# Causal Reasoning from Meta-reinforcement Learning

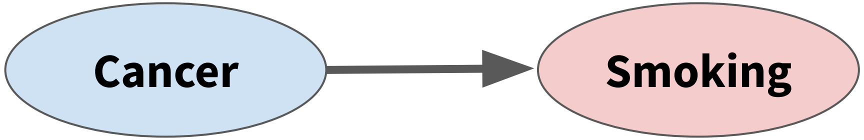
Dasgupta et al. (2018)

CS330 Student Presentation

# Background: Why Causal Reasoning?

There is only so much of the world we can understand via observation.

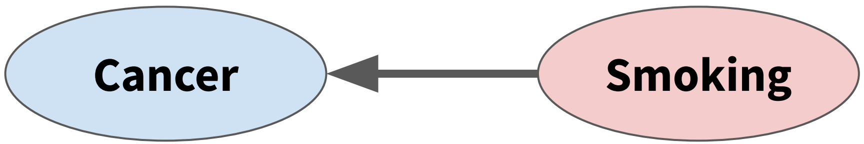
- Cancer **(correlates to)** Smoking → Cancer **(causes)** Smoking?
- Cancer **(correlates to)** Smoking → Smoking **(causes)** Cancer?
- Cancer **(correlates to)** Smoking → Genetics **(causes)** Cancer, Smoking?



# Background: Why Causal Reasoning?

There is only so much of the world we can understand via observation.

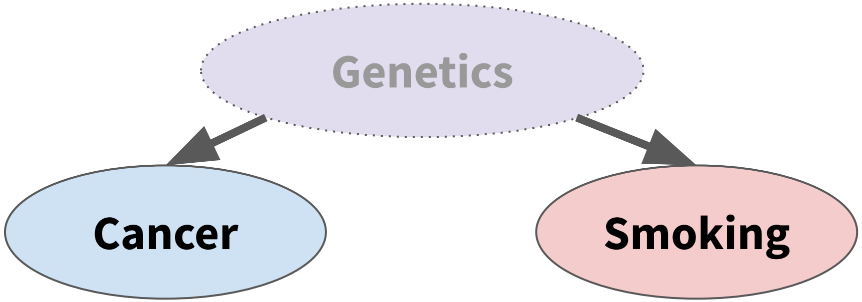
- Cancer **(correlates to)** Smoking → Cancer **(causes)** Smoking?
- Cancer **(correlates to)** Smoking → Smoking **(causes)** Cancer?
- Cancer **(correlates to)** Smoking → Genetics **(causes)** Cancer, Smoking?



# Background: Why Causal Reasoning?

There is only so much of the world we can understand via observation.

- Cancer **(correlates to)** Smoking → Cancer **(causes)** Smoking?
- Cancer **(correlates to)** Smoking → Smoking **(causes)** Cancer?
- Cancer **(correlates to)** Smoking → Genetics **(causes)** Cancer, Smoking?



# Background: Why Causal Reasoning?

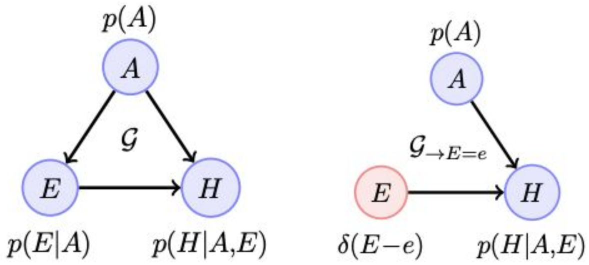
Fig. 1: Tank hidden in grass. Photos taken on a sunny day.

Fig. 2: No tank present. Photos taken on a cloudy day.

- Limits of ML from observational data: the "tank classification" story.
- If we want machine learning algorithms to **affect** the world (especially RL agents), they need a good understanding of cause and effect!

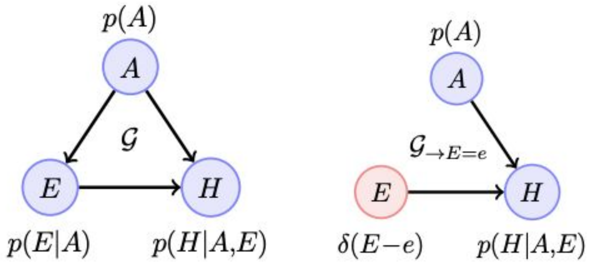
# Background: Causal Inference and the Do-Calculus

- Rather than: **P(A | B=b, C=c)**
- We might say: **P(A | do(B=b), C=c)** to represent an **intervention** where the random variable **B** is manipulated to be equal to **b**. This is completely different from an observational sample!
- Observing interventions lets us infer the causal structure of the data: a Causal Bayesian Network, or CBN.



# Method Overview - Dataset

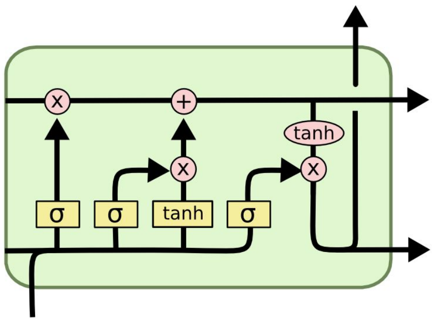
- Causal Bayesian Networks - directed acyclic graph that captures both *independence* and *causal* relations.
  - Nodes are Random Variables
  - Edges indicate one RV's causal effect on another




- Generated all graphs with 5 nodes ~ 60,000
- Each node was a Gaussian Random Variable. Parentless nodes had distribution N(0.0, 0.1), and child nodes had conditional distributions with mean equal to weighted sum of parents'
- One root node was always hidden to allow for an unobserved confounder

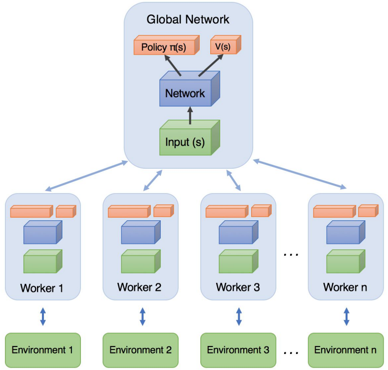
# Method Overview - Agent Architecture

- LSTM network (192 hidden units)
- Input: concatenated vector $[o_t, a_{t-1}, r_{t-1}]$
    - $o_t$ - “observation vector” composed of values of nodes + one-hot encoding of external intervention during the quiz phase
    - $a_{t-1}$ - previous action as a one-hot encoding
    - $r_{t-1}$ - previous reward as a single real-value
- Output: policy logits plus a scalar baseline. Next action sampled from a softmax over these logits.



# Method Overview - Learning Procedure

- Information phase (*meta-train)*
  - Output action $a_i$ sets value of $X_i$ to 5. Agent observes new values of RV's
  - Agent given T - 1 = 4 information steps
- Quiz phase (*meta-test)*
  - One hidden node selected at random and set to -5.
  - Agent informed of which node was set, and then asked to select the node with the highest sampled value
- Used *asynchronous advantage actor-critic* framework



# Experiments

**Settings:**

1. Observational
2. Interventional
3. Counterfactual

**Notation:**

- $\mathcal{G}$ : CBN with confounders
- $\mathcal{G}_{\rightarrow X_j}$ : Intervened CBN, where $X_j$ is the node being intervened on

# Experiment 1: *observational*

**Setup**: not allowed to intervene or observe external interventions ( $\mathcal{G}$ , not $\mathcal{G}_{\rightarrow X_j}$ )

- Observational: agent's actions are ignored, and $v_t$ sampled from $\mathcal{G}$
  - Obs (T=5)
  - Long-Obs (T=20)
- Conditional: choose an observable node and set its value to 5, then take a conditional sample from $\mathcal{G}$
  - Active
  - Random
- Optimal associative baseline (not learned): can perform exact associative reasoning but not cause-effect reasoning

# Experiment 1: *observational*

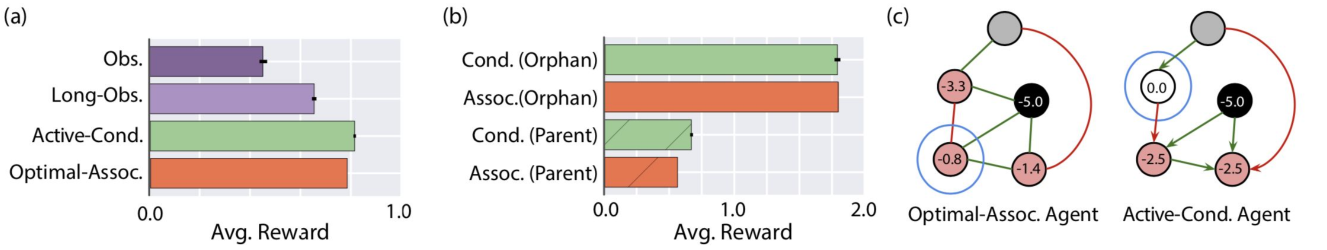




*Figure 4.* Active and Random Conditional Agents

## Questions:

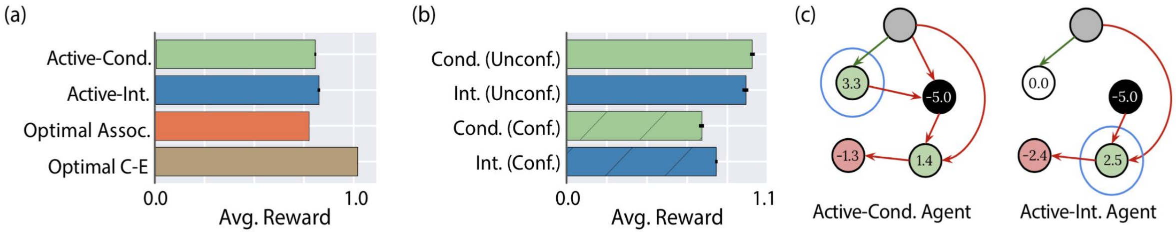
1. Do agents learn cause-effect reasoning from *observational* data?
2. Do agents learn to select useful *observations*?

# Experiment 2: *interventional*

**Setup**: allowed to make interventions in *information* phase only and observe samples from $\mathcal{G}_{\rightarrow X_j}$

- Interventional: chooses to intervene on an observable node $X_j$, and samples from the intervened graph $\mathcal{G}_{\rightarrow X_j}$
  - Active
  - Random
- Optimal Cause-Effect Baseline (not learned):
  - Receives the true CBN $\mathcal{G}$
  - In quiz phase, chooses the node with max value according to $\mathcal{G}_{\rightarrow X_j}$
  - Maximum possible score on this task

# Experiment 2: *interventional*



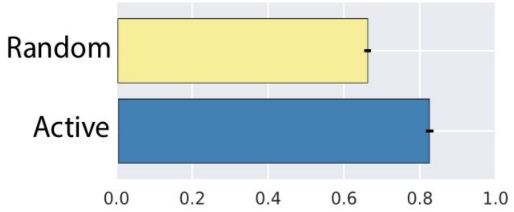




*Figure 5.* Active and Random Interventional Agents

**Questions:**

1. Do agents learn cause-effect reasoning from *interventional* data?
2. Do agents learn to select useful *interventions*?

# Experiment 3: *counterfactual*

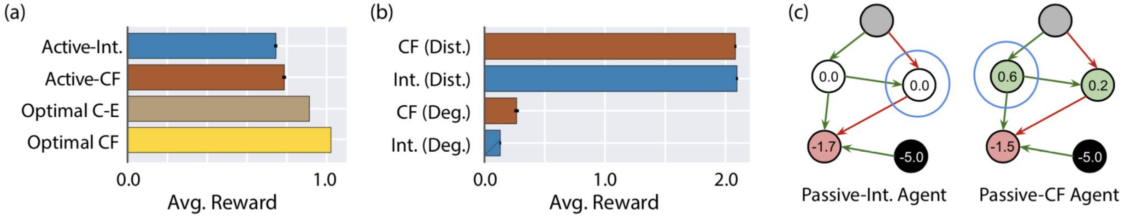
**Setup**: same as interventional setting, but tasked with answering a counterfactual question at quiz time

**Implementation:**

- Assume: $X_i = \sum_j w_{ji} X_j + \epsilon_i$
- Store some additional latent randomness in the last information phase step to use during the quiz phase
- "Which of the nodes would have had the highest value in the last step of the information phase if the intervention was different?"

**Agents:** counterfactual (active, random); optimal counterfactual baseline

# Experiment 3: *counterfactual*



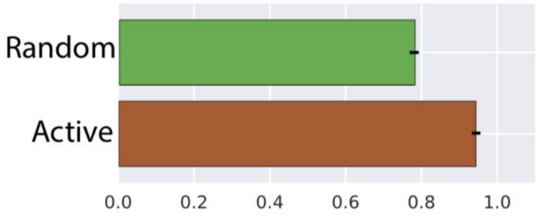




*Figure 7.* Active and Random Counterfactual Agents

**Questions:**

1. Do agents learn to do *counterfactual* inference?
2. Do agents learn to make useful interventions in the service of a *counterfactual* task?

# Strengths

- First direct demonstration of causal reasoning learning from an end-to-end model-free reinforcement learning algorithms.
- Experiments consider three grades of causal sophistication with varying levels of agent-environment interaction.
- Training these models via a meta-learning approach shifts the learning burden onto the training cycle and thus enables fast inference at test time.
- RL agents learned to more carefully gather data during the ‘information’ phase compared to a random data-collection policy: aspects of *active learning.*
- Agents also showed ability to perform *do-calculus:* agents with access to only observational data received more reward than highest possible reward achievable without causal knowledge.

# Weaknesses

- Experiment setting is quite limited: maximum of 6 nodes in the CBN graph, one hidden, edges/causal relationships were unweighted (sampled from {-1, 0, 1}), all nodes had a Gaussian distribution with the root node always having mean 0 and standard deviation 0.1 .
- Experiments are entirely performed on toy datasets. Would have been nice to see some real world demonstrations.
- Authors don’t interpret what strategy the agent is learning. Though results indicate that some causal inference is being made, to what extent and how is generally unclear.
- Perhaps outside the scope of this paper, but unclear about how well their approaches would scale to more complex datasets.
- Not clear why agent was not given more observations ($T > N$).

# Questions?