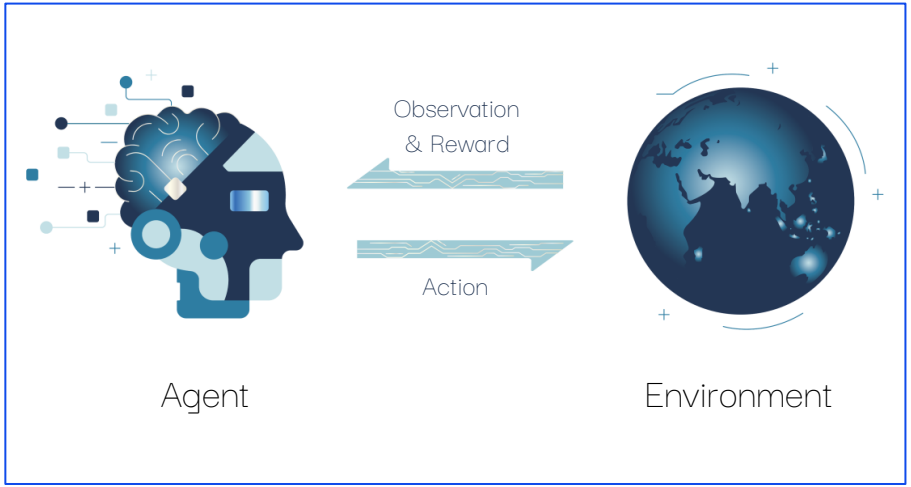
$$\pi(A_t = a|S_t = s), \forall\, S_t \in \mathcal{S},$$

where $A_t \in \mathcal{A}(s)$ is the state specific action space.

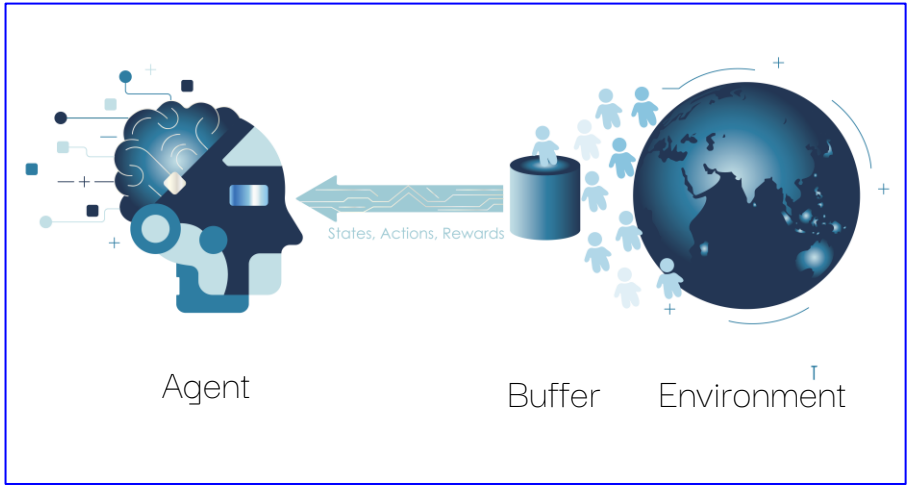


A policy π performs a soft intervention that preserves the dependency of the action on the state

Categorizing RL Methods

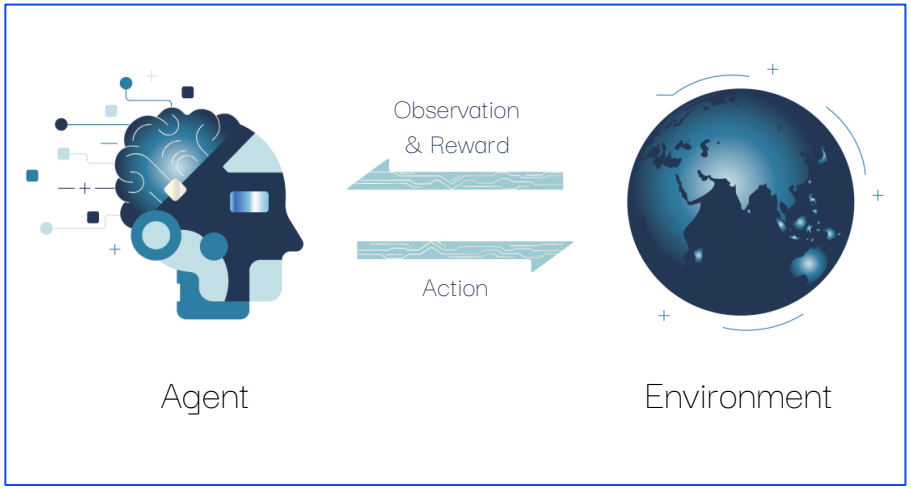


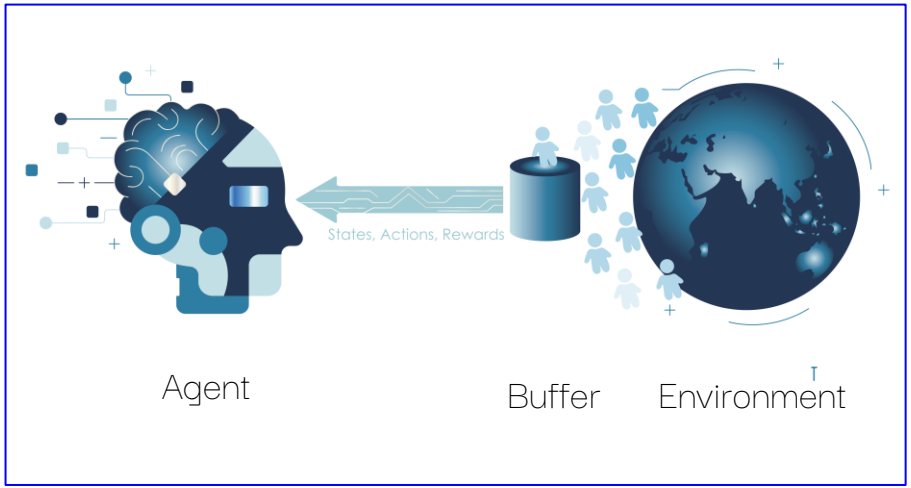
Online Reinforcement Learning



Offline Reinforcement Learning

Categorizing RL Methods





Online Reinforcement Learning

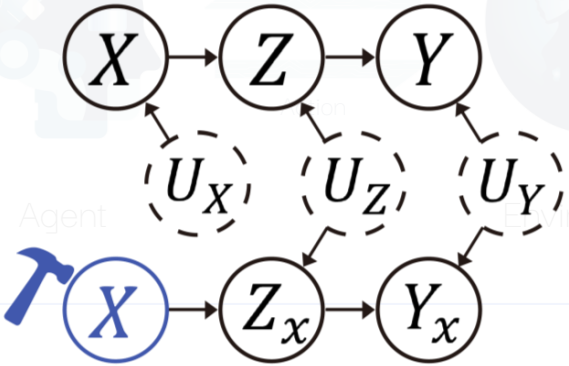
The agent can actively intervene in the environment.

Offline Reinforcement Learning

The agent can only passively observe the outcomes.

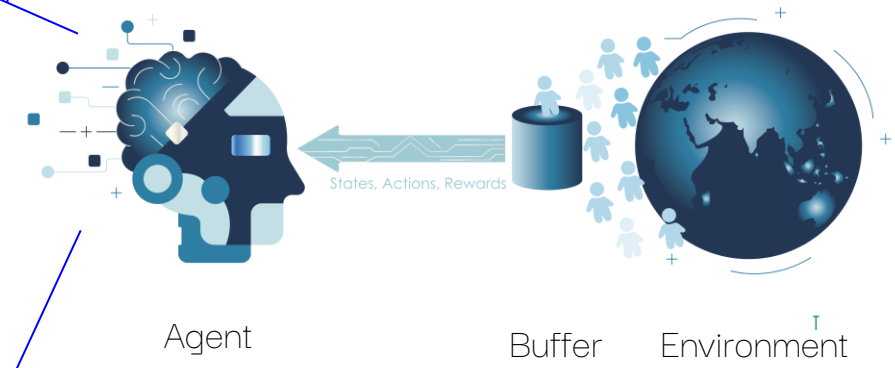
Categorizing RL Methods

Q: Considering that they consumed rotten meat in reality, would sailors have been protected from scurvy if they had consumed enough citrus fruit?



Counterfactual World

$P(Y_{X=}$  $|X =$  $, Y =$  $)$



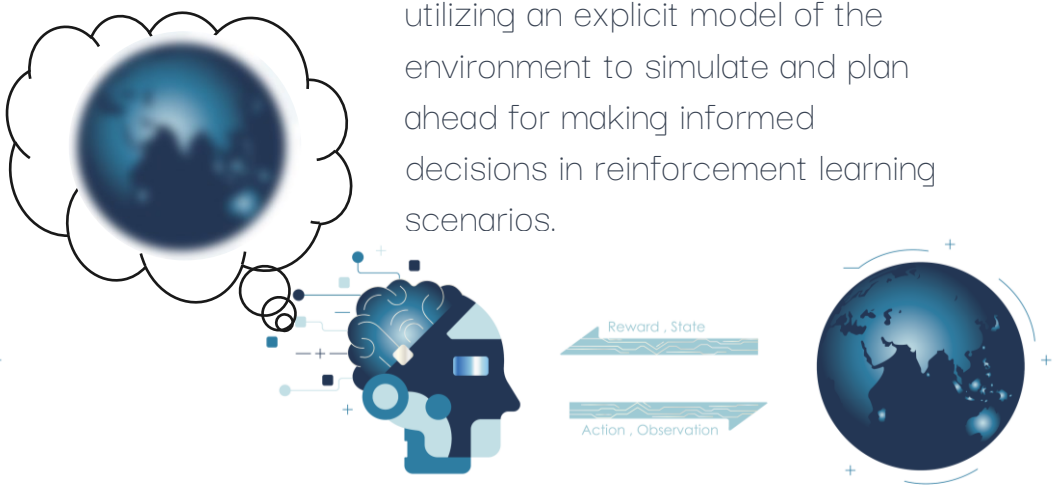
Offline Reinforcement Learning

The agent can only passively observe the outcomes.

Categorizing RL Methods

Model-free methods involve learning optimal policies or value functions directly from interaction with the environment without explicitly building a model of the environment's dynamics.

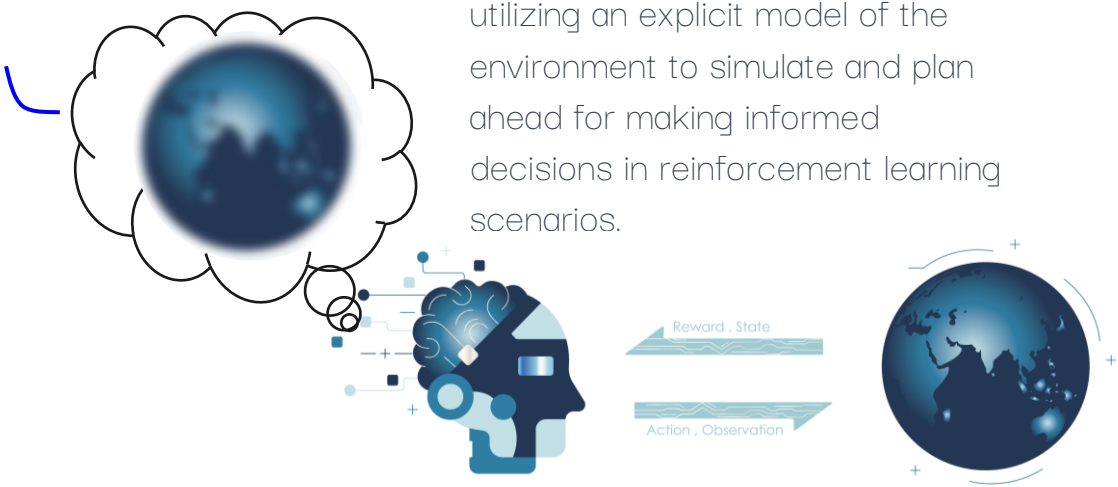
Model-based methods, on the other hand, revolve around creating and utilizing an explicit model of the environment to simulate and plan ahead for making informed decisions in reinforcement learning scenarios.



Categorizing RL Methods

Yann Lecun

The ability to learn "world models" – internal models of how the world works – may be the key to build human-level AI.



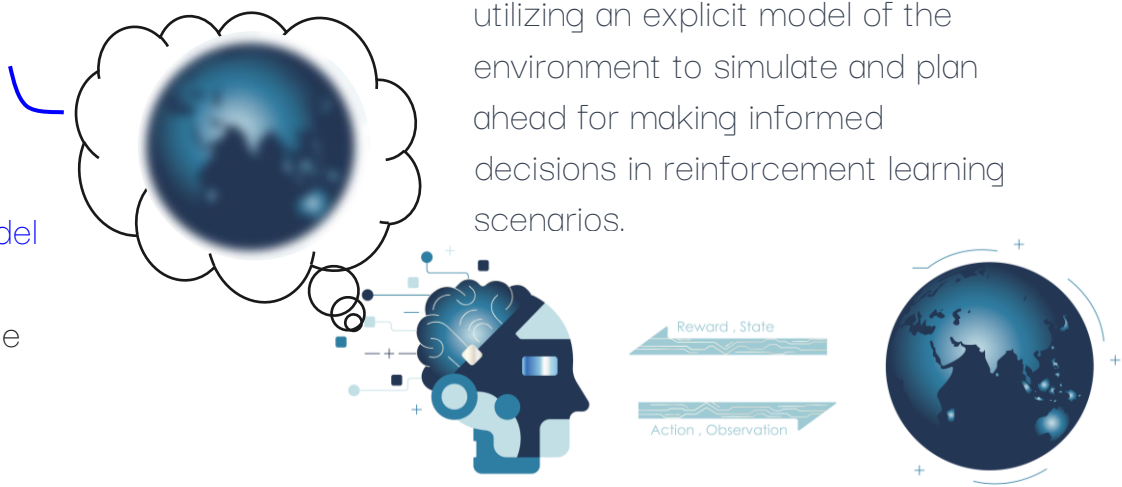
Model-based methods, on the other hand, revolve around creating and utilizing an explicit model of the environment to simulate and plan ahead for making informed decisions in reinforcement learning scenarios.

Categorizing RL Methods

Yann Lecun

The ability to learn "world models" – internal models of how the world works – may be the key to build human-level AI.

Q: How to construct an internal causal model that describes the causal relationships between variables (concepts) governing the data generation process in our world?



Model-based methods, on the other hand, revolve around creating and utilizing an explicit model of the environment to simulate and plan ahead for making informed decisions in reinforcement learning scenarios.

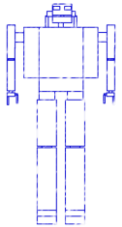
Causal Reinforcement Learning

Causal Reinforcement Learning

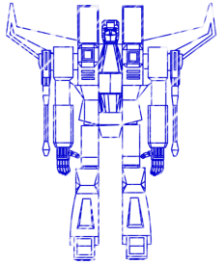
Definition (Causal Reinforcement Learning)**.** Causal RL is an umbrella term for RL approaches that incorporate additional assumptions or prior knowledge to analyze and understand the causal mechanisms underlying actions and their consequences, enabling agents to make more informed and effective decisions.

Causal Reinforcement Learning

Definition (Causal Reinforcement Learning)**.** Causal RL is an umbrella term for RL approaches that incorporate additional assumptions or prior knowledge to analyze and understand the causal mechanisms underlying actions and their consequences, enabling agents to make more informed and effective decisions.



Traditional RL methods focus on learning the optimal policies through interactions with the environment, without explicitly considering the causal relationships between actions and outcomes.

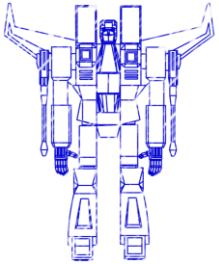


Causal RL, in contrast, go beyond the traditional framework by incorporating additional assumptions or prior knowledge about causality, empowering agents with a deeper understanding of the underlying dynamics of the world.

Causal Reinforcement Learning

Definition (Causal Reinforcement Learning)**.** Causal RL is an umbrella term for RL approaches that incorporate additional assumptions or prior knowledge to analyze and understand the causal mechanisms underlying actions and their consequences, enabling agents to make more informed and effective decisions.

Go beyond the evidence of memory and senses!



Causal RL, in contrast, go beyond the traditional framework by incorporating additional assumptions or prior knowledge about causality, empowering agents with a deeper understanding of the underlying dynamics of the world.

Summary

- Reinforcement Learning (RL) focuses on sequential decision-making problems, where an agent intervenes in an environment with the goal of maximizing cumulative rewards.
- A Markov Decision Process (MDP) describes the dynamics of the environment during interaction, and it can also be represented as an SCM.
- A policy guides an agent's decision-making by mapping states to appropriate actions.
- RL methods can be categorized in various ways, such as online vs. offline and model-free vs. model-based methods.
- Causal RL aims to incorporate causality into RL, enabling agents to understand the world better and make more informed decisions.

Summary

- Reinforcement Learning (RL) focuses on sequential decision-making problems, where an agent intervenes in an environment with the goal of maximizing cumulative rewards.
- A Markov Decision Process (MDP) describes the dynamics of the environment during interaction, and it can also be represented as an SCM.
- A policy guides an agent's decision-making by mapping states to appropriate actions.
- RL methods can be categorized in various ways, such as online vs. offline and model-free vs. model-based methods.
- Causal RL aims to incorporate causality into RL, enabling agents to understand the world better and make more informed decisions.

Summary

- Reinforcement Learning (RL) focuses on sequential decision-making problems, where an agent intervenes in an environment with the goal of maximizing cumulative rewards.
- A Markov Decision Process (MDP) describes the dynamics of the environment during interaction, and it can also be represented as an SCM.
- A policy guides an agent's decision-making by mapping states to appropriate actions.
- RL methods can be categorized in various ways, such as online vs. offline and model-free vs. model-based methods.
- Causal RL aims to incorporate causality into RL, enabling agents to understand the world better and make more informed decisions.

Summary

- Reinforcement Learning (RL) focuses on sequential decision-making problems, where an agent intervenes in an environment with the goal of maximizing cumulative rewards.
- A Markov Decision Process (MDP) describes the dynamics of the environment during interaction, and it can also be represented as an SCM.
- A policy guides an agent's decision-making by mapping states to appropriate actions.
- RL methods can be categorized in various ways, such as online vs. offline and model-free vs. model-based methods.
- Causal RL aims to incorporate causality into RL, enabling agents to understand the world better and make more informed decisions.

Summary

- Reinforcement Learning (RL) focuses on sequential decision-making problems, where an agent intervenes in an environment with the goal of maximizing cumulative rewards.
- A Markov Decision Process (MDP) describes the dynamics of the environment during interaction, and it can also be represented as an SCM.
- A policy guides an agent's decision-making by mapping states to appropriate actions.
- RL methods can be categorized in various ways, such as online vs. offline and model-free vs. model-based methods.
- Causal RL aims to incorporate causality into RL, enabling agents to understand the world better and make more informed decisions.

Summary

- Reinforcement Learning (RL) focuses on sequential decision-making problems, where an agent intervenes in an environment with the goal of maximizing cumulative rewards.
- A Markov Decision Process (MDP) describes the dynamics of the environment during interaction, and it can also be represented as an SCM.
- A policy guides an agent's decision-making by mapping states to appropriate actions.
- RL methods can be categorized in various ways, such as online vs. offline and model-free vs. model-based methods.
- Causal RL aims to incorporate causality into RL, enabling agents to understand the world better and make more informed decisions.

Tutorial Outline

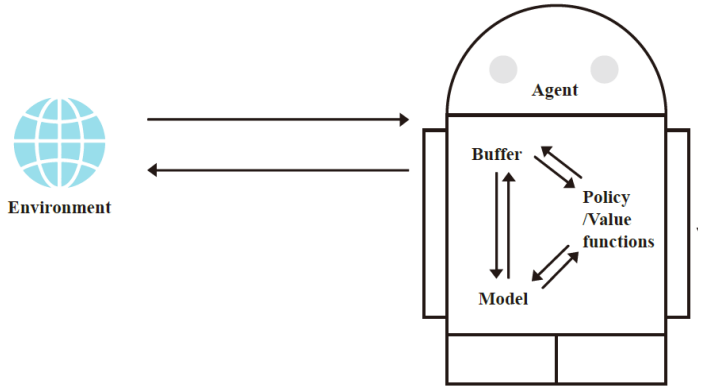
Part 1

- [x] Introduction
- [x] Causality Background
- [x] Reinforcement Learning Background

Part 2

- [ ] Sample Efficiency
- [ ] Generalization Ability
- [ ] Reliability

Causal RL



For more information, check "Causal Reinforcement Learning: A Survey", TMLR 2023

Causal RL

For more information, check "Causal Reinforcement Learning: A Survey", TMLR 2023

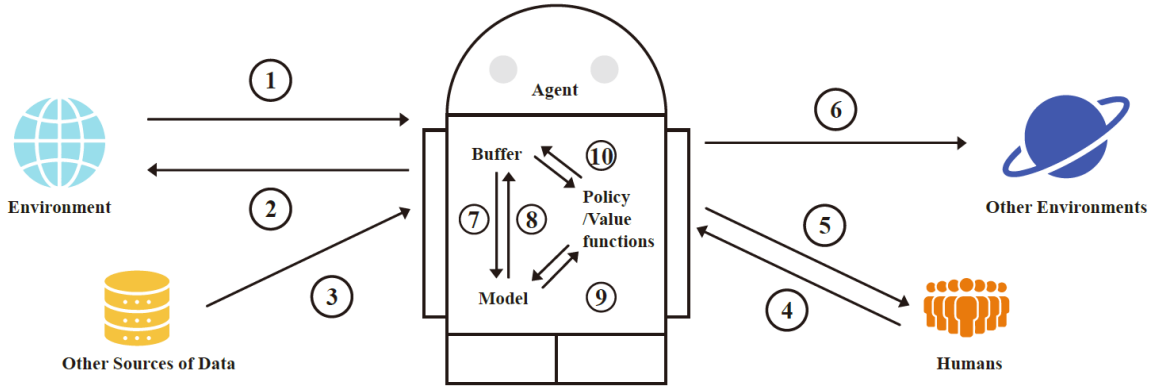


Figure 10: A schematic diagram illustrating the integration of causality into the reinforcement learning process. The numbered edges represent some key components: 1) Abstraction and extraction of causal representations from raw observations; 2) Directed exploration guided by causal knowledge; 3) Fusing (possibly confounded) data; 4) Incorporating causal assumptions or knowledge from humans. 5) Providing causality-based explanations; 6) Generalization and knowledge transfer; 7) Learning causal world models; 8) Counterfactual data generation; 9) Planning with world models; 10) Enhanced training of policies and value functions with causal reasoning.

Sample Efficiency

AlphaGo

3×10^7 games of self-play

AlphaStar

2×10^2 years of self-play

OpenAI Rubik's Cube

10^4 years of simulation

Sample Efficiency

AlphaGo

3×10^7 games of self-play

AlphaStar

2×10^2 years of self-play

OpenAI Rubik's Cube

10^4 years of simulation

Does human learning also require such a large sample size?

Sample Efficiency

RQ1

How can exploration be made more efficient?

Is all unexplored area in the state space equally important?

RQ2

What are the causal variables that govern the environmental dynamics?

Can we accelerate the learning process utilizing these factors?

RQ3

How can agents be equipped with introspective capabilities?

Can agents effectively learn from imaginative experiences?

Directed Exploration

RQ1

How can exploration be made more efficient?

Is all unexplored area in the state space equally important?

Directed Exploration

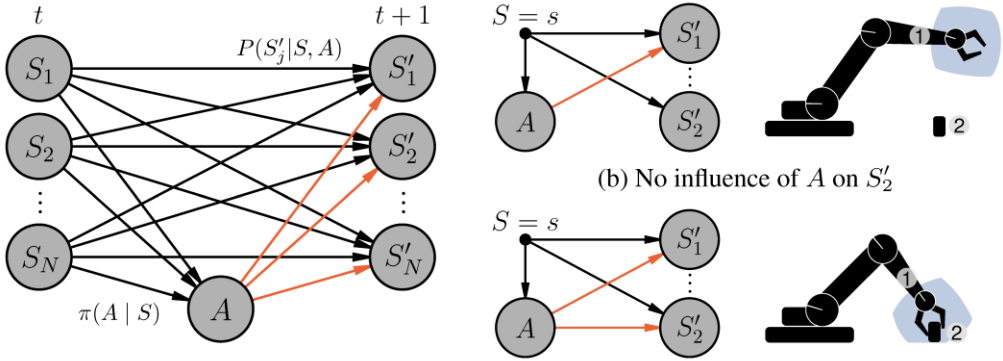
Causal Graph

For more information, check "Causal Influence Detection for Improving Efficiency in Reinforcement Learning", NeurIPS 2021

RQ1

How can exploration be made more efficient?

Is all unexplored area in the state space equally important?



(a) Causal Graph $\mathcal{G}$

(b) No influence of A on S'_2

(c) Influence of A on S'_1 and S'_2

Figure 1: Causal graphical model capturing the environment transition from state S to S' by action A, factorized into state components. (a): Viewed globally over all time steps, all components of the state and the action can influence all state components at the next time step. (b, c): Given a situation $S = s$, some influences may or may not not hold in the *local causal graph* $\mathcal{G}_{S=s}$. In this paper, our aim is to detect which influence the action has on S', i.e. the presence of the red arrows.

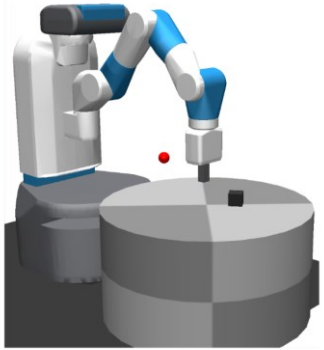


Figure 8: FETCHROTTABLE. The table rotates periodically.

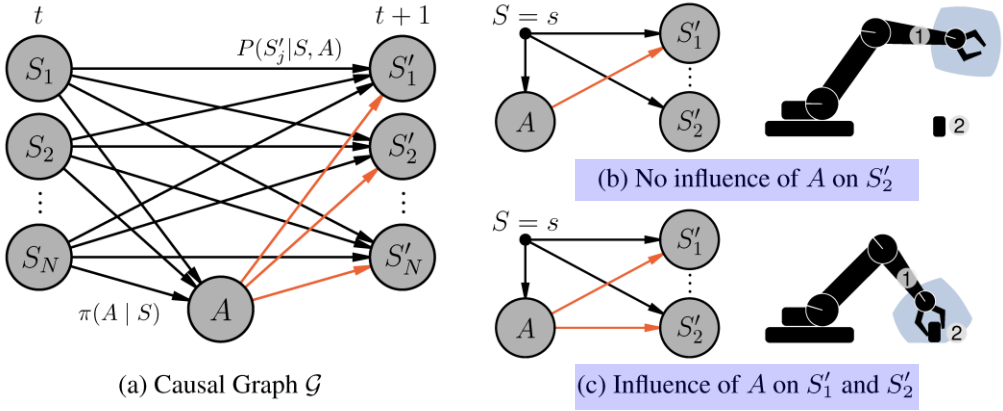
Directed Exploration

Causal Graph

RQ1

How can exploration be made more efficient?

Is all unexplored area in the state space equally important?



(a) Causal Graph $\mathcal{G}$

(b) No influence of A on S'_2

(c) Influence of A on S'_1 and S'_2

How to infer the influence the action has in a specific state configuration $S = s$?

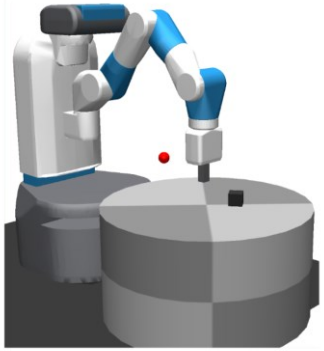


Figure 8: FETCHROT-TABLE. The table rotates periodically.

For more information, check "Causal Influence Detection for Improving Efficiency in Reinforcement Learning", NeurIPS 2021

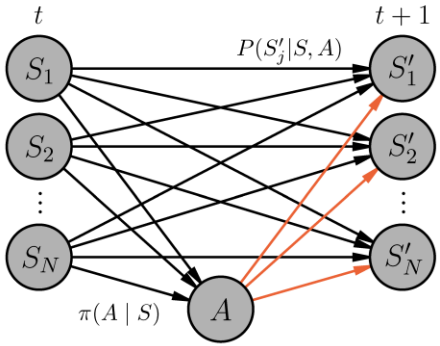
Directed Exploration

Causal Graph

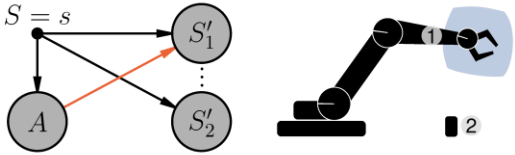
For more information, check "Causal Influence Detection for Improving Efficiency in Reinforcement Learning", NeurIPS 2021

RQ1 How can exploration be made more efficient?

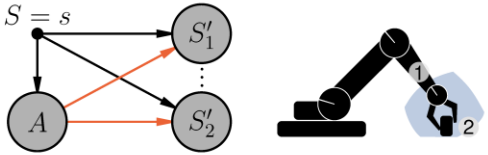
Is all unexplored area in the state space equally important?



(a) Causal Graph $\mathcal{G}$



(b) No influence of A on S'_2



(c) Influence of A on S'_1 and S'_2

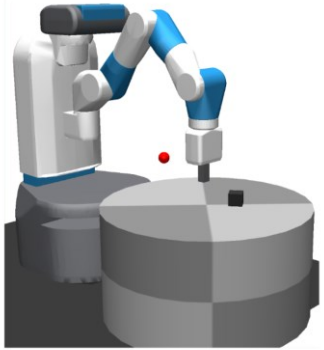


Figure 8: FETCHROT-TABLE. The table rotates periodically.

How to infer the influence the action has in a specific state configuration $S = s$?

Conditional Action Influence (CAI)

$$C^j(s) \coloneqq I\left(S'_j; A \mid S = s\right) = \mathrm{E}_{a\sim\pi}\left[\mathrm{D}_{\mathrm{KL}}\left(P_{S'_j|s,a} \,\middle\|\, P_{S'_j|s}\right)\right]$$

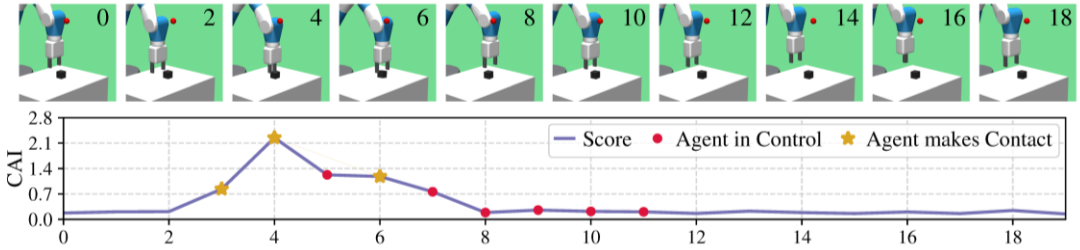
Directed Exploration

Causal Graph

RQ1

How can exploration be made more efficient?

Is all unexplored area in the state space equally important?



Visualizing CAI

- Causal influence as intrinsic motivation
- Greedy w.r.t action influence
- Causal influence as replay priority

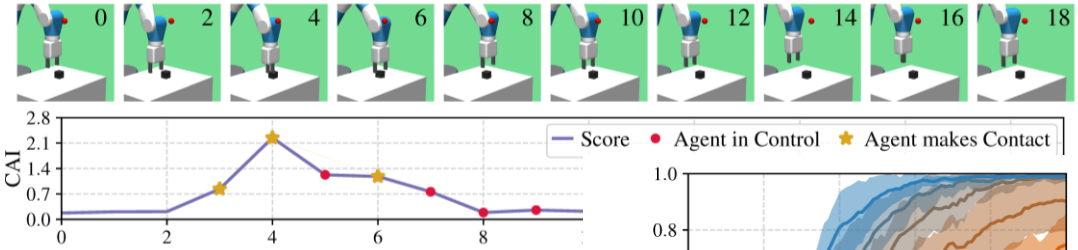
For more information, check "Causal Influence Detection for Improving Efficiency in Reinforcement Learning", NeurIPS 2021

Directed Exploration

Causal Graph

RQ1 How can exploration be made more efficient?

Is all unexplored area in the state space equally important?



- Causal influence as intrinsic motivation
- Greedy w.r.t action influence
- Causal influence as replay priority

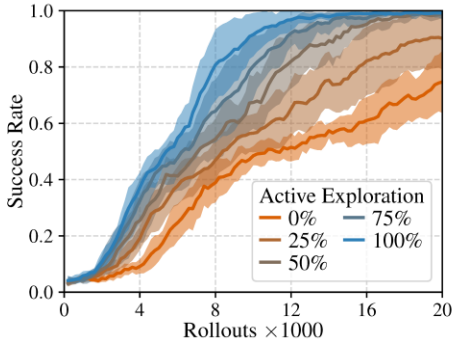


Figure 5: Performance of *active exploration* in FETCHPICKANDPLACE depending on the fraction of exploratory actions chosen actively (Eq. 6) from a total of 30% exploratory actions.

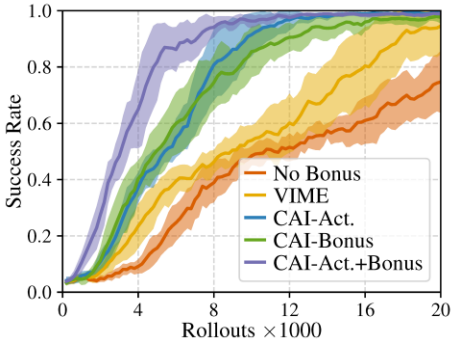


Figure 6: Experiment comparing exploration strategies on FETCHPICKANDPLACE. The combination of active exploration and reward bonus yields the largest sample efficiency.

For more information, check "Causal Influence Detection for Improving Efficiency in Reinforcement Learning", NeurIPS 2021

Causal RL

For more information, check "Causal Reinforcement Learning: A Survey", TMLR 2023

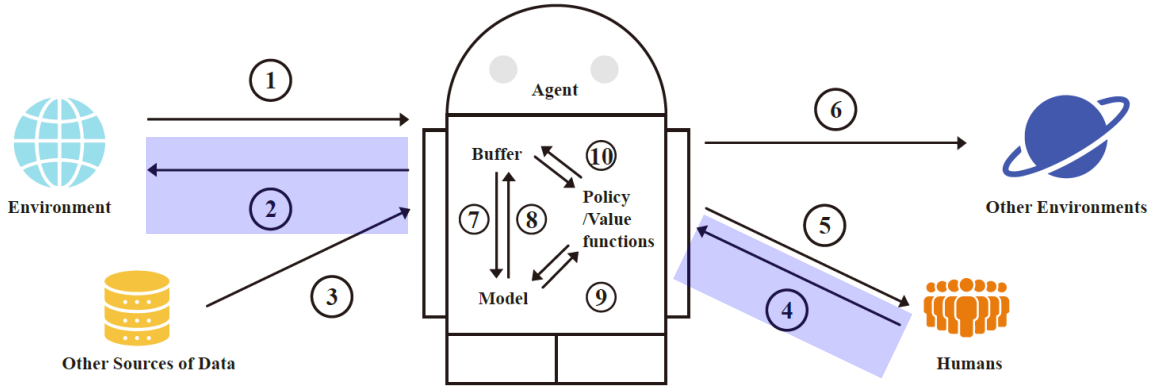


Figure 10: A schematic diagram illustrating the integration of causality into the reinforcement learning process. The numbered edges represent some key components: 1) Abstraction and extraction of causal representations from raw observations; 2) Directed exploration guided by causal knowledge; 3) Fusing (possibly confounded) data; 4) Incorporating causal assumptions or knowledge from humans. 5) Providing causality-based explanations; 6) Generalization and knowledge transfer; 7) Learning causal world models; 8) Counterfactual data generation; 9) Planning with world models; 10) Enhanced training of policies and value functions with causal reasoning.

Causal Representation

RQ2

What are the causal variables that govern the environmental dynamics?

Can we accelerate the learning process utilizing these factors?

Causal Representation

RQ2

What are the causal variables that govern the environmental dynamics?

Can we accelerate the learning process utilizing these factors?

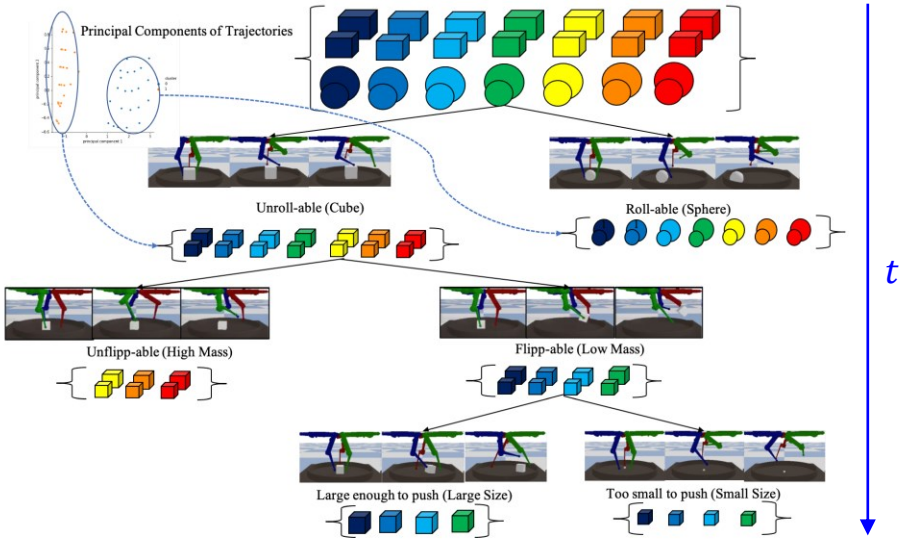


Figure 3. Discovered hierarchical latent space. The agent learns experiments that differentiate the full set of blocks in `ShapeSizeMass` into hierarchical binary clusters. At each level, the environments are divided into 2 clusters on the basis of the value of a single causal factor. We also show the principal components of the trajectories in the top left. For brevity, the full of extent of the tree is not depicted here. For each level of hierarchy k, there are 2^k number of clusters.

For more information, check "Causal Curiosity: RL Agents Discovering Self-supervised Experiments for Causal Representation Learning", ICML 2021

Causal Representation

RQ2

What are the causal variables that govern the environmental dynamics?

Can we accelerate the learning process utilizing these factors?

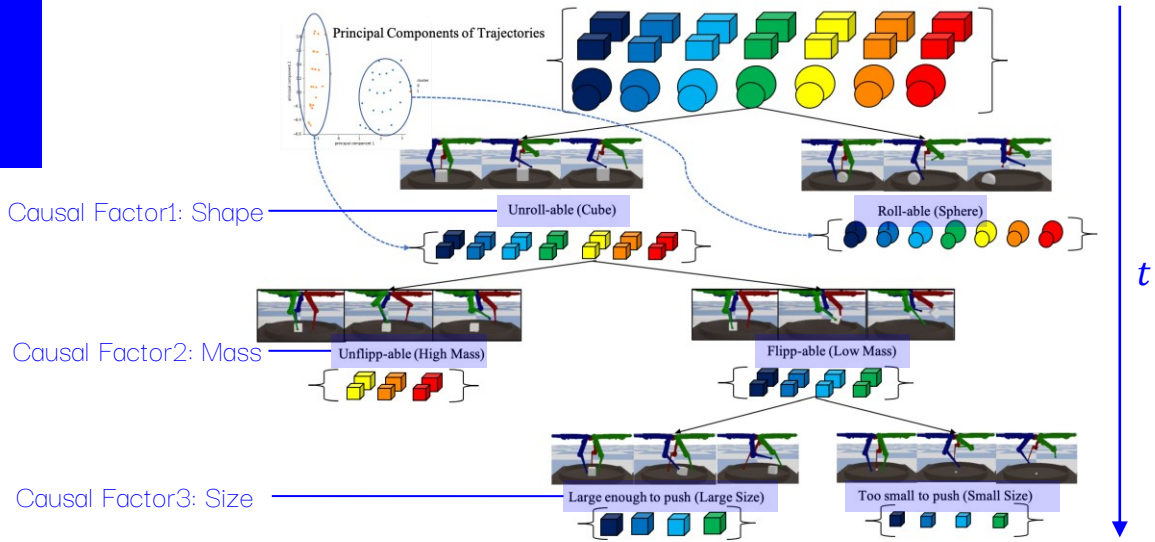


Figure 3. Discovered hierarchical latent space. The agent learns experiments that differentiate the full set of blocks in `ShapeSizeMass` into hierarchical binary clusters. At each level, the environments are divided into 2 clusters on the basis of the value of a single causal factor. We also show the principal components of the trajectories in the top left. For brevity, the full of extent of the tree is not depicted here. For each level of hierarchy k, there are 2^k number of clusters.

For more information, check "Causal Curiosity: RL Agents Discovering Self-supervised Experiments for Causal Representation Learning", ICML 2021

How to empower agents to discover semantically meaningful experimental behaviors rather than maximizing reward for a particular task?

Causal Representation

RQ2

What are the causal variables that govern the environmental dynamics?

Can we accelerate the learning process utilizing these factors?

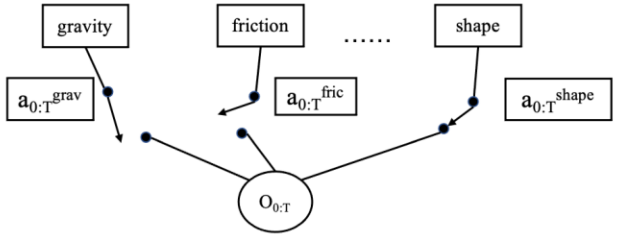


Figure 2. Gated Causal Graph. A subset of the unobserved parent causal variables influence the observed variable $\mathbf{O}$. The action sequence $\mathbf{a}_{0:T}$ serves a gating mechanism, allowing or blocking particular edges of the causal graph using the implicit Causal Selector Function (Equation 4).

For more information, check "Causal Curiosity: RL Agents Discovering Self-supervised Experiments for Causal Representation Learning", ICML 2021

How to empower agents to discover semantically meaningful experimental behaviors rather than maximizing reward for a particular task?

Causal Representation ICM

For more information, check "Causal Curiosity: RL Agents Discovering Self-supervised Experiments for Causal Representation Learning", ICML 2021

RQ2 What are the causal variables that govern the environmental dynamics?

Can we accelerate the learning process utilizing these factors?

Independent Causal Mechanism

The information in an observed trajectory is the sum of information "injected" into it from the multiple causes

The information content will be greater for a larger number of causal parents in the graph.

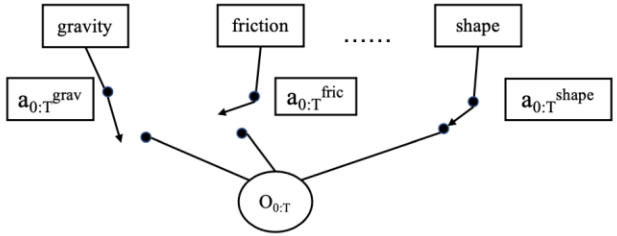


Figure 2. Gated Causal Graph. A subset of the unobserved parent causal variables influence the observed variable **O**. The action sequence $\mathbf{a}_{0:T}$ serves a gating mechanism, allowing or blocking particular edges of the causal graph using the implicit Causal Selector Function (Equation 4).

How to empower agents to discover semantically meaningful experimental behaviors rather than maximizing reward for a particular task?

Causal Representation ICM

For more information, check "Causal Curiosity: RL Agents Discovering Self-supervised Experiments for Causal Representation Learning", ICML 2021

RQ2 What are the causal variables that govern the environmental dynamics?

Can we accelerate the learning process utilizing these factors?

One-Factor-at-A-Time

We want to search for one causal factor at a time.

Find an action sequence for which the number of causal parents of is low, i.e., the complexity of $\mathbf{O}$.

Independent Causal Mechanism

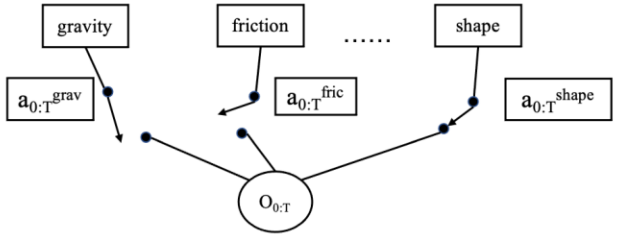


Figure 2. Gated Causal Graph. A subset of the unobserved parent causal variables influence the observed variable $\mathbf{O}$. The action sequence $\mathbf{a}_{0:T}$ serves a gating mechanism, allowing or blocking particular edges of the causal graph using the implicit Causal Selector Function (Equation 4).

How to empower agents to discover semantically meaningful experimental behaviors rather than maximizing reward for a particular task?

Causal Representation ICM

RQ2 What are the causal variables that govern the environmental dynamics?

Can we accelerate the learning process utilizing these factors?

For more information, check "Causal Curiosity: RL Agents Discovering Self-supervised Experiments for Causal Representation Learning", ICML 2021

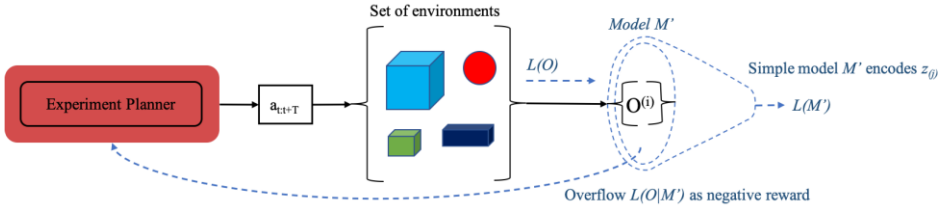


Figure 9. Overview of training. The experiment planner generates a trajectory of actions which is applied to each of the environments with varying causal factors namely mass, shape and size of blocks. For each environment, an observation trajectory or state $\mathbf{O}^i \in O$ is obtained. A simple model with fixed low expressive power is used to approximate the generative model for $\mathbf{O}$. The "information overflow" $L(\mathbf{O}|\mathbf{M})$ is returned as negative reward forcing O to be caused by few causal factors.

How to empower agents to discover semantically meaningful experimental behaviors rather than maximizing reward for a particular task?

Minimizing the complexity → Maximizing the causal curiosity

$$a^*_{0:T} = \underset{a_{0:T}}{\operatorname{argmin}}\, L(O|M) \qquad a^*_{0:T} = \underset{a_{0:T}}{\operatorname{argmax}} -L(O|M)$$

Causal Representation ICM

RQ2 What are the causal variables that govern the environmental dynamics?

Can we accelerate the learning process utilizing these factors?

For more information, check "Causal Curiosity: RL Agents Discovering Self-supervised Experiments for Causal Representation Learning", ICML 2021

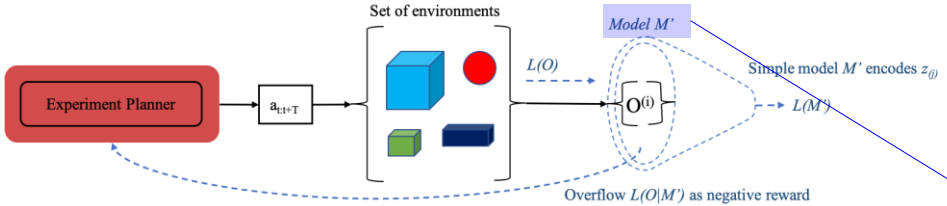


Figure 9. Overview of training. The experiment planner generates a trajectory of actions which is applied to each of the environments with varying causal factors namely mass, shape and size of blocks. For each environment, an observation trajectory or state $\mathbf{O}^i \in O$ is obtained. A simple model with fixed low expressive power is used to approximate the generative model for $\mathbf{O}$. The "information overflow" $L(\mathbf{O}|\mathbf{M})$ is returned as negative reward forcing O to be caused by few causal factors.

How to empower agents to discover semantically meaningful experimental behaviors rather than maximizing reward for a particular task?

Assume $\mathbf{M}$ is a bimodal clustering model.

Minimizing the complexity

$$a^*_{0:T} = \underset{a_{0:T}}{\operatorname{argmin}}\, L(O|M)$$

Maximizing the causal curiosity

$$a^*_{0:T} = \underset{a_{0:T}}{\operatorname{argmax}}\, -L(O|M)$$

Causal Representation

ICM

RQ2

What are the causal variables that govern the environmental dynamics?

Can we accelerate the learning process utilizing these factors?

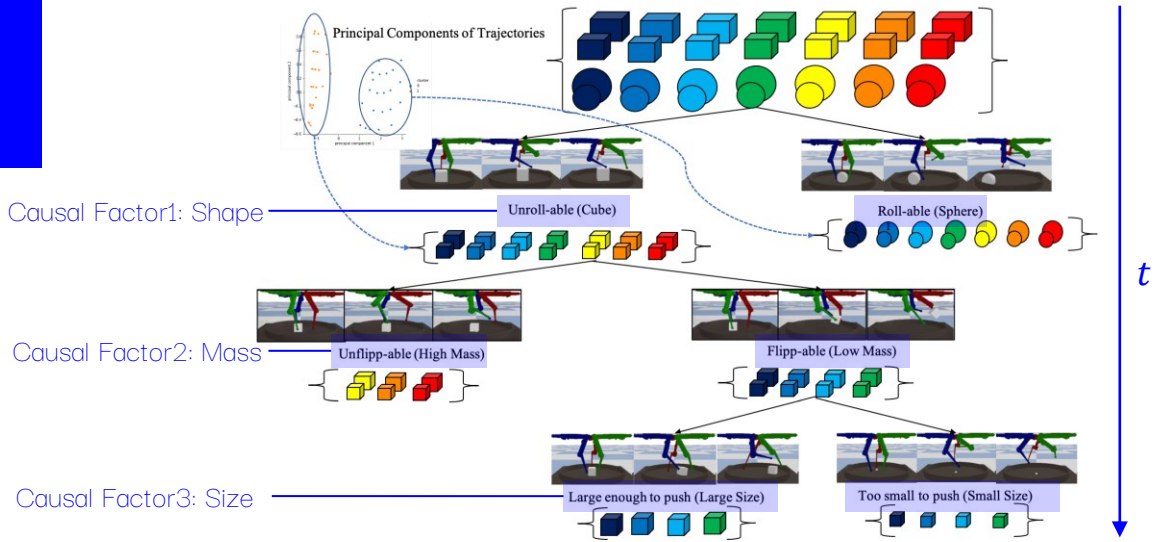
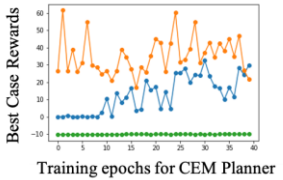
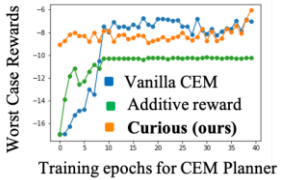


Figure 3. Discovered hierarchical latent space. The agent learns experiments that differentiate the full set of blocks in `ShapeSizeMass` into hierarchical binary clusters. At each level, the environments are divided into 2 clusters on the basis of the value of a single causal factor. We also show the principal components of the trajectories in the top left. For brevity, the full of extent of the tree is not depicted here. For each level of hierarchy k, there are 2^k number of clusters.

For more information, check "Causal Curiosity: RL Agents Discovering Self-supervised Experiments for Causal Representation Learning", ICML 2021

How to empower agents to discover semantically meaningful experimental behaviors rather than maximizing reward for a particular task?





The behaviors discovered by the agents while optimizing causal curiosity show high zero-shot Generalization Ability and converge to the same performance as conventional planners for downstream tasks.

Causal RL

For more information, check "Causal Reinforcement Learning: A Survey", TMLR 2023

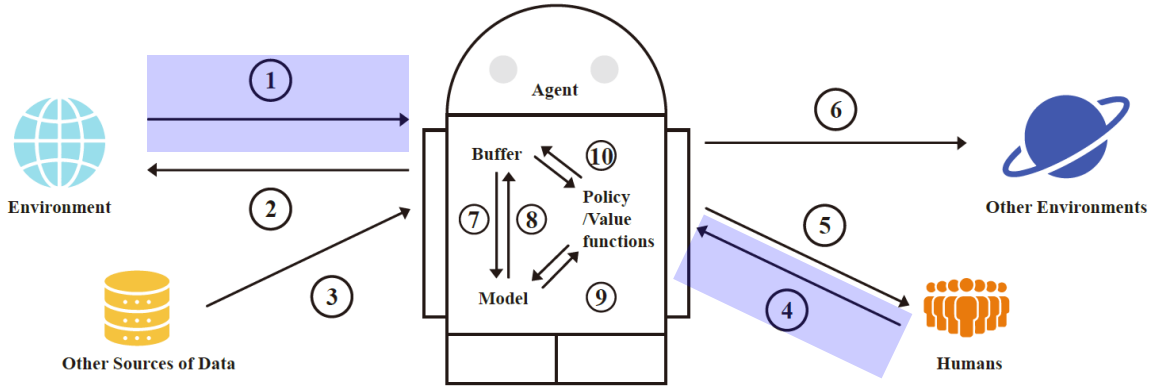


Figure 10: A schematic diagram illustrating the integration of causality into the reinforcement learning process. The numbered edges represent some key components: 1) Abstraction and extraction of causal representations from raw observations; 2) Directed exploration guided by causal knowledge; 3) Fusing (possibly confounded) data; 4) Incorporating causal assumptions or knowledge from humans. 5) Providing causality-based explanations; 6) Generalization and knowledge transfer; 7) Learning causal world models; 8) Counterfactual data generation; 9) Planning with world models; 10) Enhanced training of policies and value functions with causal reasoning.

Counterfactual Reasoning

RQ3

How can agents be equipped with introspective capabilities?

Can agents effectively learn from imaginative experiences?

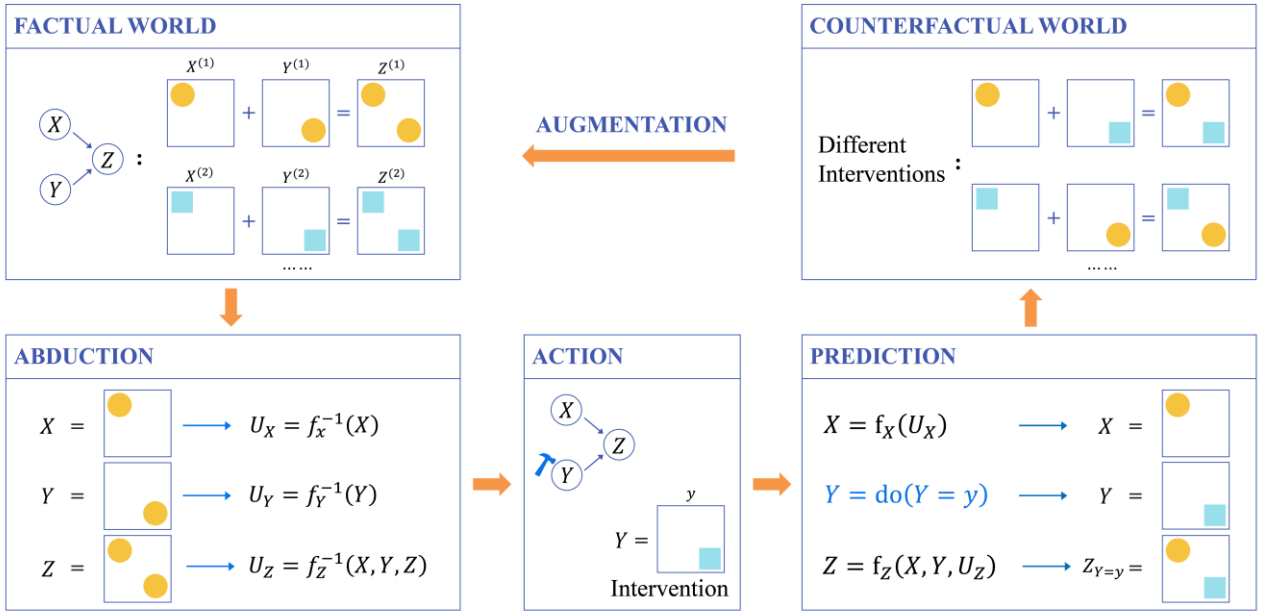
For more information, check "Causal Reinforcement Learning: A Survey", TMLR 2023

Counterfactual Reasoning

RQ3

How can agents be equipped with introspective capabilities?

Can agents effectively learn from imaginative experiences?



For more information, check "Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search", ICLR 2019

Counterfactual Reasoning SCM

RQ3 How can agents be equipped with introspective capabilities?

Can agents effectively learn from imaginative experiences?

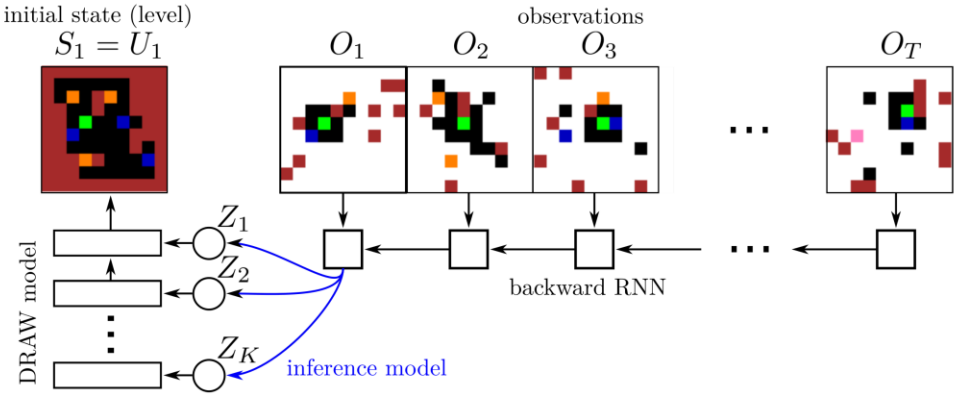
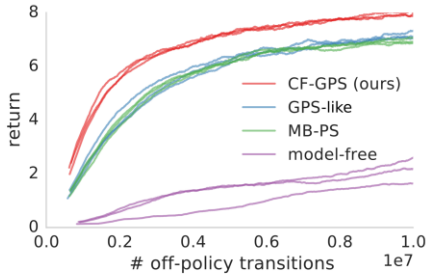


Figure 3: **Top: PO-SOKOBAN**. Shown on the left is a procedurally generated initial state. The agent is shown in green, boxes in yellow, targets in blue and walls in red. The agent does not observe this state but a sequence of observations, which are masked by iid noise with 0.9 probability, except a 3x3 window around the agent. **Bottom: Inference model.** For counterfactual inference in PO-SOKOBAN, we need the (approximate) inference distribution $p(U_{s1}|\hat{h}_T)$ over the initial state $U_{s1} = S_1$, conditioned on the history of observations $\hat{h}_T$. We model this distribution using a DRAW generative model with latent variables Z, which are conditioned on the output of a backward RNN summarizing the observation history.



Counterfactually-Guided Policy Search (CF-GPS) outperforms a naive model-based RL (MB-PS) algorithm as well as model-free methods

For more information, check "Causal Reinforcement Learning: A Survey", TMLR 2023

Causal RL

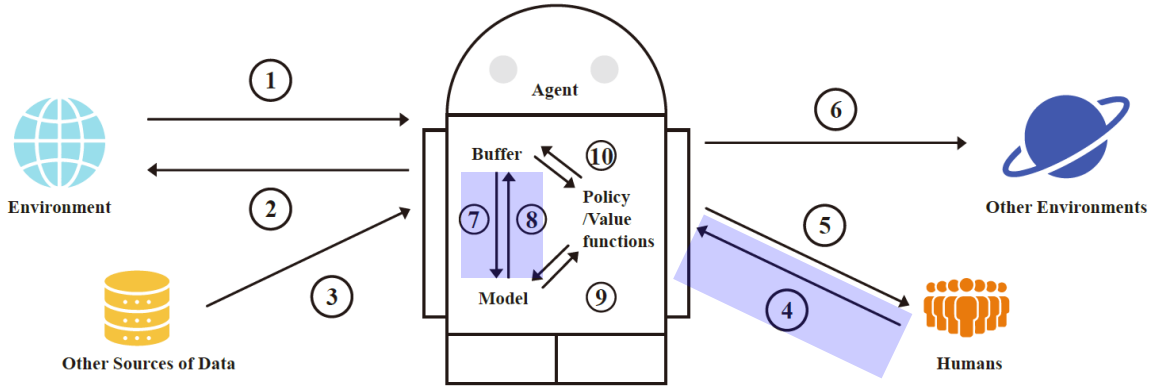


Figure 10: A schematic diagram illustrating the integration of causality into the reinforcement learning process. The numbered edges represent some key components: 1) Abstraction and extraction of causal representations from raw observations; 2) Directed exploration guided by causal knowledge; 3) Fusing (possibly confounded) data; 4) Incorporating causal assumptions or knowledge from humans. 5) Providing causality-based explanations; 6) Generalization and knowledge transfer; 7) Learning causal world models; 8) Counterfactual data generation; 9) Planning with world models; 10) Enhanced training of policies and value functions with causal reasoning.

Tutorial Outline

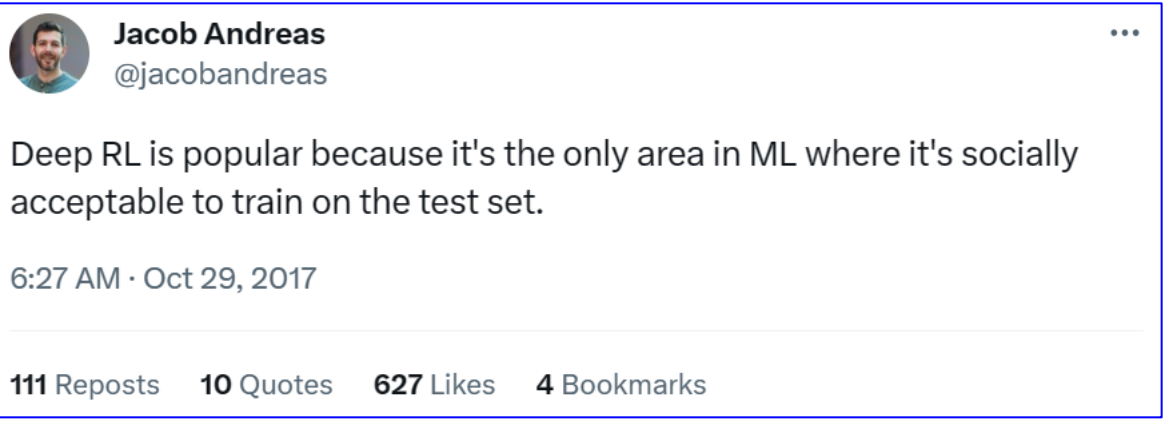
Part 1

- [x] Introduction
- [x] Causality Background
- [x] Reinforcement Learning Background

Part 2

- [x] Sample Efficiency
- [ ] Generalization Ability
- [ ] Reliability

Generalization Ability



Jacob Andreas
@jacobandreas

Deep RL is popular because it's the only area in ML where it's socially acceptable to train on the test set.

6:27 AM · Oct 29, 2017

111 Reposts **10** Quotes **627** Likes **4** Bookmarks

Generalization Ability

Jacob Andreas

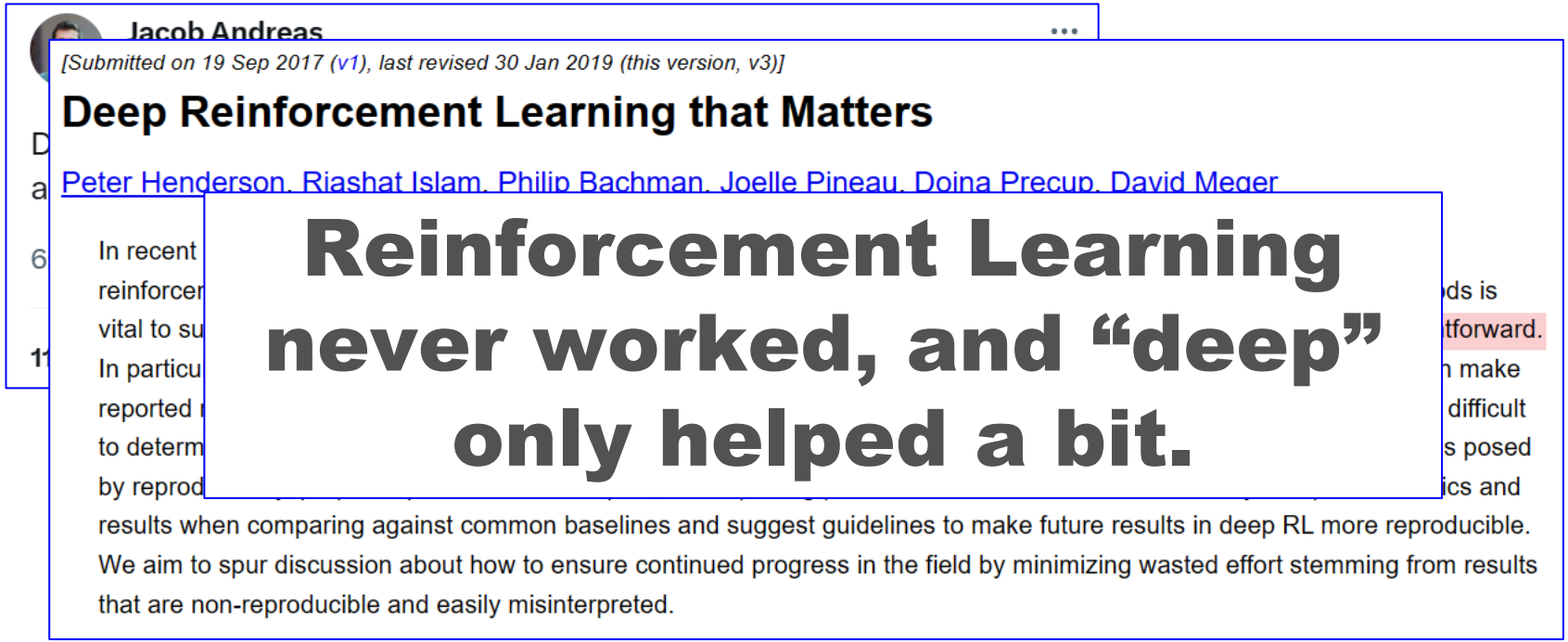
[Submitted on 19 Sep 2017 (v1), last revised 30 Jan 2019 (this version, v3)]

Deep Reinforcement Learning that Matters

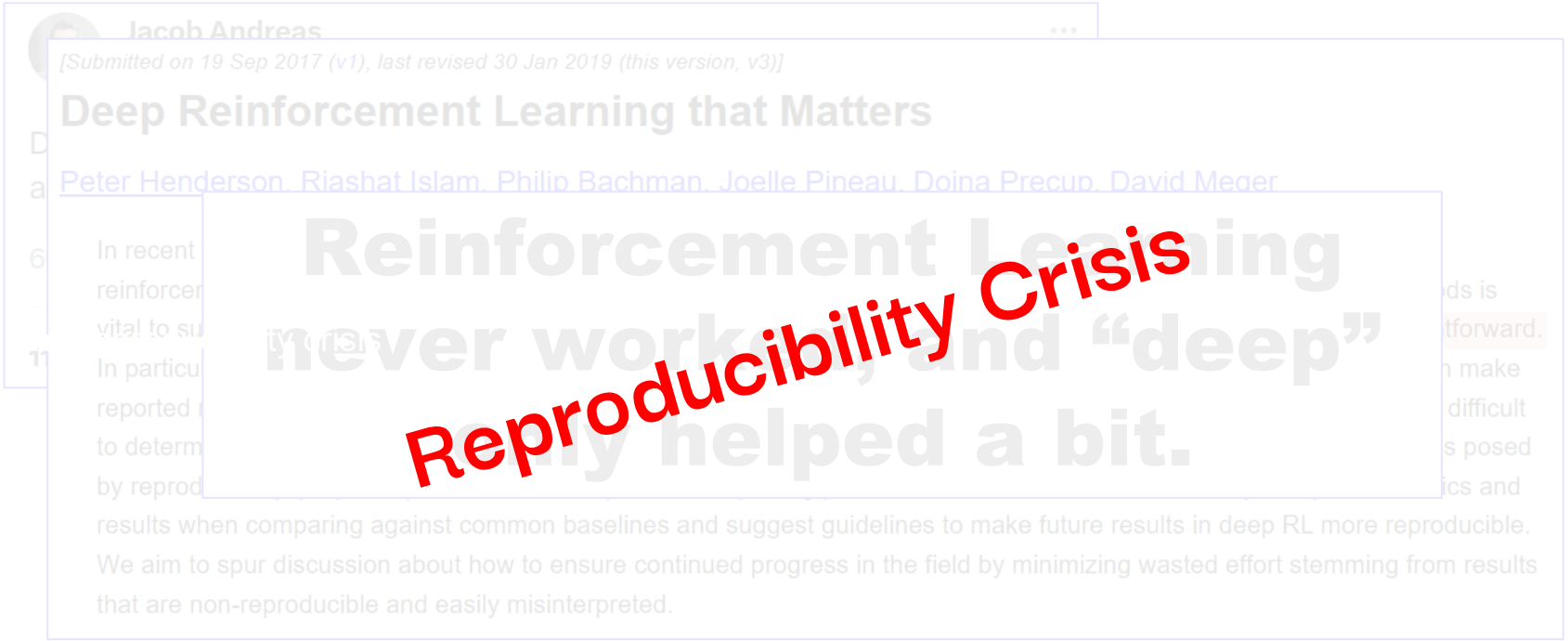
Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, David Meger

In recent years, significant progress has been made in solving challenging problems across various domains using deep reinforcement learning (RL). Reproducing existing work and accurately judging the improvements offered by novel methods is vital to sustaining this progress. Unfortunately, reproducing results for state-of-the-art deep RL methods is seldom straightforward. In particular, non-determinism in standard benchmark environments, combined with variance intrinsic to the methods, can make reported results tough to interpret. Without significance metrics and tighter standardization of experimental reporting, it is difficult to determine whether improvements over the prior state-of-the-art are meaningful. In this paper, we investigate challenges posed by reproducibility, proper experimental techniques, and reporting procedures. We illustrate the variability in reported metrics and results when comparing against common baselines and suggest guidelines to make future results in deep RL more reproducible. We aim to spur discussion about how to ensure continued progress in the field by minimizing wasted effort stemming from results that are non-reproducible and easily misinterpreted.

Generalization Ability



Generalization Ability



Generalization Ability

RQ1 What does generalization mean for agents?

How can agents achieve reliable generalization despite unknown variations?

RQ2 What knowledge can be transferred?

How can algorithms be designed to facilitate efficient adaptation?

Generalization Ability

RQ1

What does generalization mean for agents?

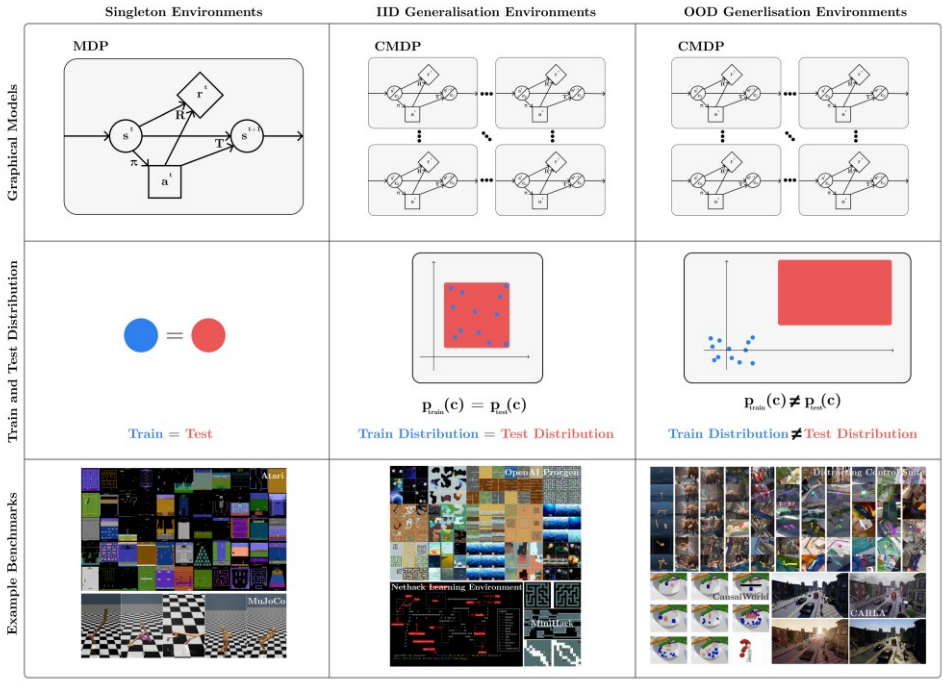
How can agents achieve reliable generalization despite unknown variations?

Generalization Ability

RQ1

What does generalization mean for agents?

How can agents achieve reliable generalization despite unknown variations?



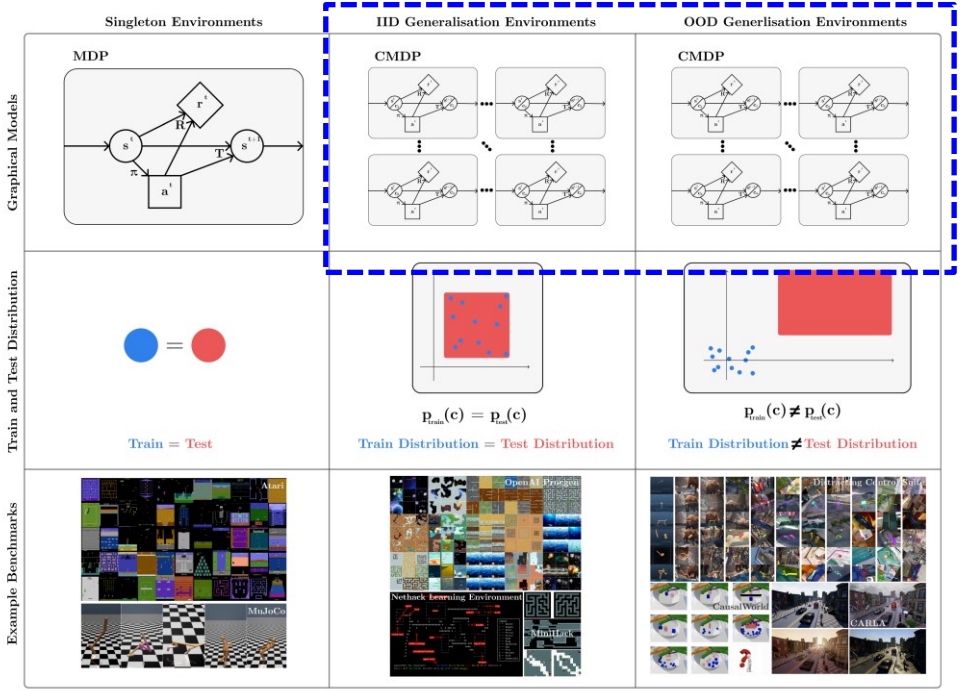
For more information, check "A Survey of Generalisation in Deep Reinforcement Learning", 2021

Generalization Ability

RQ1

What does generalization mean for agents?

How can agents achieve reliable generalization despite unknown variations?



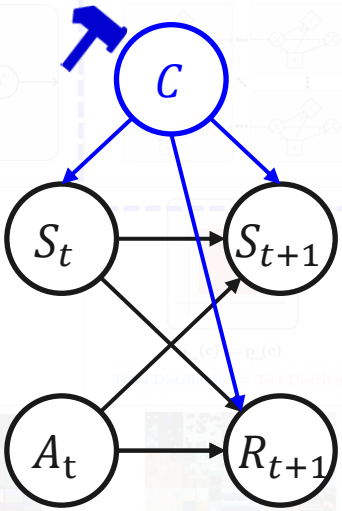
For more information, check "A Survey of Generalisation in Deep Reinforcement Learning", 2021

For more information, check "Causal Reinforcement Learning: A Survey", TMLR 2023

Generalization Ability

RQ1 What does generalization mean for agents?

How can agents achieve reliable generalization despite unknown variations?



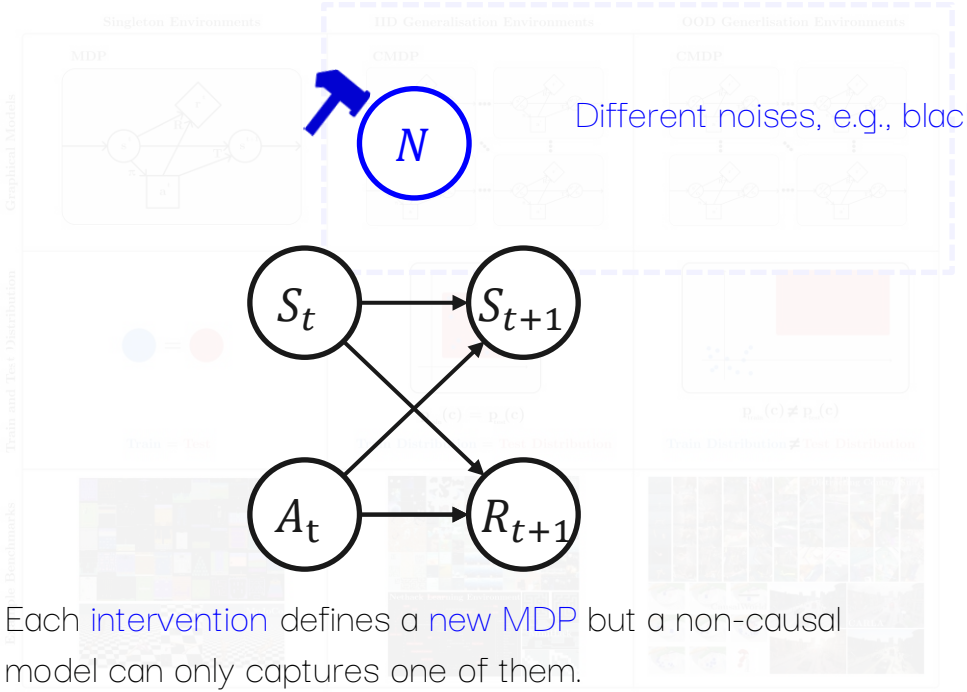
Each intervention defines a new MDP but a non-causal model can only captures one of them.

For more information, check "Causal Reinforcement Learning: A Survey", TMLR 2023

Generalization Ability

RQ1 What does generalization mean for agents?

How can agents achieve reliable generalization despite unknown variations?



Different noises, e.g., black-out perturbation.

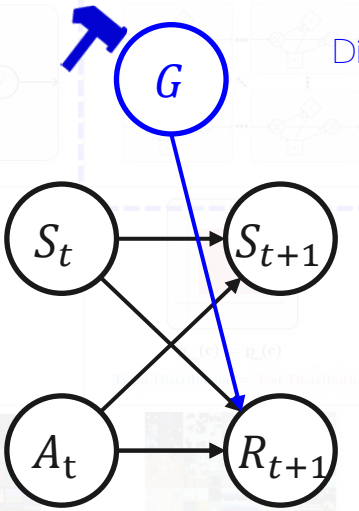
Each intervention defines a new MDP but a non-causal model can only captures one of them.

For more information, check "Causal Reinforcement Learning: A Survey", TMLR 2023

Generalization Ability

RQ1 What does generalization mean for agents?

How can agents achieve reliable generalization despite unknown variations?



Different goals, e.g., target locations.

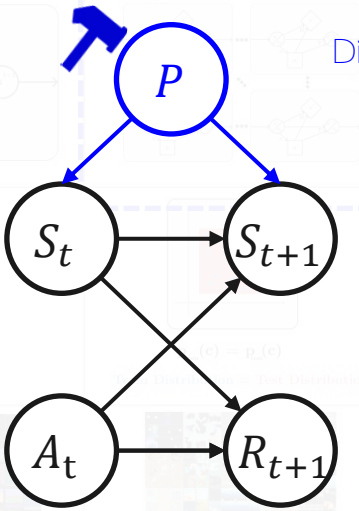
Each intervention defines a new MDP but a non-causal model can only captures one of them.

For more information, check "Causal Reinforcement Learning: A Survey", TMLR 2023

Generalization Ability

RQ1 What does generalization mean for agents?

How can agents achieve reliable generalization despite unknown variations?



Different physical properties, e.g., mass.

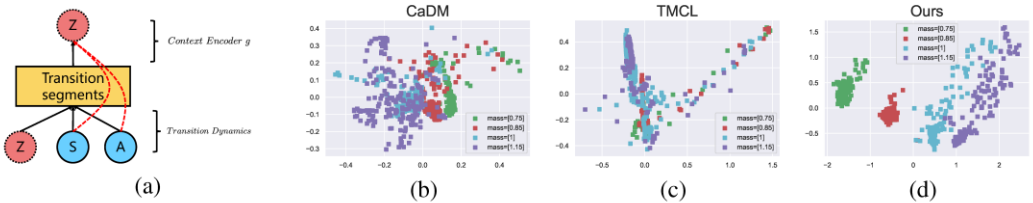
Each intervention defines a new MDP but a non-causal model can only captures one of them.

Generalization

RQ1

What does generalization mean for agents?

How can agents achieve reliable generalization despite unknown variations?



(a) (b) (c) (d)

Figure 1: (a) The illustration of why historical states and actions are encoded in environment-specified factor Z, (b)(c)(d) The PCA visulization of estimated context (environmental-specific) vectors in **Pendulum** task, where the dots with different colors denote that the context vector (after PCA) estimated from different environments. More visualization results are given at Appendix A.13.

For more information, check "A Relational Intervention Approach for Unsupervised Dynamics Generalization in Model-Based Reinforcement Learning", ICLR 2022

Generalization

For more information, check "A Relational Intervention Approach for Unsupervised Dynamics Generalization in Model-Based Reinforcement Learning", ICLR 2022

RQ1

What does generalization mean for agents?

How can agents achieve reliable generalization despite unknown variations?

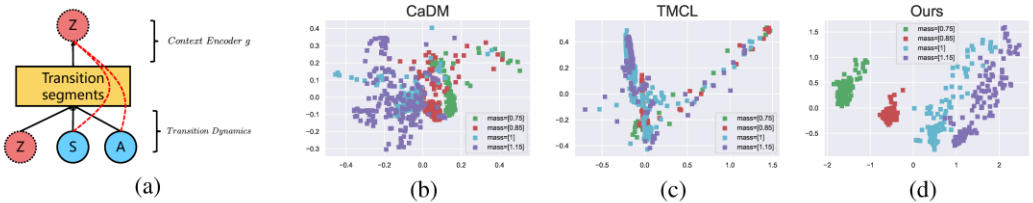


Figure 1: (a) The illustration of why historical states and actions are encoded in environment-specified factor Z, (b)(c)(d) The PCA visulization of estimated context (environmental-specific) vectors in **Pendulum** task, where the dots with different colors denote that the context vector (after PCA) estimated from different environments. More visualization results are given at Appendix A.13.

How can the extraction of contextual variable Z from past transition segments be improved to ensure that Z maintains its crucial property of being similar in the same environment and dissimilar in different ones for effective model generalization?

Generalization

For more information, check "A Relational Intervention Approach for Unsupervised Dynamics Generalization in Model-Based Reinforcement Learning", ICLR 2022

RQ1

What does generalization mean for agents?

How can agents achieve reliable generalization despite unknown variations?

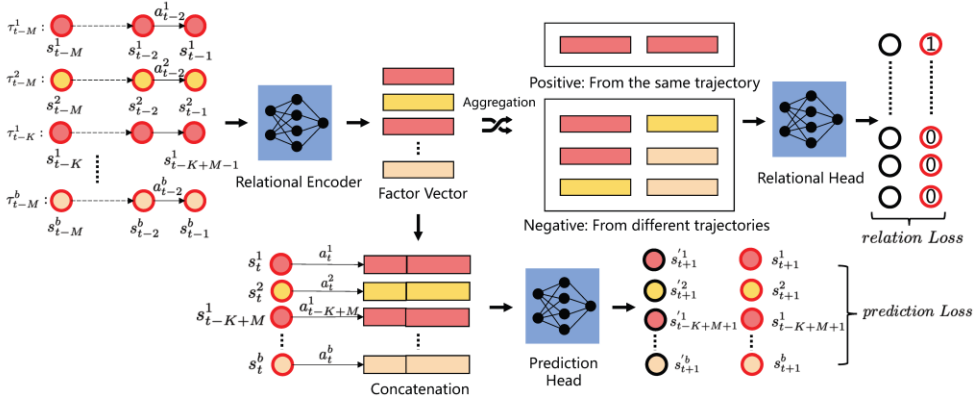


Figure 2: An overview of our Relational Intervention approach, where Relational Encoder, Prediction Head and Relational Head are three learnable functions, and the circles denote states (Ground-Truths are with red boundary, and estimated states are with black boundary), and the rectangles denote the estimated vectors. Specifically, *prediction Loss* enables the estimated environmental-specified factor can help the Prediction head to predict the next states, and the *relation Loss* aims to enforce the similarity between factors estimated from the same trajectory or similar trajectories.

How can the extraction of contextual variable Z from past transition segments be improved to ensure that Z maintains its crucial property of being similar in the same environment and dissimilar in different ones for effective model generalization?

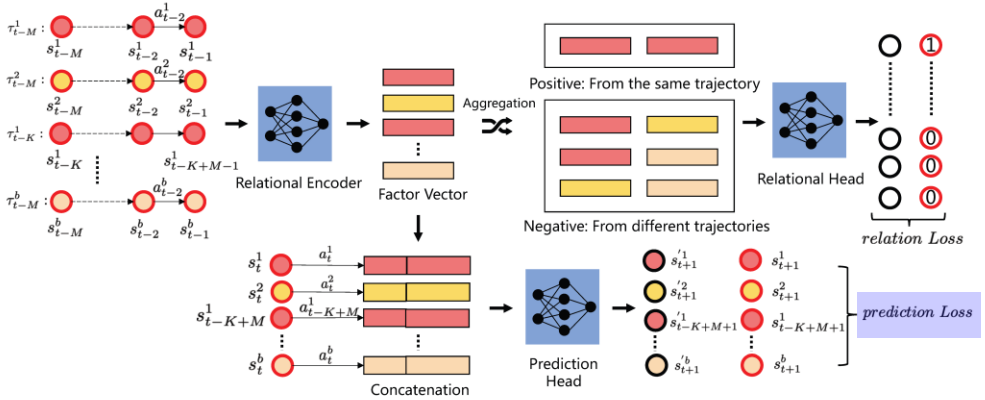
Generalization

For more information, check "A Relational Intervention Approach for Unsupervised Dynamics Generalization in Model-Based Reinforcement Learning", ICLR 2022

RQ1

What does generalization mean for agents?

How can agents achieve reliable generalization despite unknown variations?



How can the extraction of contextual variable Z from past transition segments be improved to ensure that Z maintains its crucial property of being similar in the same environment and dissimilar in different ones for effective model generalization?

$$\mathcal{L}^{pred}_{\theta,\phi} = -\frac{1}{N}\sum_{i=1}^{N} \log\ \hat{f}(s^i_{t+1}|s^i_t, a^i_t, g(\tau^i_{t-k:t-1};\phi);\theta)$$

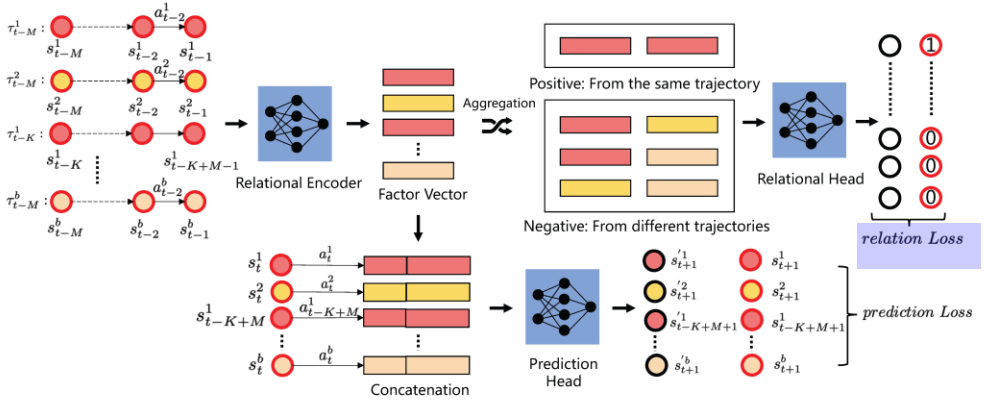
Generalization

For more information, check "A Relational Intervention Approach for Unsupervised Dynamics Generalization in Model-Based Reinforcement Learning", ICLR 2022

RQ1

What does generalization mean for agents?

How can agents achieve reliable generalization despite unknown variations?



How can the extraction of contextual variable Z from past transition segments be improved to ensure that Z maintains its crucial property of being similar in the same environment and dissimilar in different ones for effective model generalization?

$$\mathcal{L}_{\varphi,\phi}^{relation} = -\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[y^{i,j}\cdot\log\ h([\hat{z}^i,\hat{z}^j];\varphi)+(1-y^{i,j})\cdot\log\left(1-h([\hat{z}^i,\hat{z}^j];\varphi)\right)\right]$$

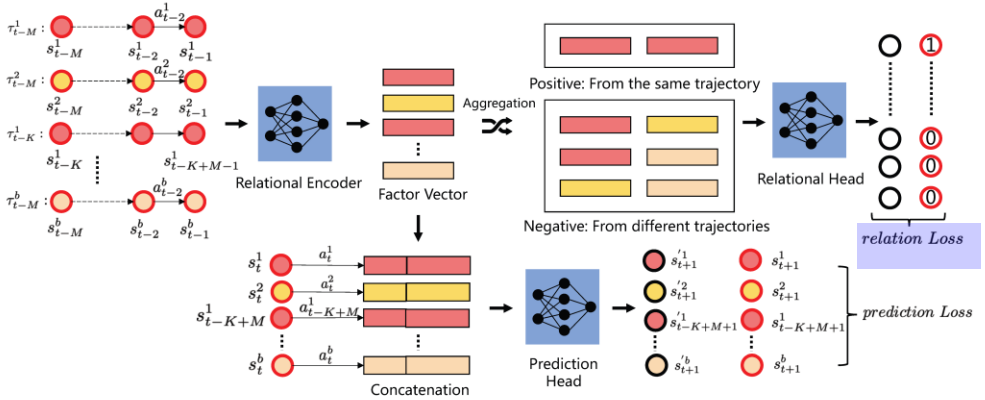
Generalization

For more information, check "A Relational Intervention Approach for Unsupervised Dynamics Generalization in Model-Based Reinforcement Learning", ICLR 2022

RQ1

What does generalization mean for agents?

How can agents achieve reliable generalization despite unknown variations?



How can the extraction of contextual variable Z from past transition segments be improved to ensure that Z maintains its crucial property of being similar in the same environment and dissimilar in different ones for effective model generalization?

$$\mathcal{L}^{relation}_{\varphi,\phi} = -\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[y^{i,j}\cdot\log\ h([\hat{z}^i,\hat{z}^j];\varphi)+(1-y^{i,j})\cdot\log\left(1-h([\hat{z}^i,\hat{z}^j];\varphi)\right)\right]$$

Estimating trajectory invariant information is insufficient because the estimated $\hat{Z}$s in the same environment will also be pushed away, which may undermine the cluster compactness for estimated $\hat{Z}$s.

Generalization

The causal effects induced by the same context are similar.

For more information, check "A Relational Intervention Approach for Unsupervised Dynamics Generalization in Model-Based Reinforcement Learning", ICLR 2022

RQ1

What does generalization mean for agents?

How can agents achieve reliable generalization despite unknown variations?

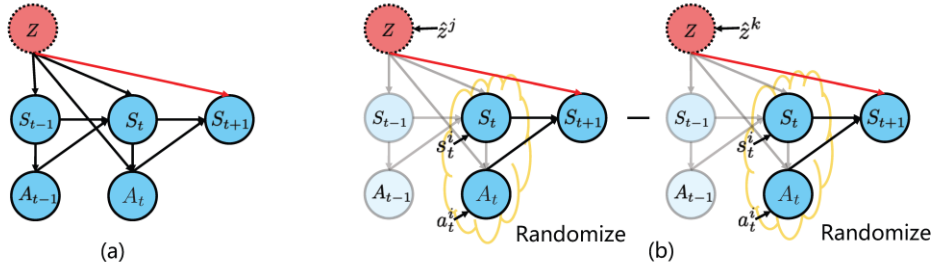


Figure 3: (a) The illustration of causal graph, and the red line denotes the direct causal effect from Z to S_{t+1}. (b) The illustration of estimating the controlled causal effect.

How can the extraction of contextual variable Z from past transition segments be improved to ensure that Z maintains its crucial property of being similar in the same environment and dissimilar in different ones for effective model generalization?

similarity

$$\mathcal{L}_{\varphi,\phi}^{i-relation} = -\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}\Big[\,[y^{i,j}+(1-y^{i,j})\cdot w^{i,j}]\cdot\log\, h([\hat{z}^i,\hat{z}^j];\varphi) + (1-y^{i,j})\cdot(1-w^{i,j})\cdot\log\,(1-h([\hat{z}^i,\hat{z}^j];\varphi))\Big],$$

$$ACDE_{\hat{z}^j,\hat{z}^k} = \frac{1}{N}\sum_{i=1}^{N}|CDE_{\hat{z}^j,\hat{z}^k}(s_t^i,a_t^i)|$$

The estimated $\hat{Z}$s in the same environment should have similar causal effect on S_{t+1}.

Generalization

The causal effects induced by the same context are similar.

For more information, check "A Relational Intervention Approach for Unsupervised Dynamics Generalization in Model-Based Reinforcement Learning", ICLR 2022

RQ1

What does generalization mean for agents?

How can agents achieve reliable generalization despite unknown variations?

Table 3: The prediction errors of methods on test environments

	CaDM (Lee et al., 2020)	TMCL (Seo et al., 2020)	Ours
Hopper	0.0551± 0.0236	0.0316 ± 0.0138	**0.0271 ± 0.0011**
Ant	0.3850 ± 0.0256	0.1560 ± 0.0106	**0.1381 ± 0.0047**
C_Halfcheetah	0.0815 ± 0.0029	0.0751 ±0.0123	**0.0525 ± 0.0061**
HalfCheetah	0.6151 ± 0.0251	1.0136 ± 0.6241	**0.4513 ±0.2147**
Pendulum	0.0160 ±0.0036	0.0130± 0.0835	**0.0030 ± 0.0012**
Slim_Humanoid	0.8842 ± 0.2388	0.3243 ± 0.0027	**0.3032 ± 0.0046**

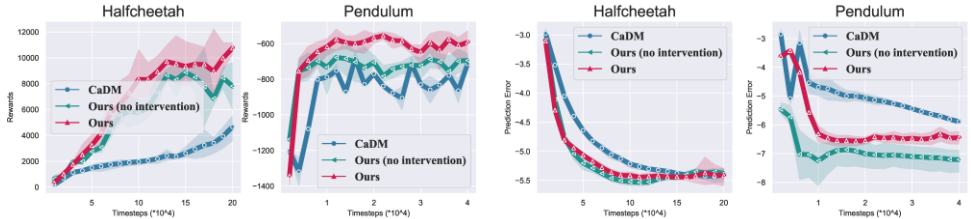


Figure 6: (a) The average rewards of trained model-based RL agents on unseen environments. The results show the mean and standard deviation of returns averaged over three runs. (b) The average prediction errors over the training procedure.

How can the extraction of contextual variable Z from past transition segments be improved to ensure that Z maintains its crucial property of being similar in the same environment and dissimilar in different ones for effective model generalization?

Causal RL

For more information, check "Causal Reinforcement Learning: A Survey", TMLR 2023

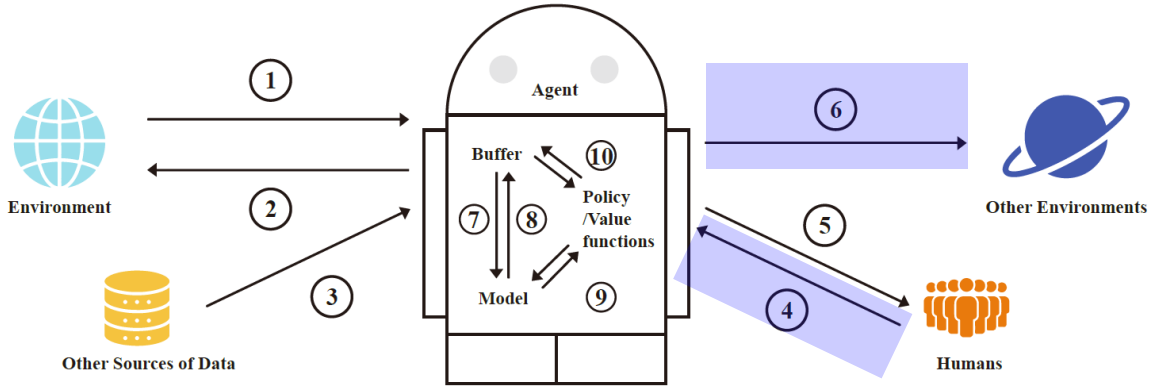


Figure 10: A schematic diagram illustrating the integration of causality into the reinforcement learning process. The numbered edges represent some key components: 1) Abstraction and extraction of causal representations from raw observations; 2) Directed exploration guided by causal knowledge; 3) Fusing (possibly confounded) data; 4) Incorporating causal assumptions or knowledge from humans. 5) Providing causality-based explanations; 6) Generalization and knowledge transfer; 7) Learning causal world models; 8) Counterfactual data generation; 9) Planning with world models; 10) Enhanced training of policies and value functions with causal reasoning.

Knowledge Transfer

RQ2

What knowledge can be transferred?

How can algorithms be designed to facilitate efficient adaptation?

Knowledge Transfer

RQ2

What knowledge can be transferred?

How can algorithms be designed to facilitate efficient adaptation?

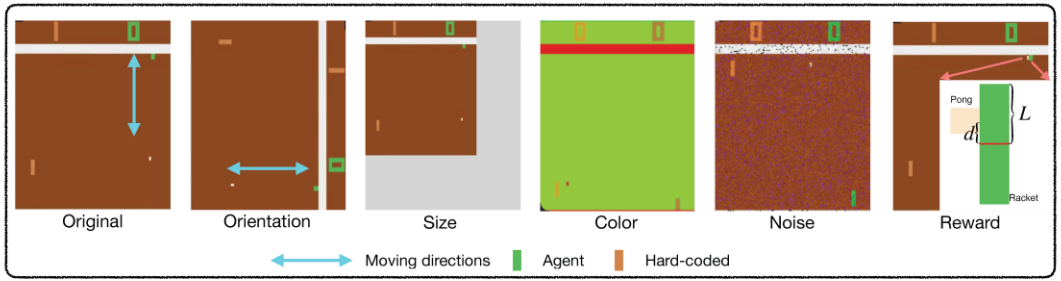


Figure A8: Visual example of the original Pong game and the various change factors. The light blue arrows are added to show the direction in which the agent can move.

For more information, check "AdaRL: What, Where, And How to Adapt in Transfer Reinforcement Learning", ICLR 2022

Knowledge Transfer

RQ2

What knowledge can be transferred?

How can algorithms be designed to facilitate efficient adaptation?

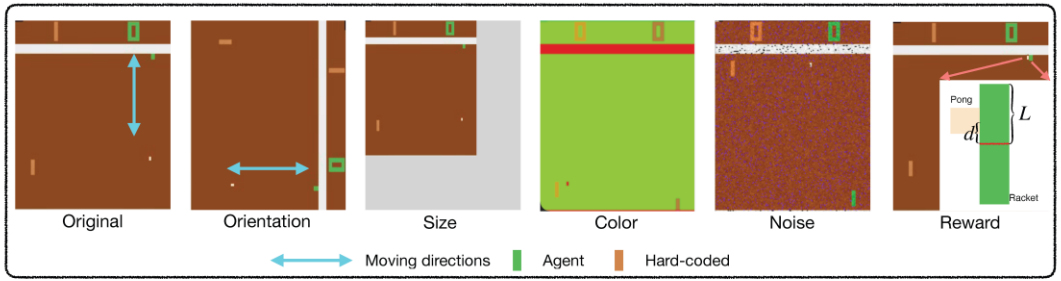


Figure A8: Visual example of the original Pong game and the various change factors. The light blue arrows are added to show the direction in which the agent can move.

For more information, check "AdaRL: What, Where, And How to Adapt in Transfer Reinforcement Learning", ICLR 2022

How to adapt reliably and efficiently to changes across domains with a few samples from the target domain, even in partially observable environments?

Knowledge Transfer

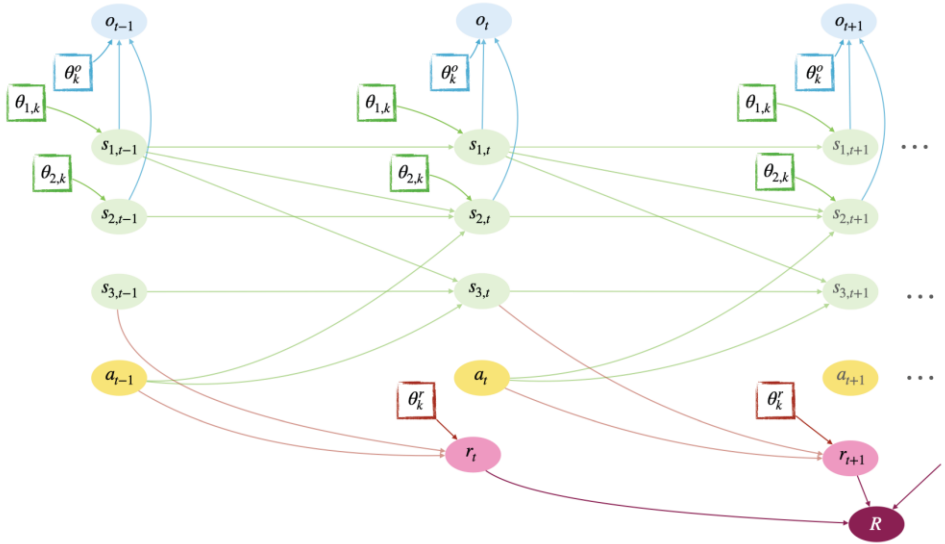
Sparse Mechanisms Shift

For more information, check "AdaRL: What, Where, And How to Adapt in Transfer Reinforcement Learning", ICLR 2022

RQ2

What knowledge can be transferred?

How can algorithms be designed to facilitate efficient adaptation?



How to adapt reliably and efficiently to changes across domains with a few samples from the target domain, even in partially observable environments?

Knowledge Transfer

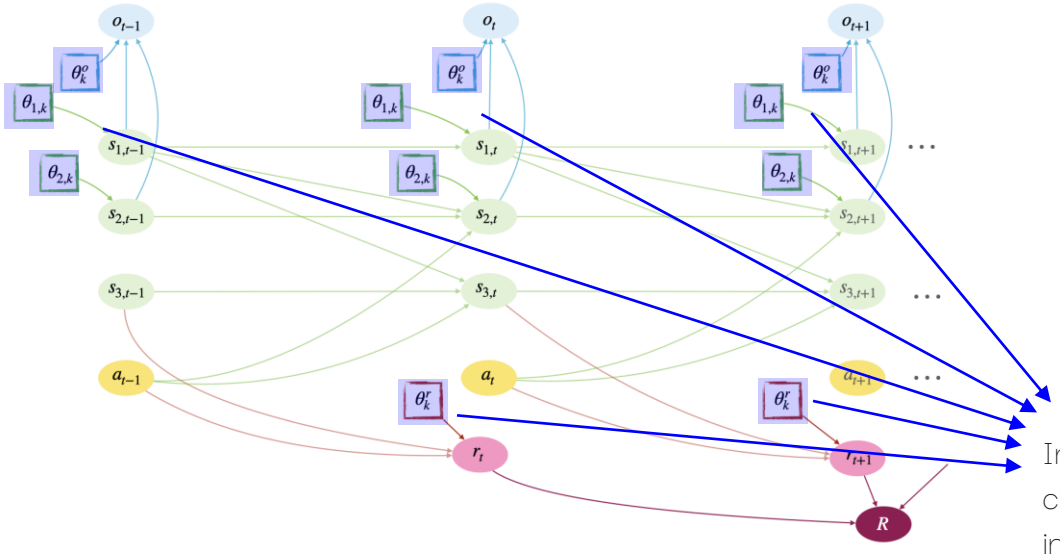
Sparse Mechanisms Shift

For more information, check "AdaRL: What, Where, And How to Adapt in Transfer Reinforcement Learning", ICLR 2022

RQ2

What knowledge can be transferred?

How can algorithms be designed to facilitate efficient adaptation?



How to adapt reliably and efficiently to changes across domains with a few samples from the target domain, even in partially observable environments?

Introduce a low-dimensional vector $\boldsymbol{\theta_k}$ to characterize the domain-specific information in a compact way.

Knowledge Transfer

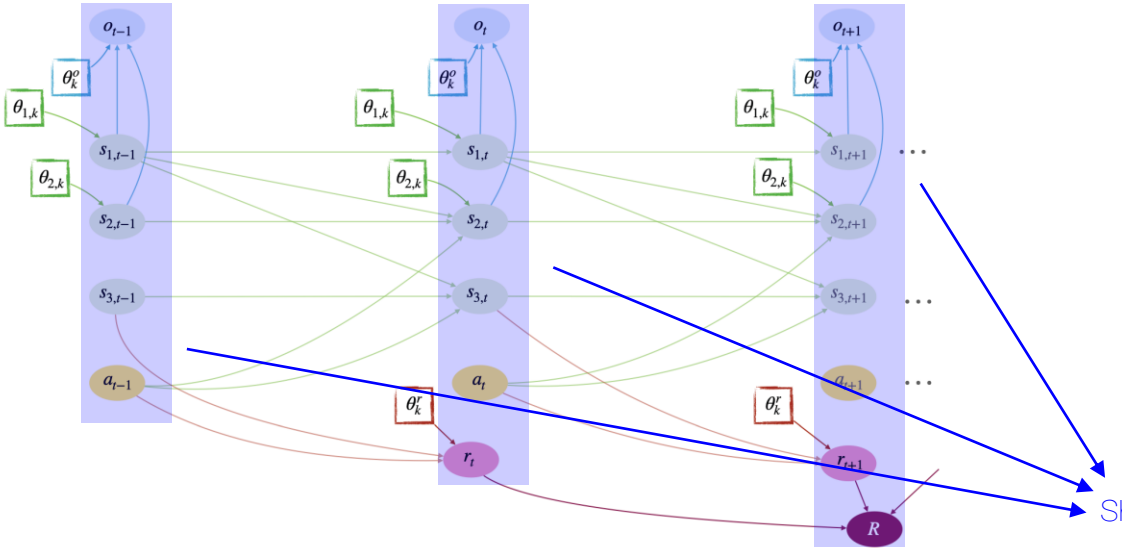
Sparse Mechanisms Shift

For more information, check "AdaRL: What, Where, And How to Adapt in Transfer Reinforcement Learning", ICLR 2022

RQ2

What knowledge can be transferred?

How can algorithms be designed to facilitate efficient adaptation?



How to adapt reliably and efficiently to changes across domains with a few samples from the target domain, even in partially observable environments?

Shared across different domains.

Knowledge Transfer

Sparse Mechanisms Shift

For more information, check "AdaRL: What, Where, And How to Adapt in Transfer Reinforcement Learning", ICLR 2022

RQ2

What knowledge can be transferred?

How can algorithms be designed to facilitate efficient adaptation?

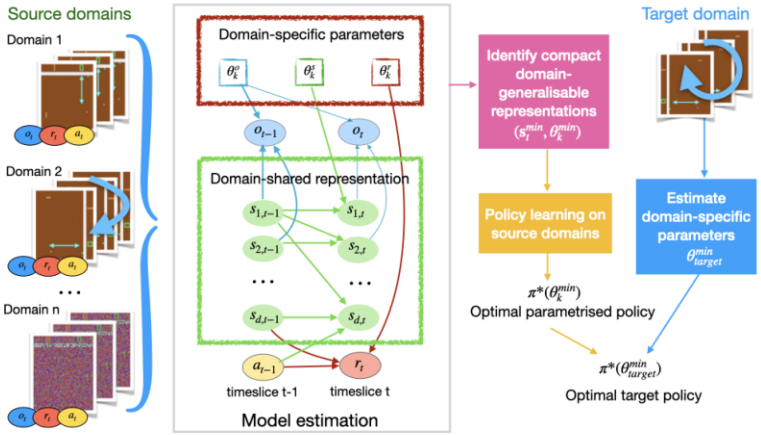


Figure 1: The overall AdaRL framework. We learn a Dynamic Bayesian Network (DBN) over the observations, latent states, reward, actions and domain-specific change factors that is shared across the domains. We then characterize a minimal set of representations that suffice for policy transfer, so that we can quickly adapt the optimal source policy with only a few samples from the target domain.

How to adapt reliably and efficiently to changes across domains with a few samples from the target domain, even in partially observable environments?

All we need to update in the target domain is the low-dimensional θ_k.

Knowledge Transfer

Sparse Mechanisms Shift

For more information, check "AdaRL: What, Where, And How to Adapt in Transfer Reinforcement Learning", ICLR 2022

RQ2

What knowledge can be transferred?

How can algorithms be designed to facilitate efficient adaptation?

	Oracle Upper bound	Non-t lower bound	PNN (Rusu et al., 2016)	PSM (Agarwal et al., 2021)	MTQ (Fakoor et al., 2020)	AdaRL* Ours w/o masks	AdaRL Ours
O_in	18.65 (±2.43)	6.18 ● (±2.43)	9.70 ● (±2.09)	11.61 ● (±3.85)	15.79 ● (±3.26)	14.27 ● (±1.93)	**18.97** (±2.00)
O_out	19.86 (±1.09)	6.40 ● (±3.17)	9.54 ● (±2.78)	10.82 ● (±3.29)	10.82 ● (±4.13)	12.67 ● (±2.49)	**15.75** (±3.80)
C_in	19.35 (±0.45)	8.53 ● (±2.08)	14.44 ● (±2.37)	19.02 (±1.17)	16.97 ● (±2.02)	18.52 ● (±1.41)	19.14 (±1.05)
C_out	19.78 (±0.25)	8.26 ● (±3.45)	14.84 ● (±1.98)	17.66 ● (±2.46)	15.45 ● (±3.30)	17.92 (±1.83)	19.03 (±0.97)
S_in	18.32 (±1.18)	6.91 ● (±2.02)	11.80 ● (±3.25)	12.65 ● (±3.72)	13.68 ● (±3.49)	14.23 ● (±3.19)	**16.65** (±1.72)
S_out	19.01 (±1.04)	6.60 ● (±3.11)	9.07 ● (±4.58)	8.45 ● (±4.51)	11.45 ● (±2.46)	12.80 ● (±2.62)	**17.82** (±2.35)
N_in	18.48 (±1.25)	5.51 ● (±3.88)	12.73 ● (±3.67)	11.30 ● (±2.58)	12.67 ● (±3.84)	13.78 ● (±2.15)	**16.84** (±3.13)
N_out	18.26 (±1.11)	6.02 ● (±3.19)	13.24 ● (±2.55)	11.26 ● (±3.15)	15.77 ● (±2.12)	14.65 ● (±3.01)	**18.30** (±2.24)

Table 3: Average final scores on modified Pong (POMDP) with $N_{target} = 50$. The best non-oracle are marked in **red**. O, C, S, and N denote the orientation, color, size, and noise factors, respectively.

How to adapt reliably and efficiently to changes across domains with a few samples from the target domain, even in partially observable environments?

Causal RL

For more information, check "Causal Reinforcement Learning: A Survey", TMLR 2023

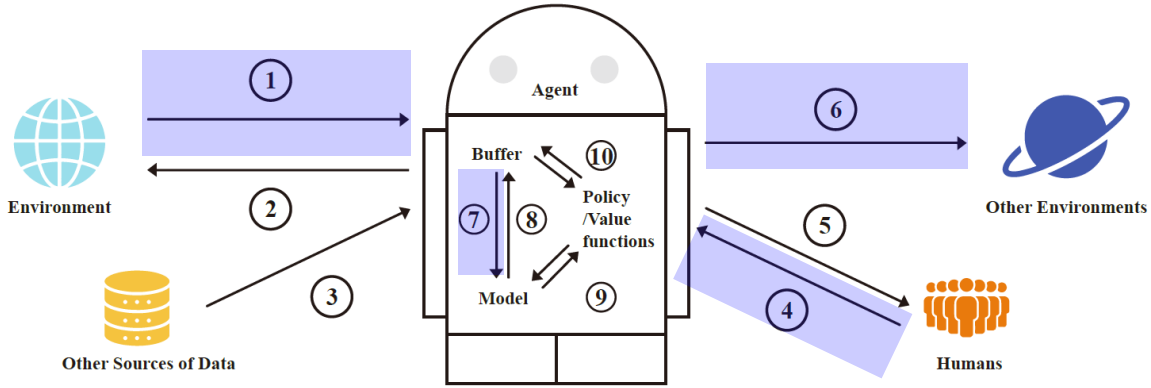


Figure 10: A schematic diagram illustrating the integration of causality into the reinforcement learning process. The numbered edges represent some key components: 1) Abstraction and extraction of causal representations from raw observations; 2) Directed exploration guided by causal knowledge; 3) Fusing (possibly confounded) data; 4) Incorporating causal assumptions or knowledge from humans. 5) Providing causality-based explanations; 6) Generalization and knowledge transfer; 7) Learning causal world models; 8) Counterfactual data generation; 9) Planning with world models; 10) Enhanced training of policies and value functions with causal reasoning.

Tutorial Outline

Part 1

- [x] Introduction
- [x] Causality Background
- [x] Reinforcement Learning Background

Part 2

- [x] Sample Efficiency
- [x] Generalization Ability
- [ ] Reliability

Reliability

[Submitted on 10 May 2024 (v1), last revised 17 May 2024 (this version, v2)]

Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems

David "davidad" Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Alessandro Abate, Joe Halpern, Clark Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing, Joshua Tenenbaum

"Ensuring that AI systems reliably and robustly avoid harmful or dangerous behaviours is a crucial challenge, especially for AI systems with a high degree of autonomy and general intelligence, or systems used in safety-critical contexts. In this paper, we will introduce and define a family of approaches to AI safety, which we will refer to as guaranteed safe (GS) AI. The core feature of these approaches is that they aim to produce AI systems which are equipped with high-assurance quantitative safety guarantees. This is achieved by the interplay of three core components: a world model (which provides a mathematical description of how the AI system affects the outside world), a safety specification (which is a mathematical description of what effects are acceptable), and a verifier (which provides an auditable proof certificate that the AI satisfies the safety specification relative to the world model). We outline a number of approaches for creating each of these three core components, describe the main technical challenges, and suggest a number of potential solutions to them. We also argue for the necessity of this approach to AI safety, and for the inadequacy of the main alternative approaches."

Reliability

”Human drivers are ridiculously reliable. The US has around one traffic fatality per 100 million miles driven; if a human driver makes 100 decisions per mile, that gets you a worst-case reliability of ~1:10,000,000,000 or ~99.999999999%. That’s around five orders of magnitude better than a very good deep learning model, and you get that even in an open environment, where data isn’t pre-filtered and there are sometimes random mechanical failures. Matching that bar is hard! I’m sure future AI will get there, but each additional ”nine” of reliability is typically another unit of engineering effort. (Note that current self-driving systems use a mix of different models embedded in a larger framework, not one model trained end-to-end like GPT-3.)“

“Humans are very reliable agents”
-- Alyssa Vance

Reliability

”Human drivers are ridiculously reliable. The US has around one traffic fatality per 100 million miles driven; if a human driver makes 100 decisions per mile, that gets you a worst-case reliability of ~1:10,000,000,000 or ~99.999999999%. That’s around five orders of magnitude better than a very good deep learning model, and you get that even in an open environment, where data isn’t pre-filtered and there are sometimes random mechanical failures. Matching that bar is hard! I’m sure future AI will get there, but each additional ”nine” of reliability is typically another unit of engineering effort. (Note that current self-driving systems use a mix of different models embedded in a larger framework, not one model trained end-to-end like GPT-3.)“

“Humans are very reliable agents”
-- Alyssa Vance

How do humans guarantee high reliability in decision-making?

Reliable human decision-making involves understanding and utilizing causality.

Reliable human decision-making involves understanding and utilizing causality.

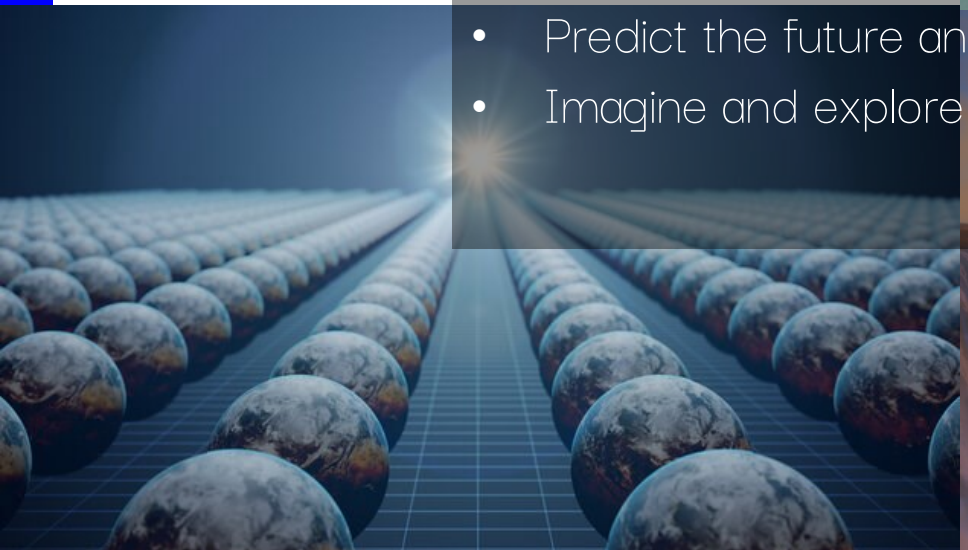
- Understand the consequences of actions

Reliable human decision-making involves understanding and utilizing causality.

- Understand the consequences of actions
- Predict the future and plan accordingly

Reliable human decision-making involves understanding and utilizing causality.

- Understand the consequences of actions
- Predict the future and plan accordingly
- Imagine and explore parallel worlds



Reliable human decision-making involves understanding and utilizing causality.

- Understand the consequences of actions
- Predict the future and plan accordingly
- Imagine and explore parallel worlds
- Reflect on and learn from the past

Reliability

RQ1

What is the connection between explainability and causality?

How to make the agent's understanding of the world more transparent?

RQ2

What does fairness mean in the context of decision-making?

How to build fairness-aware agents that can promote long-term fairness?

RQ3

What types of spurious correlations might exist in RL?

How to train agents to be robust against spurious correlations?

Explainability

RQ1

What is the connection between explainability and causality?

How to make the agent's understanding of the world more transparent?

Explainability

RQ1

What is the connection between explainability and causality?

How to make the agent's understanding of the world more transparent?

Explainability refers to the ability to explain or present the behavior of models in human-understandable terms.

For more information, check "Towards A Rigorous Science of Interpretable Machine Learning", 2017

For more information, check "Towards A Rigorous Science of Interpretable Machine Learning", 2017

Explainability

RQ1 What is the connection between explainability and causality?

How to make the agent's understanding of the world more transparent?

Explainability refers to the ability to explain or present

the behavior of models in — What to explain

human-understandable terms. — How to explain

For more information, check "Towards A Rigorous Science of Interpretable Machine Learning", 2017

Explainability

RQ1 What is the connection between explainability and causality?

How to make the agent's understanding of the world more transparent?

Explainability refers to the ability to explain or present *the behavior of models* in *human-understandable terms*.

Answer the question of 'Why'

What to explain

How to explain

For more information, check "Towards A Rigorous Science of Interpretable Machine Learning", 2017

Explainability

RQ1 What is the connection between explainability and causality?

How to make the agent's understanding of the world more transparent?

Explainability refers to the ability to explain or present *the behavior of models* in *human-understandable terms*.

Answer the question of 'Why'

What to explain

How to explain

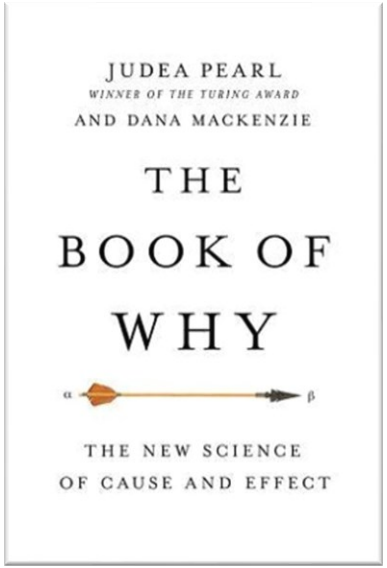
For more information, check
“The Book of Why”, 2018

Explainability

RQ1 What is the connection between explainability and causality?

How to make the agent’s understanding of the world more transparent?

Judea Pearl

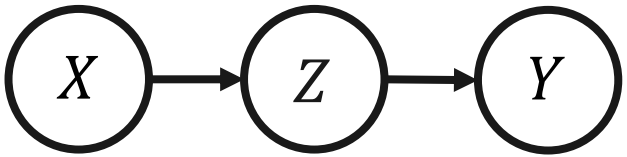


Answer the question of ‘Why’

Causality and Explainability

RQ1 What is the connection between explainability and causality?

How to make the agent's understanding of the world more transparent?



Consumption of citrus fruits — intake of vitamin C — occurrence of scurvy

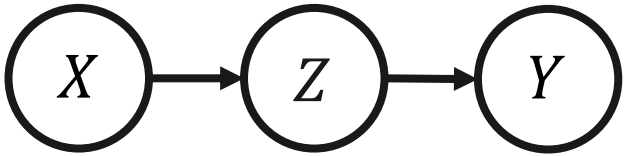
Causal Graph $\mathcal{G}$

- X causally affects Y via the mediator Z.
- Y is independent of X given Z.
- Changing Y will not affect X.
- …

Causality and Explainability

RQ1 What is the connection between explainability and causality?

How to make the agent's understanding of the world more transparent?



Consumption of citrus fruits — intake of vitamin C — occurrence of scurvy

Causal Graph $\mathcal{G}$

- X causally affects Y via the mediator Z.
- Y is independent of X given Z.
- Changing Y will not affect X.
- …

$$f_X: X = U_X$$

$$f_Z: Z = a \cdot X + U_Z$$

$$f_Y: Y = b \cdot Z + U_Y$$

Causal Model $\mathbf{F} = \{f_X, f_Z, f_Y\}$

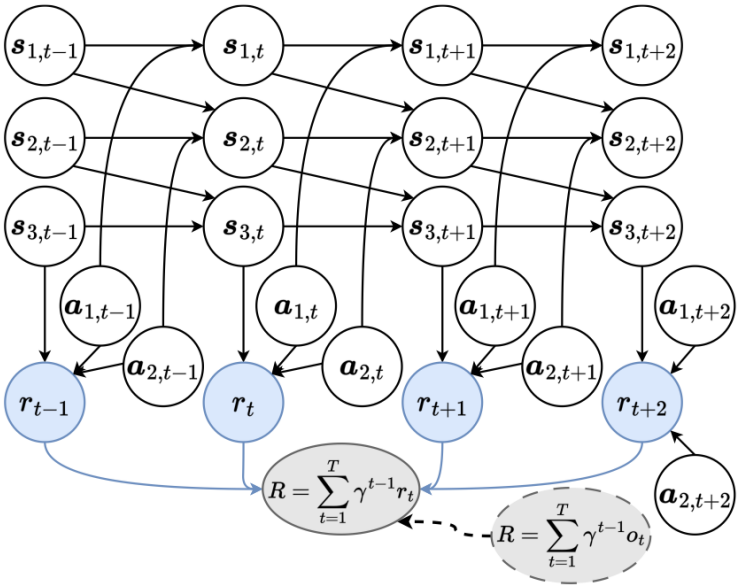
- Suppose the event $X = x, Y = y$ is observed. Would $Y = y'$ if X were x'? (Necessity)
- Suppose the event $X = x', Y = y'$ is observed. Would $Y = y$ if X were x? (Sufficiency)
- …

For more information, check "Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach", NeurIPS 2023

Explainability

RQ1 What is the connection between explainability and causality?

How to make the agent's understanding of the world more transparent?



The causal relationship among environmental variables.

The nodes denote different variables in the MDP environment, i.e., all dimensions of state $s_{\cdot,t}$ and action $a_{\cdot,t}$, Markovian rewards r_t for $t \in [1, T]$, as well as the long-term return R.

For sparse reward settings in RL, the Markovian rewards r_t are unobservable, which are represented by nodes with blue filling.

We can observe the trajectory-wise long-term return, R, which equals the discounted sum of delayed reward o_t and evaluates the performance of the agent within the whole episode.

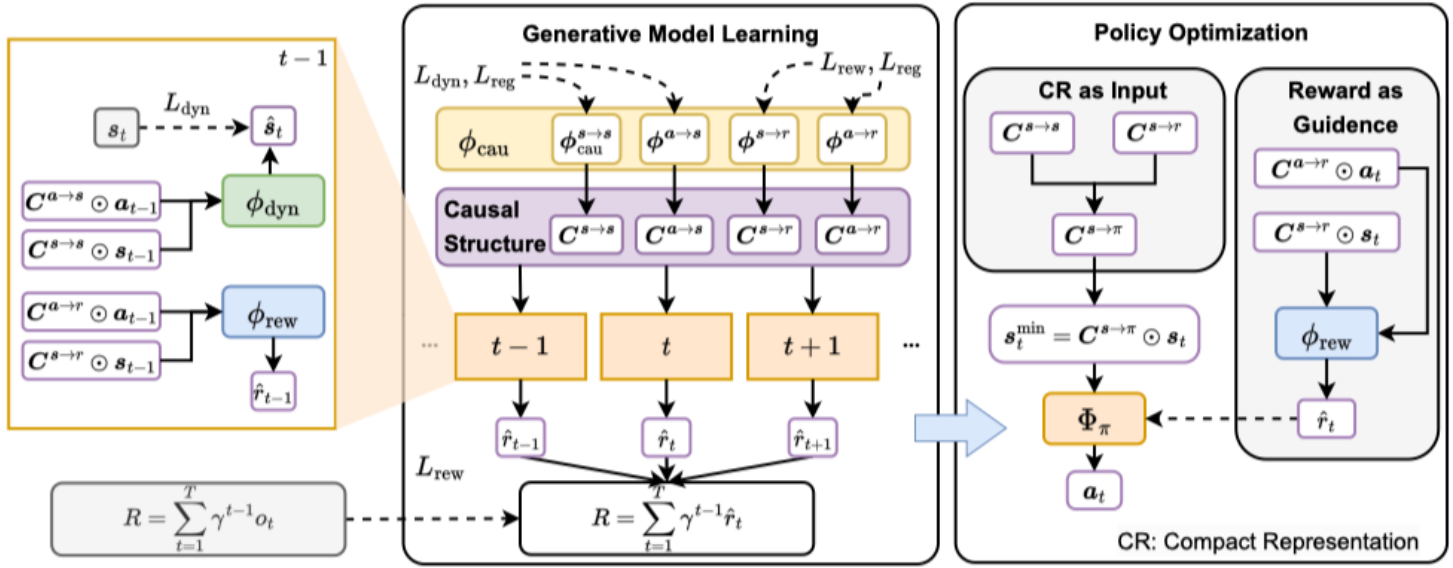
Explainability

For more information, check "Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach", NeurIPS 2023

RQ1

What is the connection between explainability and causality?

How to make the agent's understanding of the world more transparent?



The framework of the proposed method.

Explainability

For more information, check "Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach", NeurIPS 2023

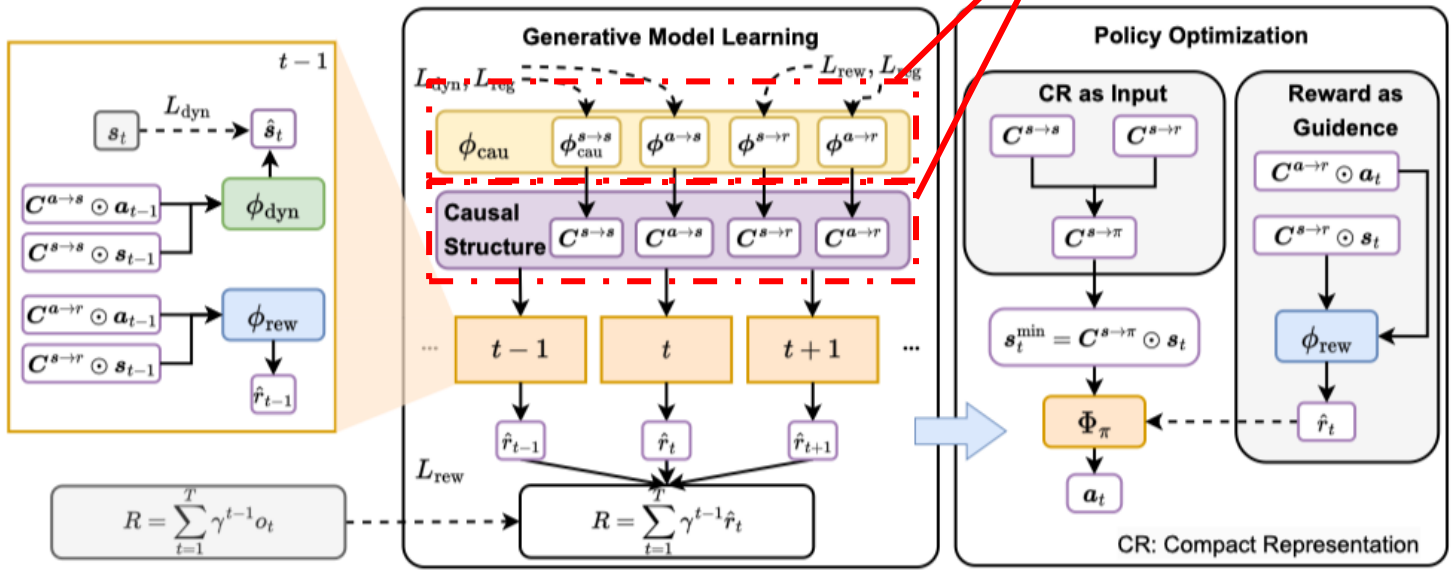
RQ1

What is the connection between explainability and causality?

How to make the agent's understanding of the world more transparent?

Learned environmental dynamics

Explainable causal structure



The framework of the proposed method.

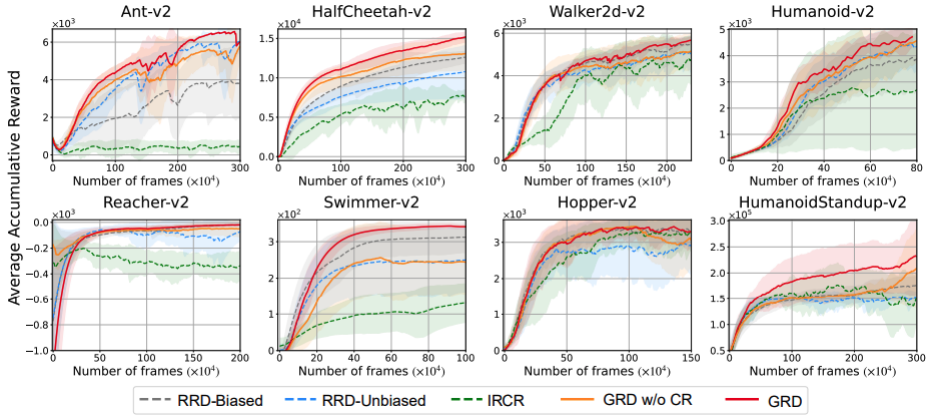
Interpretability

For more information, check "Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach", NeurIPS 2023

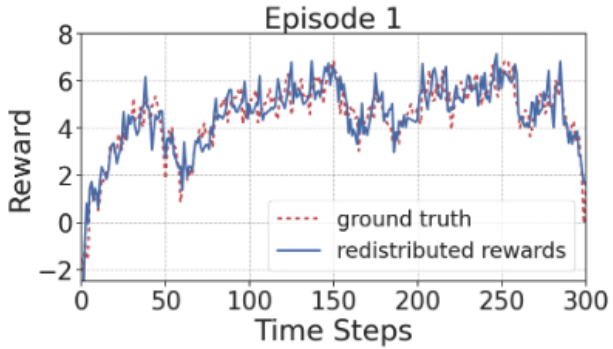
RQ1

What is the connection between explainability and causality?

How to make the agent's understanding of the world more transparent?



Learning curves on a suite of MuJoCo benchmark tasks.



Visualization of Decomposed Rewards (blue) and Ground Truth Rewards (red).

Causal RL

For more information, check "Causal Reinforcement Learning: A Survey", TMLR 2023

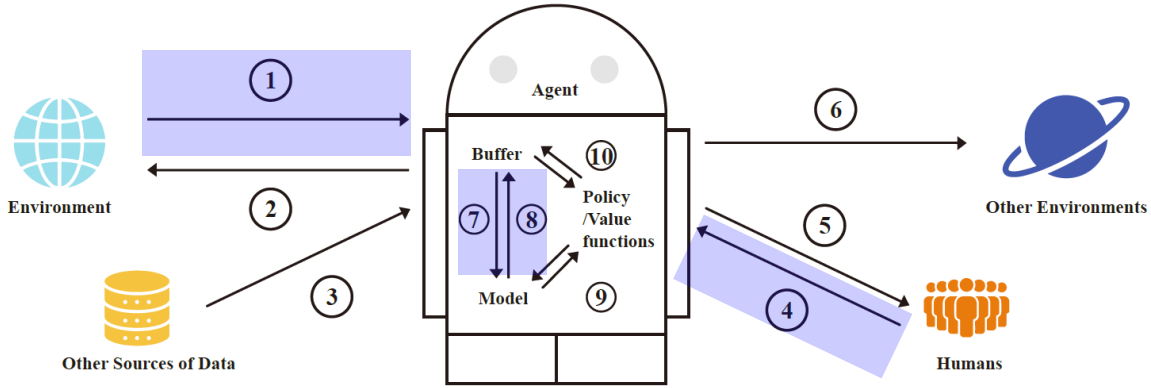


Figure 10: A schematic diagram illustrating the integration of causality into the reinforcement learning process. The numbered edges represent some key components: 1) Abstraction and extraction of causal representations from raw observations; 2) Directed exploration guided by causal knowledge; 3) Fusing (possibly confounded) data; 4) Incorporating causal assumptions or knowledge from humans. 5) Providing causality-based explanations; 6) Generalization and knowledge transfer; 7) Learning causal world models; 8) Counterfactual data generation; 9) Planning with world models; 10) Enhanced training of policies and value functions with causal reasoning.

Fairness

RQ2

What does fairness mean in the context of decision-making?

How to build fairness-aware agents that can promote long-term fairness?

Fairness

For more information, check "What Hides behind Unfairness? Exploring Dynamics Fairness in Reinforcement Learning", IJCAI 2024

RQ2

What does fairness mean in the context of decision-making?

How to build fairness-aware agents that can promote long-term fairness?

Yann Lecun

People are biased.

Data is biased, in part because people are biased.

Algorithms trained on biased data are biased.

Fairness

For more information, check "What Hides behind Unfairness? Exploring Dynamics Fairness in Reinforcement Learning", IJCAI 2024

RQ2 What does fairness mean in the context of decision-making?

How to build fairness-aware agents that can promote long-term fairness?

Yann Lecun

People are biased.

Data is biased, in part because people are biased.

Algorithms trained on biased data are biased.

Agents trained in biased environments are biased.

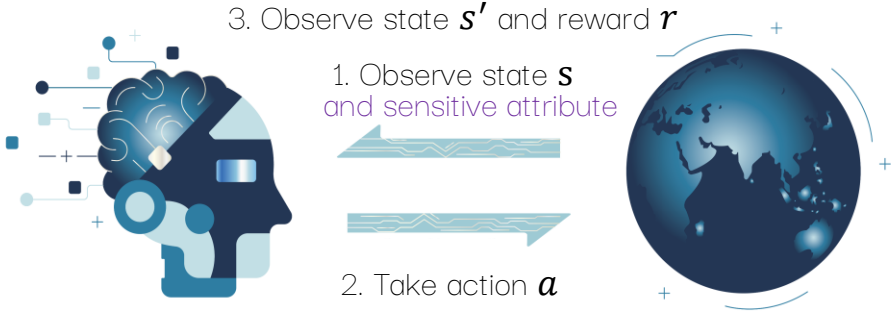
Fairness

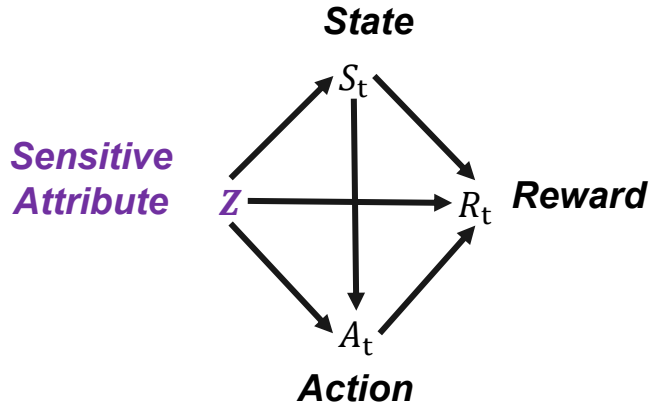
For more information, check "What Hides behind Unfairness? Exploring Dynamics Fairness in Reinforcement Learning", IJCAI 2024

RQ2

What does fairness mean in the context of decision-making?

How to build fairness-aware agents that can promote long-term fairness?



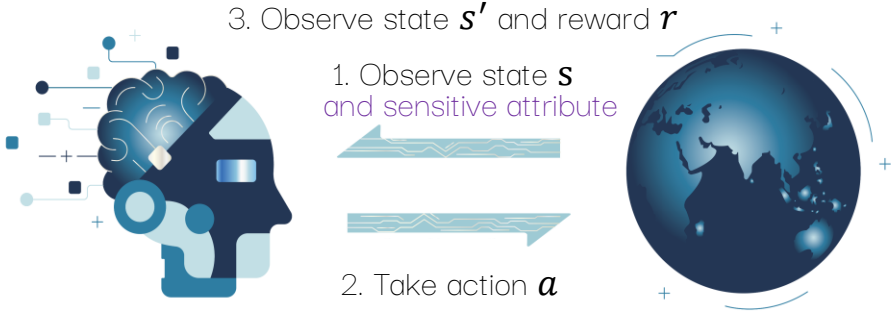


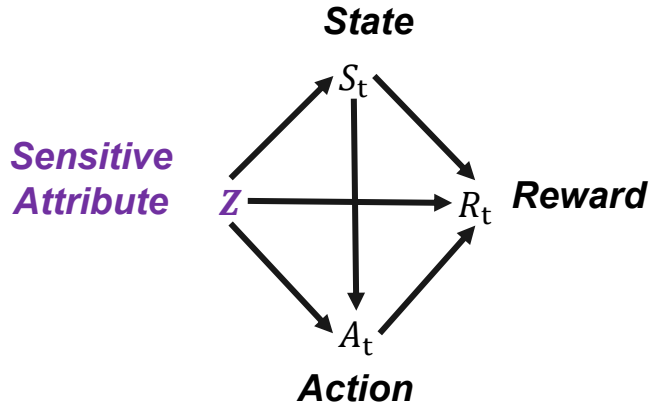
For more information, check "What Hides behind Unfairness? Exploring Dynamics Fairness in Reinforcement Learning", IJCAI 2024

Fairness

RQ2 What does fairness mean in the context of decision-making?

How to build fairness-aware agents that can promote long-term fairness?





"How a change in sensitive attribute would affect a group's future return G_t (cumulative rewards) if all other factors were held constant?", measured by the total effect of the sensitive attribute Z on future return G_t:

$$\mathrm{TE}_{z_0,z_1}(G_t) \coloneqq \mathrm{E}[G_t(z_1)] - \mathrm{E}[G_t(z_0)]$$

"The Well-Being Gap"

Fairness

For more information, check "What Hides behind Unfairness? Exploring Dynamics Fairness in Reinforcement Learning", IJCAI 2024

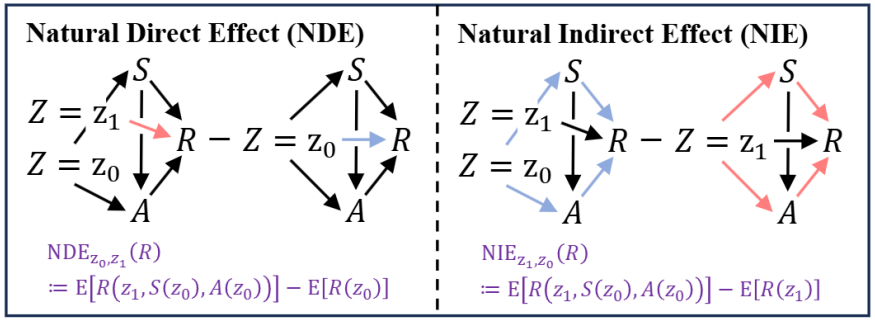
RQ2

What does fairness mean in the context of decision-making?

How to build fairness-aware agents that can promote long-term fairness?

Theorem 1 (Causal Decomposition of Well-being Gap)**.** The well-being gap $\mathrm{TE}_{z_0,z_1}(G_t)$ can be decomposed as follows:

$$\mathrm{TE}_{z_0,z_1}(G_t) = \sum_{k=0}^{\infty} \gamma^k \left(\mathrm{NDE}_{z_0,z_1}(R) - \mathrm{NIE}_{z_1,z_0}(R)\right).$$



For more information, check "What Hides behind Unfairness? Exploring Dynamics Fairness in Reinforcement Learning", IJCAI 2024

Fairness

RQ2 What does fairness mean in the context of decision-making?

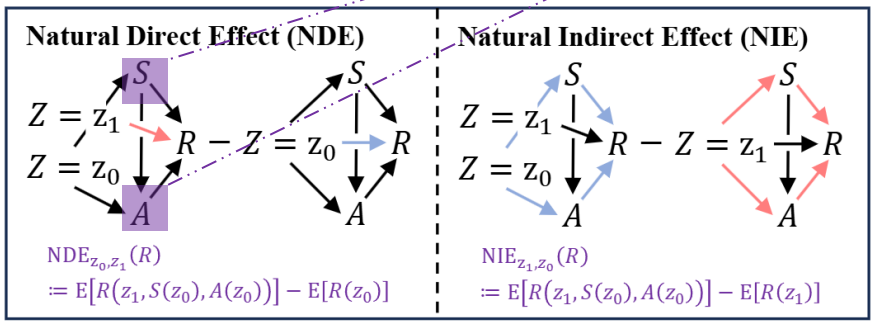
How to build fairness-aware agents that can promote long-term fairness?

Theorem 1 (Causal Decomposition of Well-being Gap). The well-being gap $\mathrm{TE}_{z_0,z_1}(G_t)$ can be decomposed as follows:

$$\mathrm{TE}_{z_0,z_1}(G_t) = \sum_{k=0}^{\infty} \gamma^k \left(\mathrm{NDE}_{z_0,z_1}(R) - \mathrm{NIE}_{z_1,z_0}(R)\right).$$

Mediators

Causal effects that does not go through any mediator.



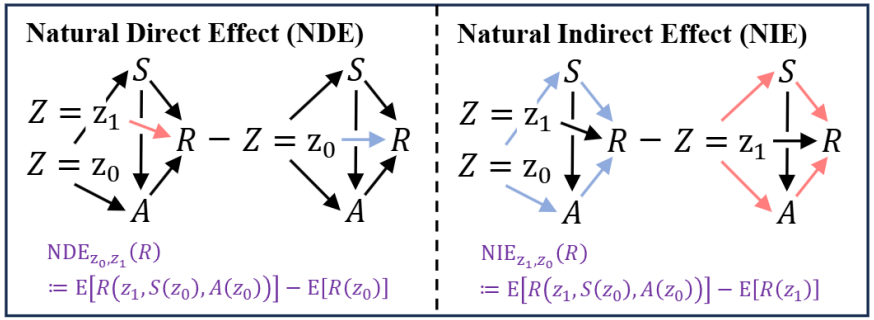
Causal effects that only go through mediators.

Fairness

For more information, check "What Hides behind Unfairness? Exploring Dynamics Fairness in Reinforcement Learning", IJCAI 2024

RQ2 What does fairness mean in the context of decision-making?

How to build fairness-aware agents that can promote long-term fairness?



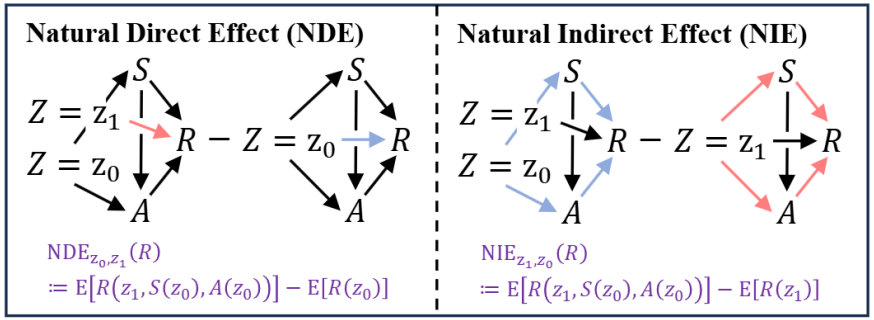
By keeping the state S and the decision A to the level they would gained under $Z = z_0$, $\mathrm{NDE}_{z_0,z_1}(R)$ measures the advantage that the individuals in the disadvantaged group $Z = z_0$ would have gained if there were in the advantaged group $Z = z_1$.

Fairness

For more information, check "What Hides behind Unfairness? Exploring Dynamics Fairness in Reinforcement Learning", IJCAI 2024

RQ2 What does fairness mean in the context of decision-making?

How to build fairness-aware agents that can promote long-term fairness?



By keeping the state S and the decision A to the level they would gained under $Z = z_0$, $\mathrm{NDE}_{z_0,z_1}(R)$ measures the advantage that the individuals in the disadvantaged group $Z = z_0$ would have gained if there were in the advantaged group $Z = z_1$.

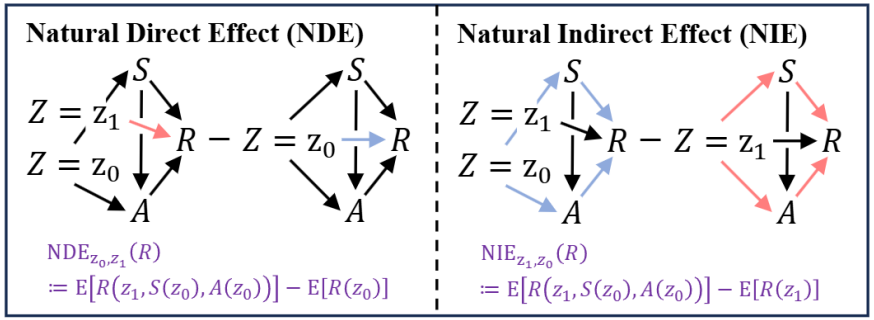
This is a nested counterfactual, a causal quantity that we can never observe in the factual world, simply because we cannot observe an individual that belongs to two different demographic groups at the same time.

For more information, check "What Hides behind Unfairness? Exploring Dynamics Fairness in Reinforcement Learning", IJCAI 2024

Fairness

RQ2 What does fairness mean in the context of decision-making?

How to build fairness-aware agents that can promote long-term fairness?



By keeping the state S and the decision A to the level they would gained under $Z = z_0$, $\mathrm{NDE}_{z_0,z_1}(R)$ measures the advantage that the individuals in the disadvantaged group $Z = z_0$ would have gained if there were in the advantaged group $Z = z_1$.

This is a nested counterfactual, a causal quantity that we can never observe in the factual world, simply because we cannot observe an individual that belongs to two different demographic groups at the same time.

Is it possible to estimate these quantities in practice?

Fairness

For more information, check "What Hides behind Unfairness? Exploring Dynamics Fairness in Reinforcement Learning", IJCAI 2024

RQ2

What does fairness mean in the context of decision-making?

How to build fairness-aware agents that can promote long-term fairness?

Theorem 3 (Identification of Dynamics Fairness)**.** Under the structural assumptions implied by the augmented dynamics model depicted by the local causal diagram, the natural direct effects of Z on R and S' can be identified as follows:

$$\text{NDE}_{z_0,z_1}(R) = \sum_{s,a} \big(\mathbb{E}[R|Z=z_1,S=s,A=a] - \mathbb{E}[R|Z=z_0,S=s,A=a]\big)P(S=s,A=a|Z=z_0).$$

$$\text{NDE}_{z_0,z_1}(S') = \sum_{s,a} \big(\mathbb{E}[S'|Z=z_1,S=s,A=a] - \mathbb{E}[S'|Z=z_0,S=s,A=a]\big)P(S=s,A=a|Z=z_0).$$

All quantities on the right-hand side are expressed using conditional expectations or probabilities rather than counterfactuals – they can be estimated from observational data using standard statistical methods! We can do this by training a model-based RL agent.

Fairness

For more information, check "What Hides behind Unfairness? Exploring Dynamics Fairness in Reinforcement Learning", IJCAI 2024

RQ2 What does fairness mean in the context of decision-making?

How to build fairness-aware agents that can promote long-term fairness?

Theorem 3 (Identification of Dynamics Fairness)**.** Under the structural assumptions implied by the augmented dynamics model depicted by the local causal diagram, the natural direct effects of Z on R and S' can be identified as follows:

$$\text{NDE}_{z_0,z_1}(R) = \sum_{s,a} \big(\mathbb{E}[R|Z=z_1,S=s,A=a] - \mathbb{E}[R|Z=z_0,S=s,A=a]\big)P(S=s,A=a|Z=z_0).$$

$$\text{NDE}_{z_0,z_1}(S') = \sum_{s,a} \big(\mathbb{E}[S'|Z=z_1,S=s,A=a] - \mathbb{E}[S'|Z=z_0,S=s,A=a]\big)P(S=s,A=a|Z=z_0).$$

All quantities on the right-hand side are expressed using conditional expectations or probabilities rather than counterfactuals – they can be estimated from observational data using standard statistical methods! We can do this by training a model-based RL agent.

If both $\text{NDE}_{z_0,z_1}(R)$ and $\text{NDE}_{z_0,z_1}(S')$ equal zero, then the environmental dynamics is considered fair because the sensitive attribute Z does not affect the reward assignment and state transition processes; otherwise, the agent needs to choose different actions for the groups to promote long-term fairness.

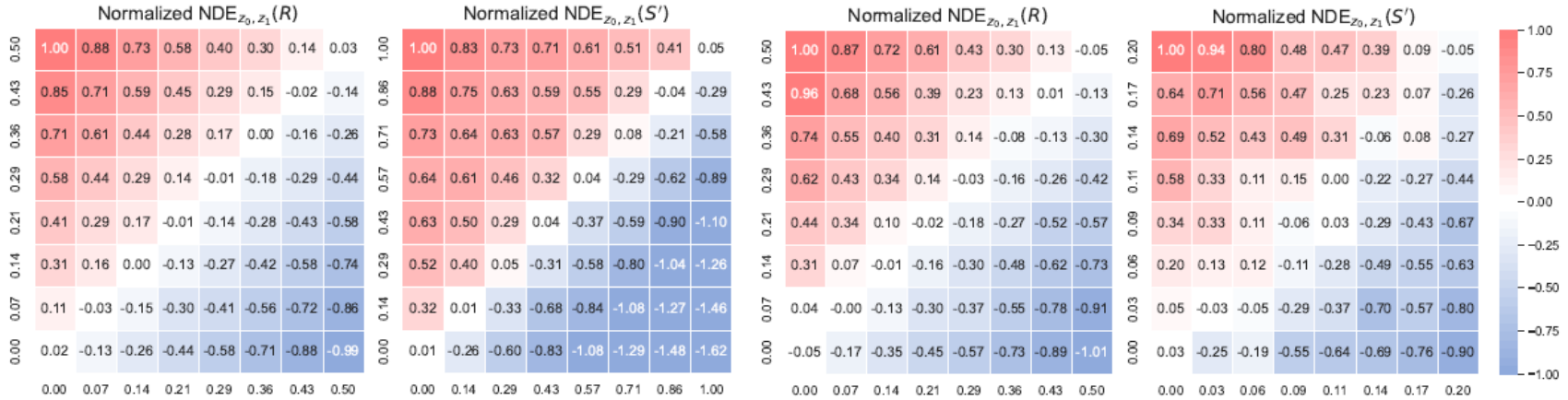
For more information, check "What Hides behind Unfairness? Exploring Dynamics Fairness in Reinforcement Learning", IJCAI 2024

Fairness

RQ2

What does fairness mean in the context of decision-making?

How to build fairness-aware agents that can promote long-term fairness?



Results of evaluating dynamics fairness using the proposed identification formula.

Each tile represents the estimated natural direct effects, $\mathrm{NDE}_{z_0,z_1}(R)$ or $\mathrm{NDE}_{z_0,z_1}(S')$, under specific parameter configurations indicated by the row and column. Lighter colors signify values closer to zero (satisfying dynamics fairness), while darker colors represent larger absolute values (violating dynamics fairness). Red denotes the second demographic group is advantaged, while blue denotes disadvantaged.

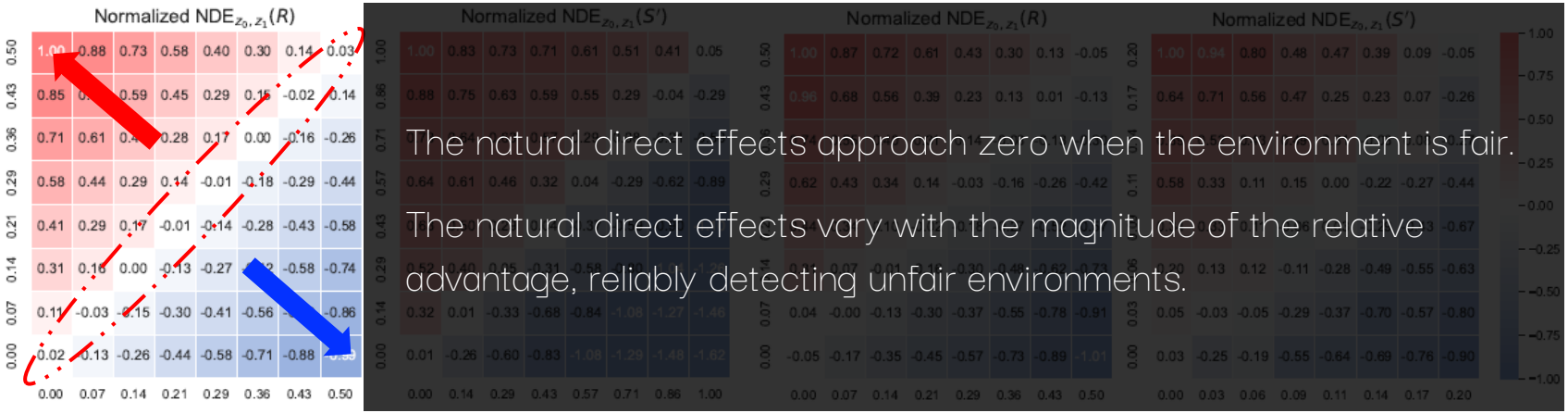
For more information, check "What Hides behind Unfairness? Exploring Dynamics Fairness in Reinforcement Learning", IJCAI 2024

Fairness

RQ2

What does fairness mean in the context of decision-making?

How to build fairness-aware agents that can promote long-term fairness?



Results of evaluating dynamics fairness using the proposed identification formula.

Each tile represents the estimated natural direct effects, $\mathrm{NDE}_{z_0,z_1}(R)$ or $\mathrm{NDE}_{z_0,z_1}(S')$, under specific parameter configurations indicated by the row and column. Lighter colors signify values closer to zero (satisfying dynamics fairness), while darker colors represent larger absolute values (violating dynamics fairness). Red denotes the second demographic group is advantaged, while blue denotes disadvantaged.

Causal RL

For more information, check "Causal Reinforcement Learning: A Survey", TMLR 2023

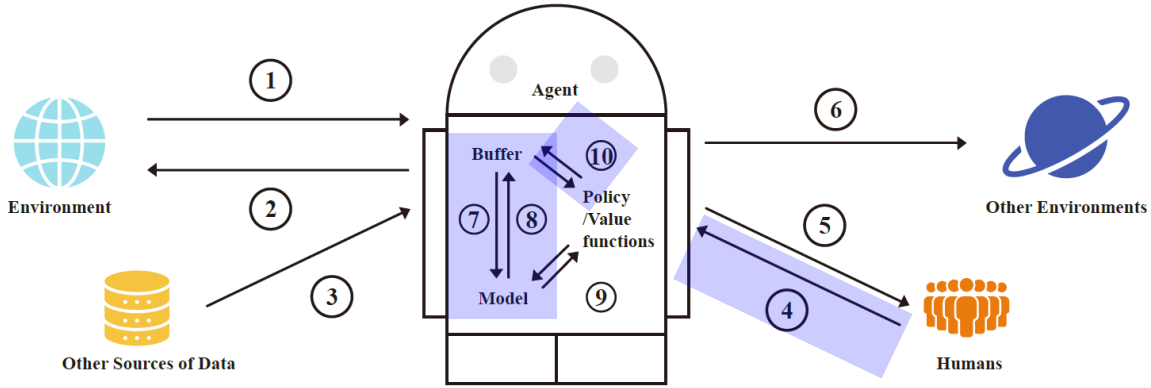


Figure 10: A schematic diagram illustrating the integration of causality into the reinforcement learning process. The numbered edges represent some key components: 1) Abstraction and extraction of causal representations from raw observations; 2) Directed exploration guided by causal knowledge; 3) Fusing (possibly confounded) data; 4) Incorporating causal assumptions or knowledge from humans. 5) Providing causality-based explanations; 6) Generalization and knowledge transfer; 7) Learning causal world models; 8) Counterfactual data generation; 9) Planning with world models; 10) Enhanced training of policies and value functions with causal reasoning.

Spurious Correlation

RQ3

What types of spurious correlations might exist in RL?

How to train agents to be robust against spurious correlations?

Correlation does not imply causation

Consuming rotten meat correlates with scurvy

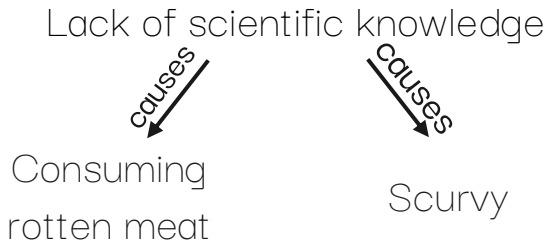
Lack of scientific knowledge

causes → Consuming rotten meat

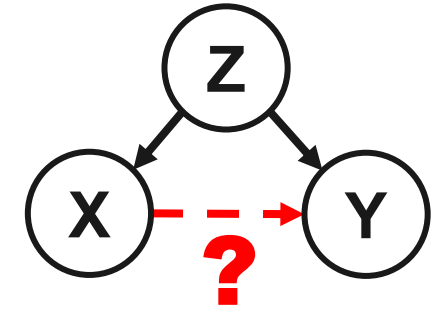
causes → Scurvy

Correlation does not imply causation

Consuming rotten meat ⟷ (correlates with) scurvy



Correlation does not imply causation



Confounding Bias

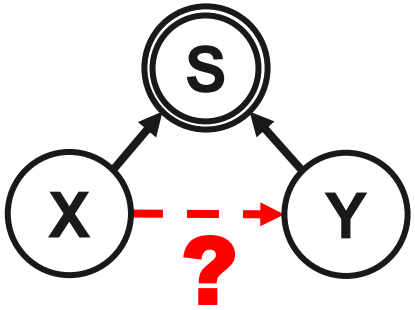
Lack of scientific knowledge

causes

causes

Consuming rotten meat

Scurvy



Selection Bias

Data selection

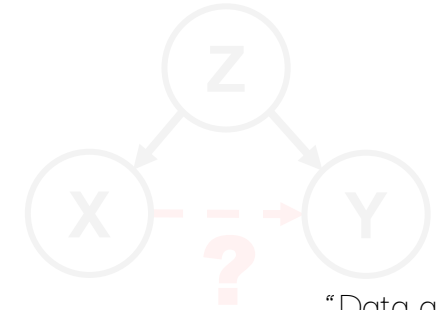
causes

causes

Consuming rotten meat

Scurvy

Correlation does not imply causation



Confounding Bias

"Data are profoundly dumb."

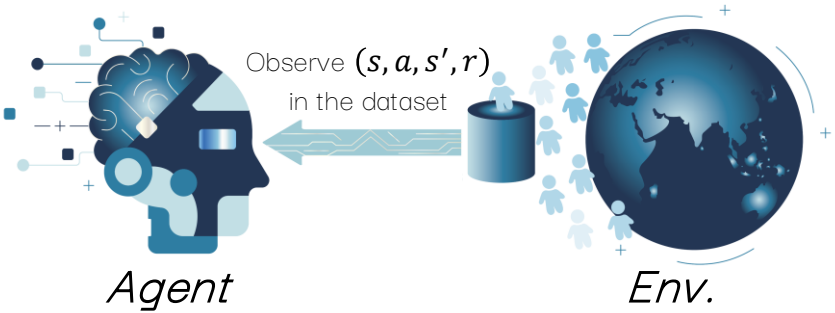
Judea Pearl

Selection Bias

Offline RL

RQ3 What types of spurious correlations might exist in RL?

How to train agents to be robust against spurious correlations?



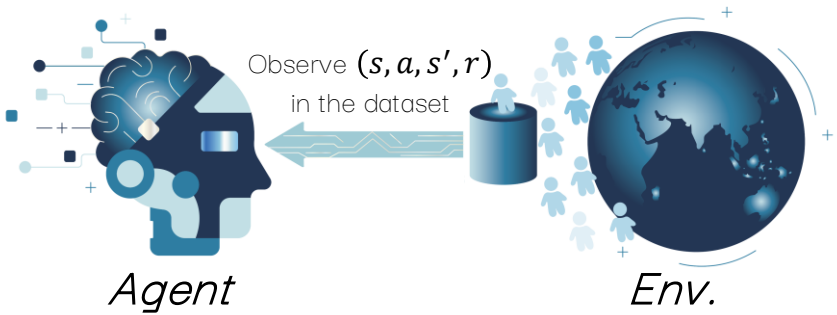
For more information, check "False Correlation Reduction for Offline Reinforcement Learning", TPAMI 2023

For more information, check "False Correlation Reduction for Offline Reinforcement Learning", TPAMI 2023

Offline RL

RQ3 What types of spurious correlations might exist in RL?

How to train agents to be robust against spurious correlations?



search the policy that maximizes the estimated return

$$\max_{\pi} V^{\pi}(s_0) = \sum_{t=0}^{\infty} E_{\pi}\left[\left\langle \hat{Q}(s_t, \cdot), \pi(\cdot\,|s_t)\right\rangle_A | s_0\right]$$

value function estimated from offline data

- Offline RL (Batch RL) is a promising approach to reuse existing experience and avoid inefficient online interactions that can be costly or even dangerous.
- The learned value function guides the agent's policy optimization process.

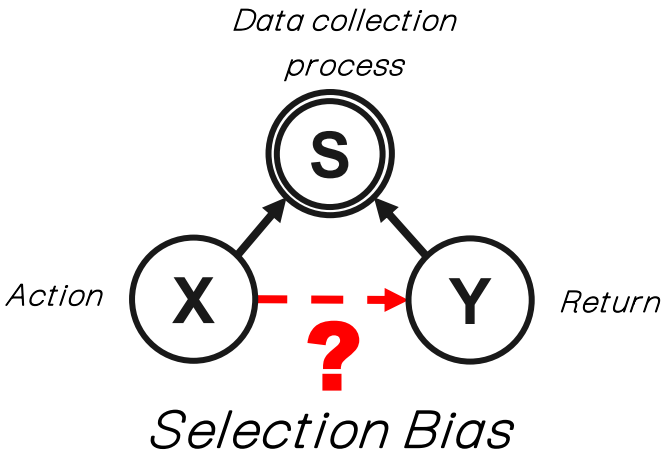
For more information, check "False Correlation Reduction for Offline Reinforcement Learning", TPAMI 2023

Spurious Correlation

Causal Graph

RQ3 What types of spurious correlations might exist in RL?

How to train agents to be robust against spurious correlations?



Does the observed correlation reflect that certain action causes the high return?

Or is it just a spurious correlation caused by selection bias?

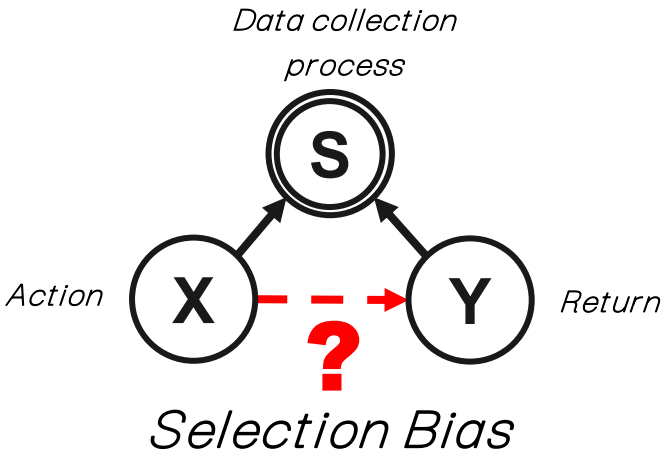
For more information, check "False Correlation Reduction for Offline Reinforcement Learning", TPAMI 2023

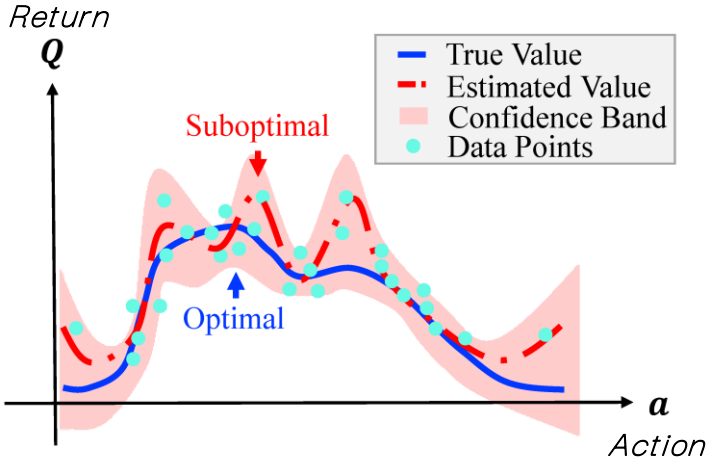
Spurious Correlation

Causal Graph

RQ3 What types of spurious correlations might exist in RL?

How to train agents to be robust against spurious correlations?





Does the observed correlation reflect that certain action causes the high return?

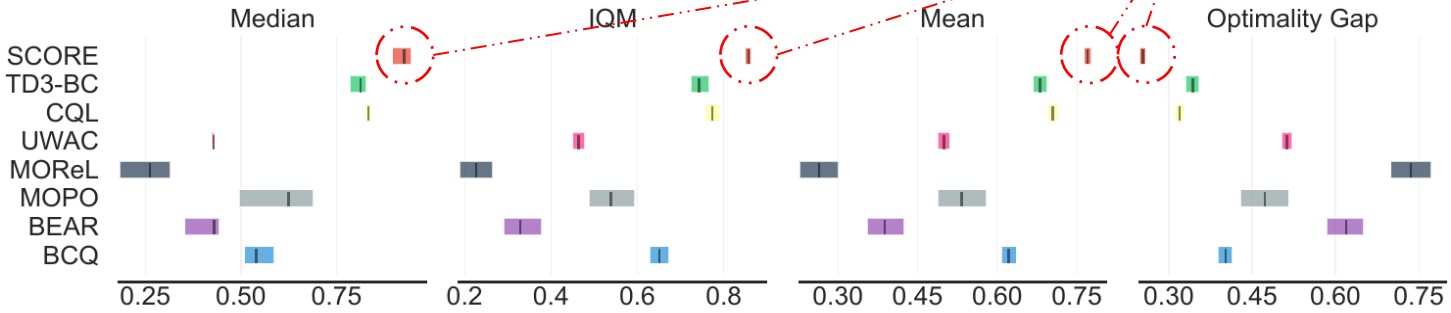
Or is it just a spurious correlation caused by selection bias?

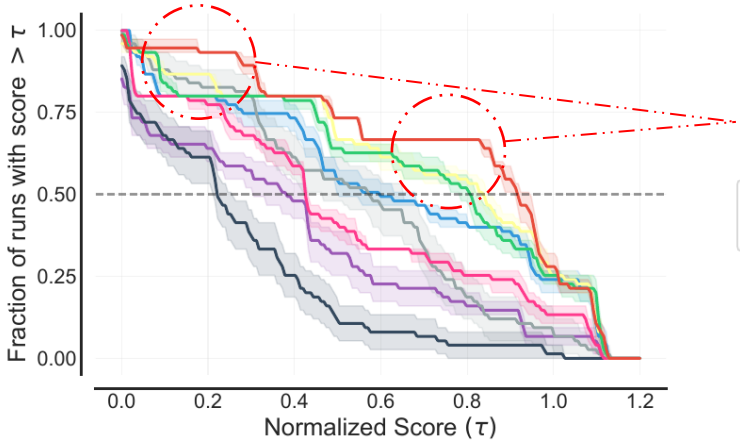
- Suboptimal actions may be overestimated due to epistemic uncertainty.
- Acting greedily w.r.t such estimations leads to suboptimal policies.

For more information, check "False Correlation Reduction for Offline Reinforcement Learning", TPAMI 2023

Spurious Correlation

best performance and stable over all runs





significant improvements on low- and medium-quality datasets

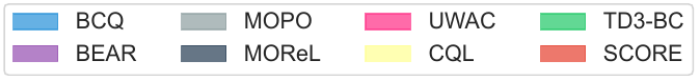


Figure 3. **Score distributions on D4RL-MuJoCo**. The distributions are computed using normalized scores obtained after training for 1M steps. Shaded regions show pointwise 95% percentile stratified bootstrap CIs.

For more information, check "Causal Reinforcement Learning: A Survey", TMLR 2023

Causal RL

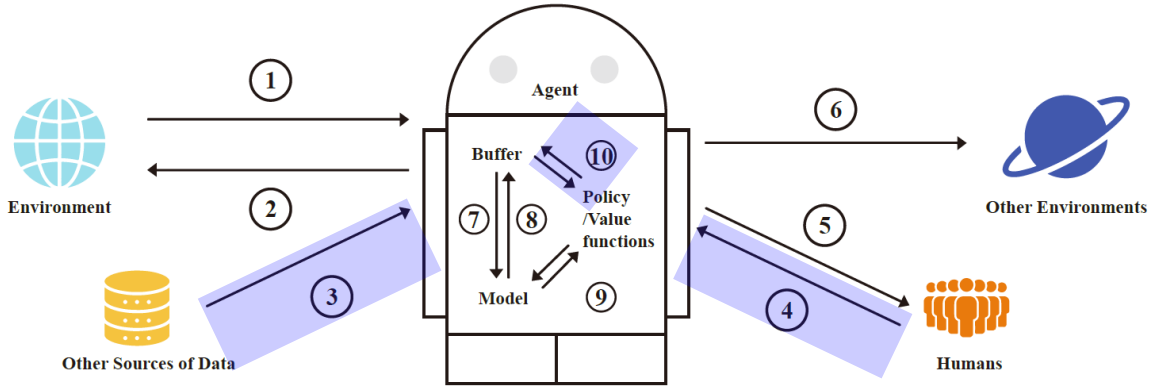


Figure 10: A schematic diagram illustrating the integration of causality into the reinforcement learning process. The numbered edges represent some key components: 1) Abstraction and extraction of causal representations from raw observations; 2) Directed exploration guided by causal knowledge; 3) Fusing (possibly confounded) data; 4) Incorporating causal assumptions or knowledge from humans. 5) Providing causality-based explanations; 6) Generalization and knowledge transfer; 7) Learning causal world models; 8) Counterfactual data generation; 9) Planning with world models; 10) Enhanced training of policies and value functions with causal reasoning.

Takeaway

1. Sample Efficiency

   👉 Explore the areas that agents can causally influence the environment.

   👉 Abstract away unnecessary details to simplify the learning problem.

   👉 Generate counterfactual rollouts for data augmentation.

2. Generalization Ability

   👉 Learn from the effects of different interventions to make the agent more generalizable..

   👉 Transfer the domain-invariant information and only adapt the changed causal mechanisms.

3. Reliability

   👉. Causal graphs and causal models provide explanations of different granularities.

   👉 Model the environmental dynamics and use mediation analysis to evaluate fairness.

   👉 Identify the factors that exhibit spurious correlations and address them.

Open Problems

1. Causal Learning in Reinforcement Learning
    - 👉 Automated causal discovery
    - 👉 Causal representation learning
    - 👉 Causal world model
2. Improved reasoning with causal knowledge
    - 👉 Planning with causal models
    - 👉 Imperfect knowledge
    - 👉 Cooperation and communication in multi-agent systems
3. Theoretical Advances
    - 👉 Identifiability, causal bounds, sensitivity analysis, and transportability
4. Real-world Applications
    - 👉 Robotics, healthcare, finance, self-driving, …

For more information, check "Causal Reinforcement Learning: A Survey", TMLR 2023

Causal Reinforcement Learning: A Survey

Zhihong Deng *zhi-hong.deng@student.uts.edu.au*
Australian Artificial Intelligence Institute, University of Technology Sydney

Jing Jiang *jing.jiang@uts.edu.au*
Australian Artificial Intelligence Institute, University of Technology Sydney

Guodong Long *guodong.long@uts.edu.auu*
Australian Artificial Intelligence Institute, University of Technology Sydney

Chengqi Zhang *chengqi.zhang@uts.edu.au*
Australian Artificial Intelligence Institute, University of Technology Sydney

Abstract

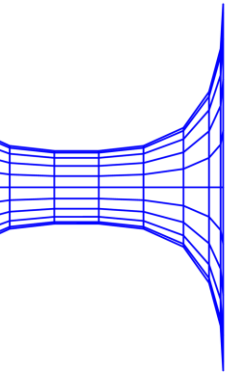
Reinforcement learning is an essential paradigm for solving sequential decision problems under uncertainty. Despite many remarkable achievements in recent decades, applying reinforcement learning methods in the real world remains challenging. One of the main obstacles is that reinforcement learning agents lack a fundamental understanding of the world and must therefore learn from scratch through numerous trial-and-error interactions. They may also face challenges in providing explanations for their decisions and generalizing the acquired knowledge. Causality, however, offers a notable advantage as it can formalize knowledge in a systematic manner and leverage invariance for effective knowledge transfer. This has led to the emergence of causal reinforcement learning, a subfield of reinforcement learning that seeks to enhance existing algorithms by incorporating causal relationships into the learning process. In this survey, we comprehensively review the literature on causal reinforcement learning. We first introduce the basic concepts of causality and reinforcement learning, and then explain how causality can address core challenges in non-causal reinforcement learning. We categorize and systematically review existing causal reinforcement learning approaches based on their target problems and methodologies. Finally, we outline open issues and future directions in this emerging field.

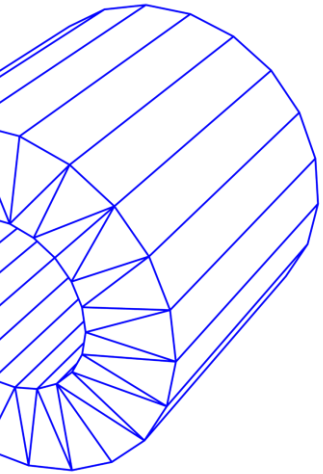
1 Introduction

"All reasonings concerning matter of fact seem to be founded on the relation of cause and effect. By means of that relation alone we can go beyond the evidence of our memory and senses."

—David Hume, An Enquiry Concerning Human Understanding.

https://arxiv.org/abs/2307.01452





Q & A

Thank you!

Zhihong Deng

Zhi-Hong.Deng@student.uts.edu.au