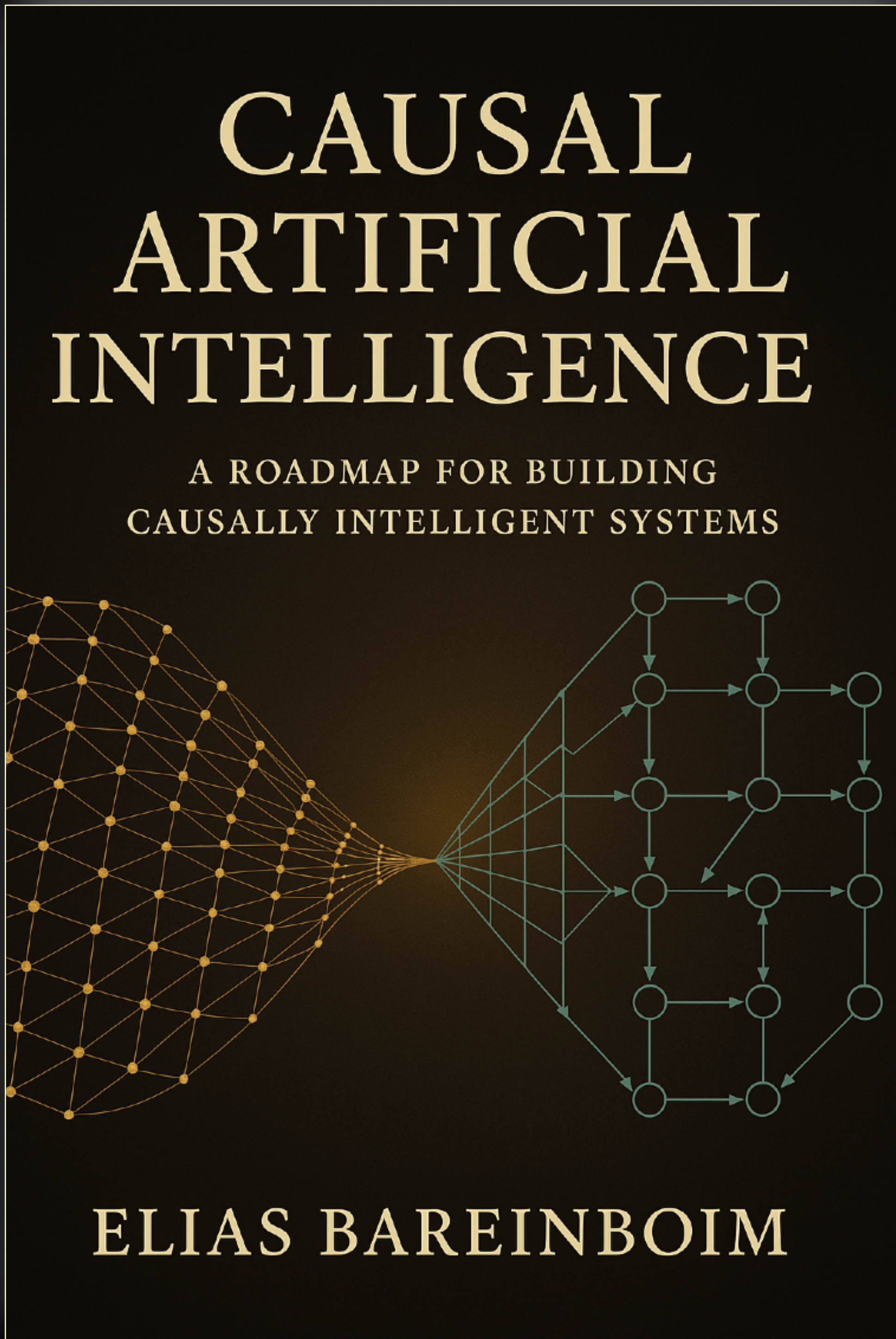
Towards Causal Artificial Intelligence

Elias Bareinboim
Causal Artificial Intelligence Lab
Columbia University
(𝕏 @eliasbareinboim)

Uncertainty in Artificial Intelligence
July, 2025



http://causalai-book.net

Recent Breakthroughs in AI

- Systems are able to perform extremely well in making predictions in high-dimensional settings.
- In particular, there has been huge progress in the fields of natural language processing, computer vision, and reinforcement learning.
  - Applications are everywhere, from medicine to business, agriculture to space exploration.

Recent Breakthroughs in AI

- Systems are able to perform extremely well in making … sional settings.
- In partic … e progress in the field … cessing, compute … ent learning.
  - Applic … om medicine to busine … exploration.

TechRepublic.

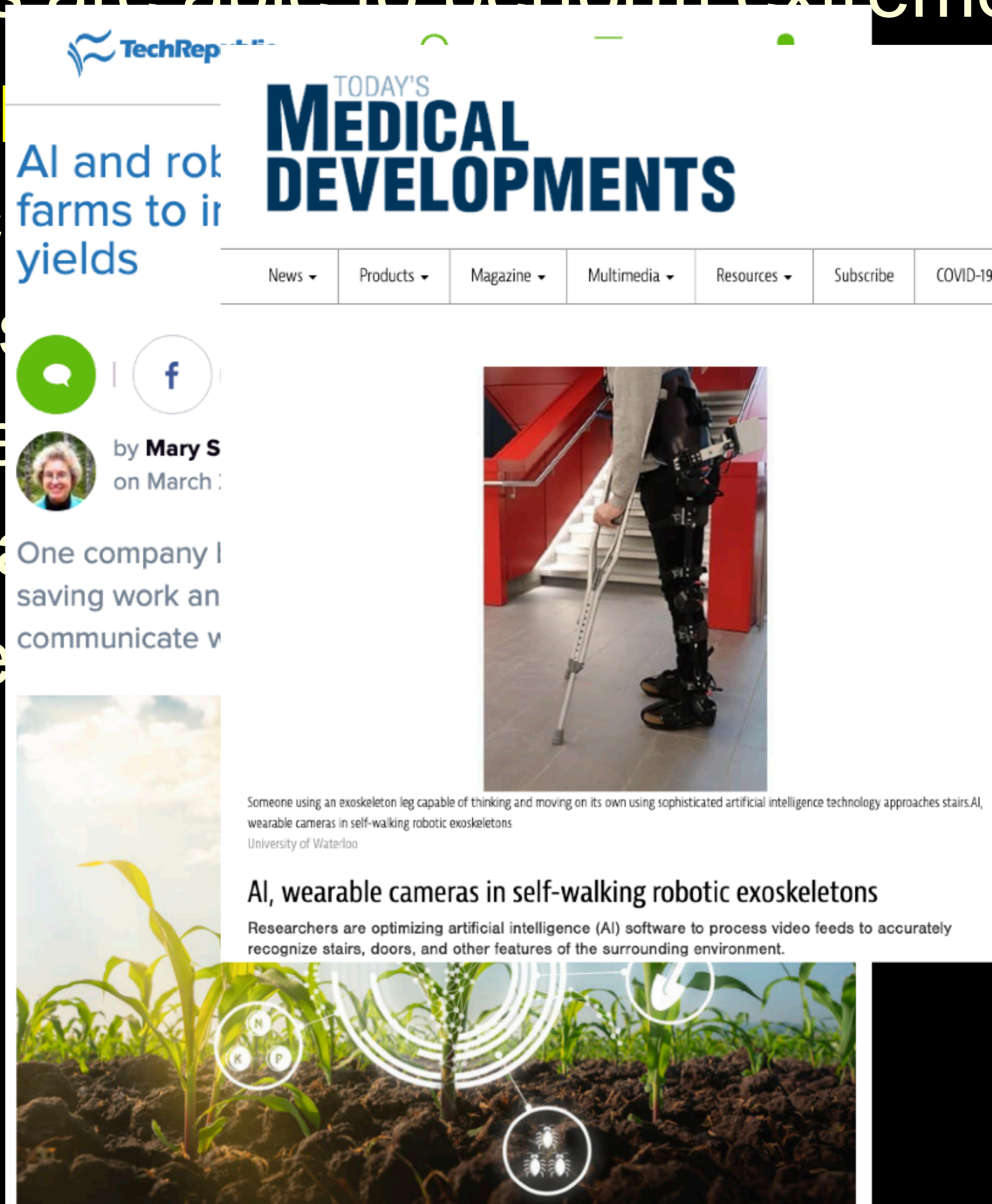
AI and robotics are helping optimize farms to increase productivity and crop yields

by Mary Shacklett in Artificial Intelligence
on March 23, 2021, 10:25 AM PST

One company built an autonomous vehicle to help haul crops, saving work and time. Others use drones and sensors to communicate with farmers.

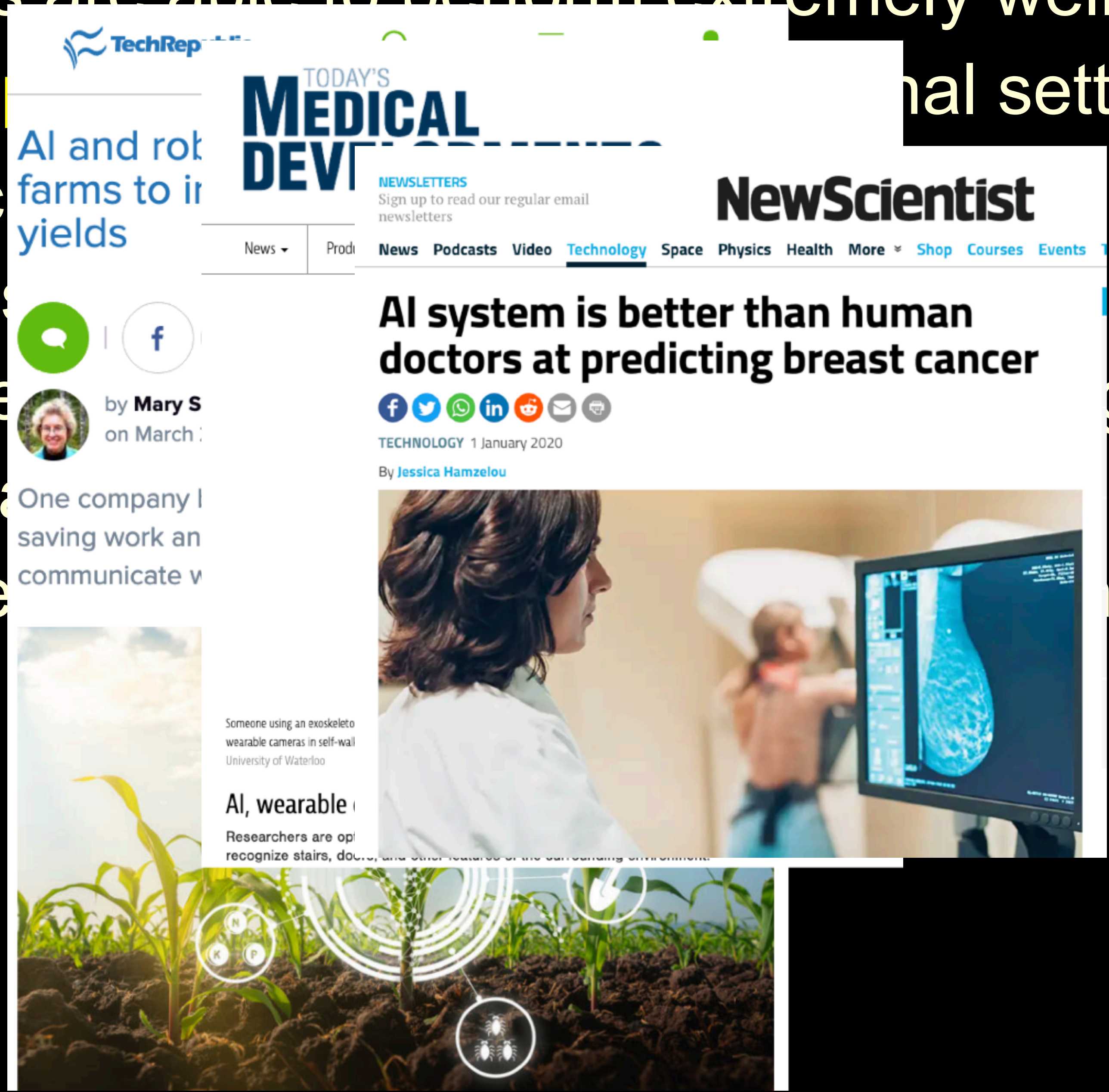
Recent Breakthroughs in AI

- Systems are able to perform extremely well in making ... nal settings.
- In partic ... ogress in the field ... sing, compute ... earning.
  - Applica ... medicine to busine ... oration.



Recent Breakthroughs in AI

- Systems are able to perform extremely well in making
nal settings.
- In partic
in the fields
compute
g.
  - Applic
ne to busine
n.



Recent Breakthroughs in AI

- Systems are able to perform extremely well in making … nal settings.
- In partic … in the field … compute … g.
  - Applic … ne to busine … n.

Recent Breakthroughs in AI

- Systems are able to perform extremely well in making ... nal settings.
- In partic... in the field... compute... g.
  - Applic... ne to busine... .

That's so awesome!

Does this mean we are done?

🙋‍♀️ (Assuming infinite compute & data.)

If not, what is missing?

Current Challenges in AI

- There are still serious foundational issues.
- Current AI systems suffer from:
  1. Lack of explainability capabilities
  2. Unfair & unethical decision-making
  3. Data inefficiency
  4. Poor generalizability
  5. Lack of controllability
- Those are thorny, long-standing problems.

Current Challenges in AI

- Th
- Cu
  1.

Advertisement

CNN Business Markets Tech Media Calculators Videos Watch Listen

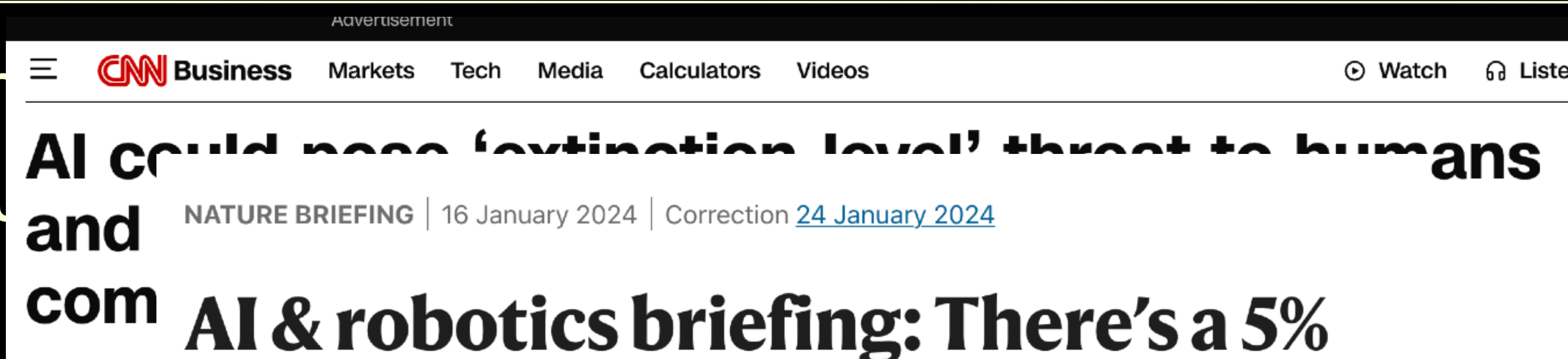
AI could pose 'extinction-level' threat to humans and the US must intervene, State Dept.-commissioned report warns

  2. Unfair & unethical decision-making
  3. Data inefficiency
  4. Poor generalizability
  5. Lack of controllability

- Those are thorny, long-standing problems.

Current Challenges in AI

- Th
- Cu
  1.
  2. Unf
  3. Dat
  4. Poc
  5. Lack of controllability
- Those are thorny, long-standing problems.



NATURE BRIEFING | 16 January 2024 | Correction 24 January 2024

AI & robotics briefing: There's a 5% risk that AI will wipe out humanity

In a survey of 2,700 AI experts, a majority said there was an at least 5% chance that superintelligent machines will destroy humanity. Plus, how medical AI fails when assessing new patients and a system that can spot similarities in a person's fingerprints.

Current Challenges in AI

- Th
- Cu
  1.
  2. U
  3. D
  4. Po
  5. Lack of controllability
- Those are thorny, long-standing problems.

Advertisement

CNN Business | Markets | Tech | Media | Calculators | Videos | Watch | Listen

AI could pose ‘extinction-level’ threat to humans and co

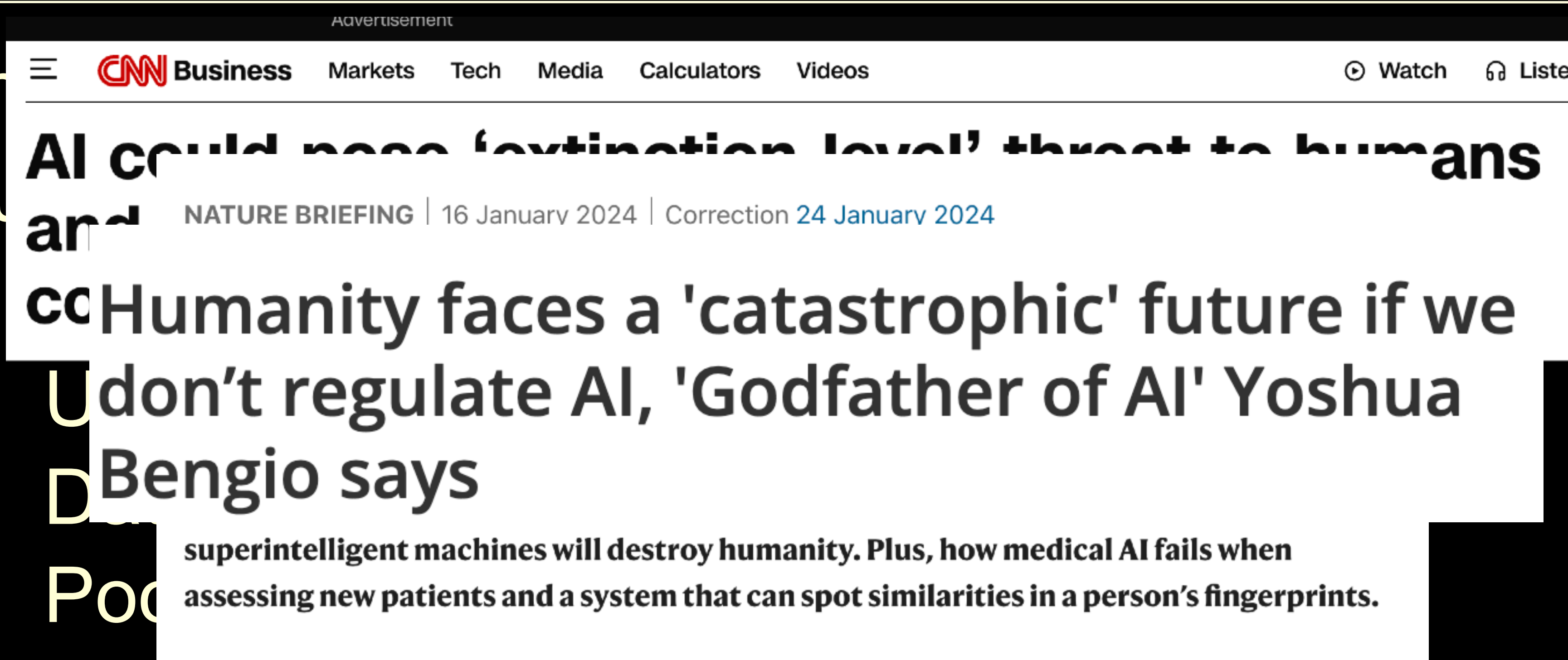
NATURE BRIEFING | 16 January 2024 | Correction 24 January 2024

Humanity faces a 'catastrophic' future if we don’t regulate AI, 'Godfather of AI' Yoshua Bengio says

superintelligent machines will destroy humanity. Plus, how medical AI fails when assessing new patients and a system that can spot similarities in a person’s fingerprints.

Current Challenges in AI

- Th
- Cu
  1.
  2. U
  3. D
  4. Po
  5. Lack of controllability



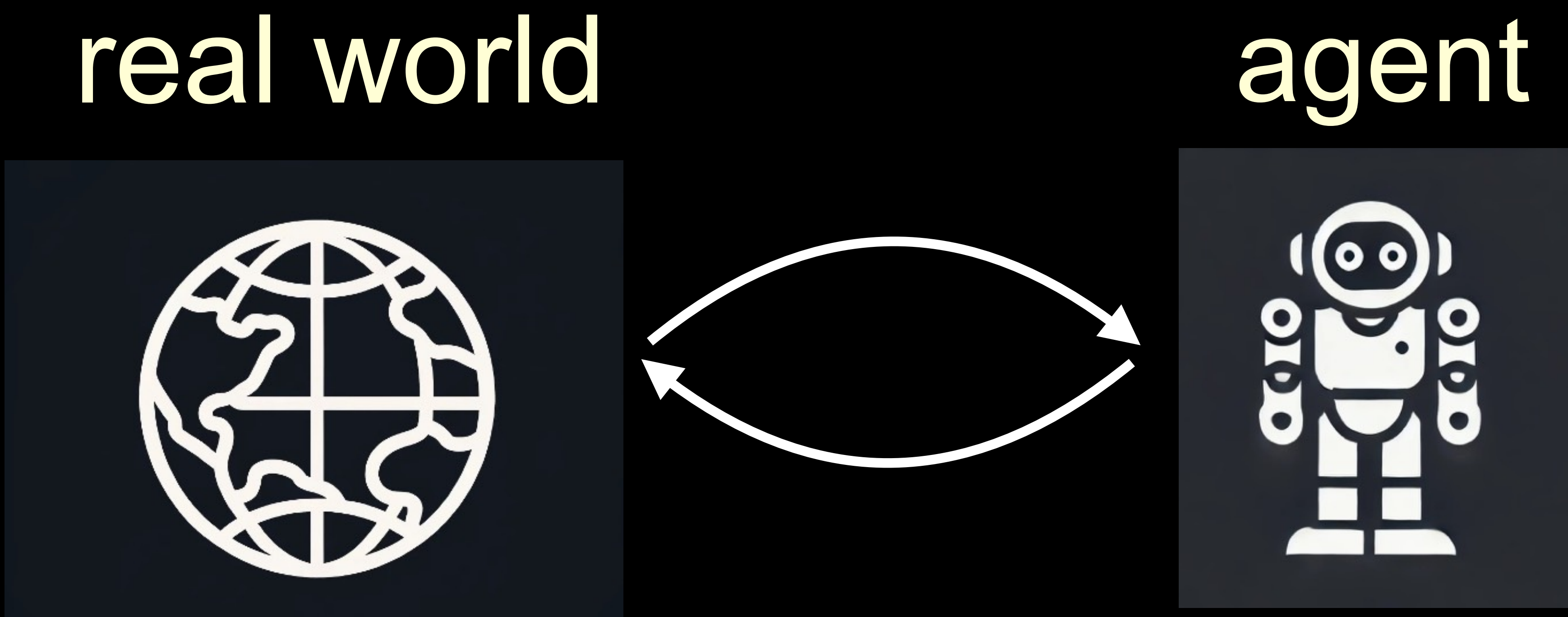
- Those are thorny, long-standing problems.
- Do these problems have anything in common?

EB: At the core of these challenges is the absence of a robust causal understanding.

Towards a Science of Artificial Intelligence

Idea: Model agent-environment ... uage!

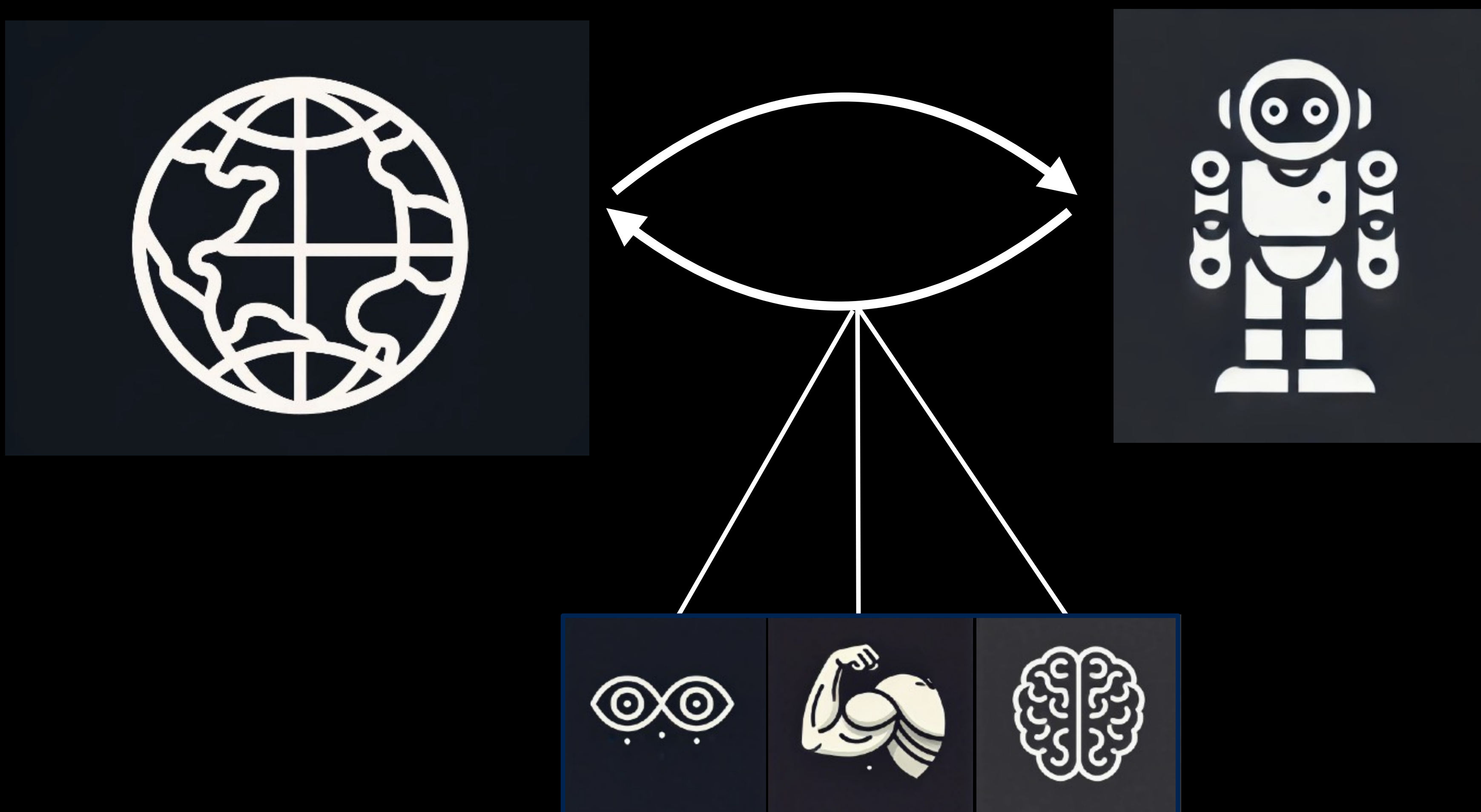
"What we want is a machine that can learn from experience." ---Alan Turing, 1947



Idea: Model agent-environment ... uage!

> “What we want is a machine that can learn from experience." ---Alan Turing, 1947

real world agent



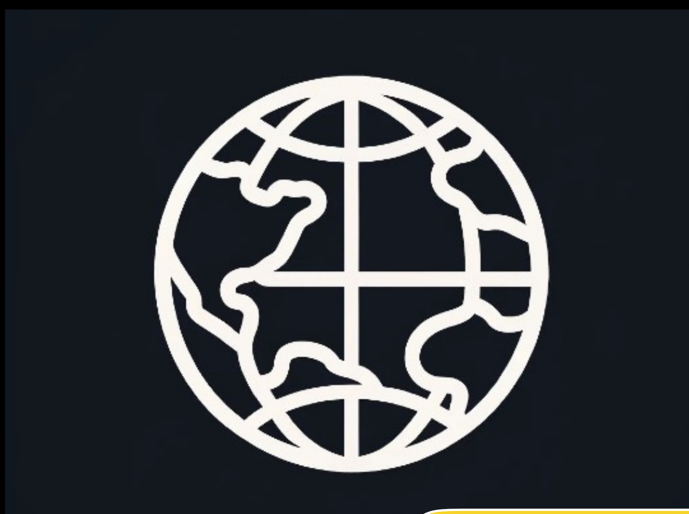
see do imagine

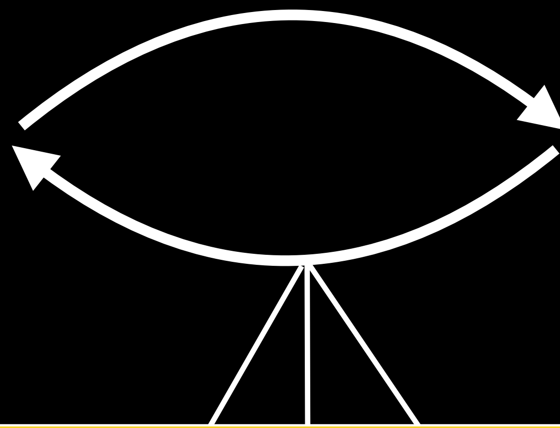
Idea: Model agent-environment
uage!

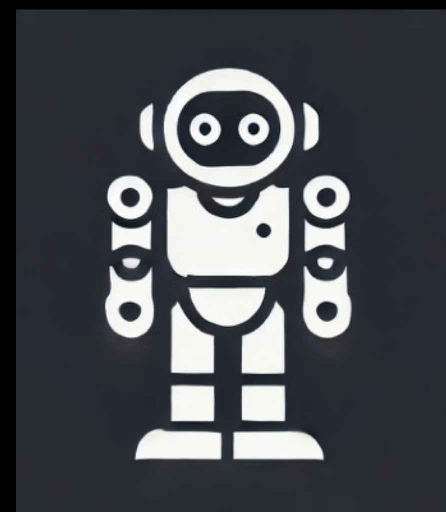
“What we want is a machine that can learn from experience." ---Alan Turing, 1947

real world

agent

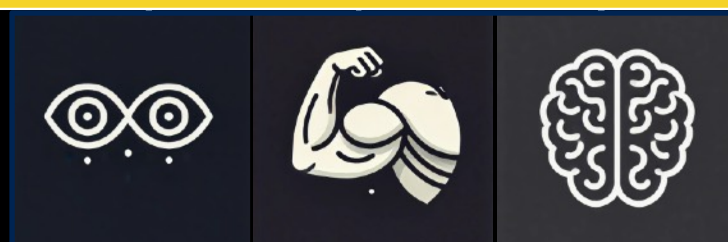






Pearl Causal Hierarchy

EB: unpacking Turing’s ‘experiencie’:



see do imagine

Causal Model → Pearl Causal Hierarchy

[Pearl & Mackenzie, 2018; Bareinboim, Correa, Ibeling, Icard 2022]

	Level (Symbol)	Typical Activity	Typical Question	Examples
1 👀	Associational $P(y \mid x)$	Seeing ML - (Un)Supervised (Bayes Net, DTree, SVM, DNN, …)	What is? How would seeing X change my belief in Y?	What does a symptom tell us about the disease?
2 💪	Interventional $P(y \mid do(x), c)$	Doing ML - Reinforcement (Causal Bayes Net, MDPs, POMDPs)	What if? What if I do X?	What if I take aspirin, will my headache be cured?
3 🧠	Counterfactual $P(y_x \mid x', y')$	Imagining, Retrospection Structural Causal Model	Why? What if I had acted differently?	Was it the aspirin that stopped my headache?

Causal Model → Pearl Causal Hierarchy

[Pearl & Mackenzie, 2018; Bareinboim, Correa, Ibeling, Icard 2022]

	Level (Symbol)	Typical Activity	Typical Question	Examples
1 👀	Associational $P(y \mid x)$	Seeing ML - (Un)Supervised (Bayes Net, DTree, SVM, DNN, …)	What is? How would seeing X change my belief in Y?	What does a symptom tell us about the disease?
2 💪	Interventional $P(y \mid do(x), c)$	Doing ML - Reinforcement (Causal Bayes Net,	What if? What if I do X?	What if I take aspirin, will my headache be
3 🧠	Counterfactua $P(y_x \mid x', y')$	S		

The formalization of the PCH provides a way to measure the capabilities (expressiveness) of different formalisms w.r.t. increasingly complex queries (see also causal hierarchy theorem).

MOVING BEYOND TRADITIONAL ML

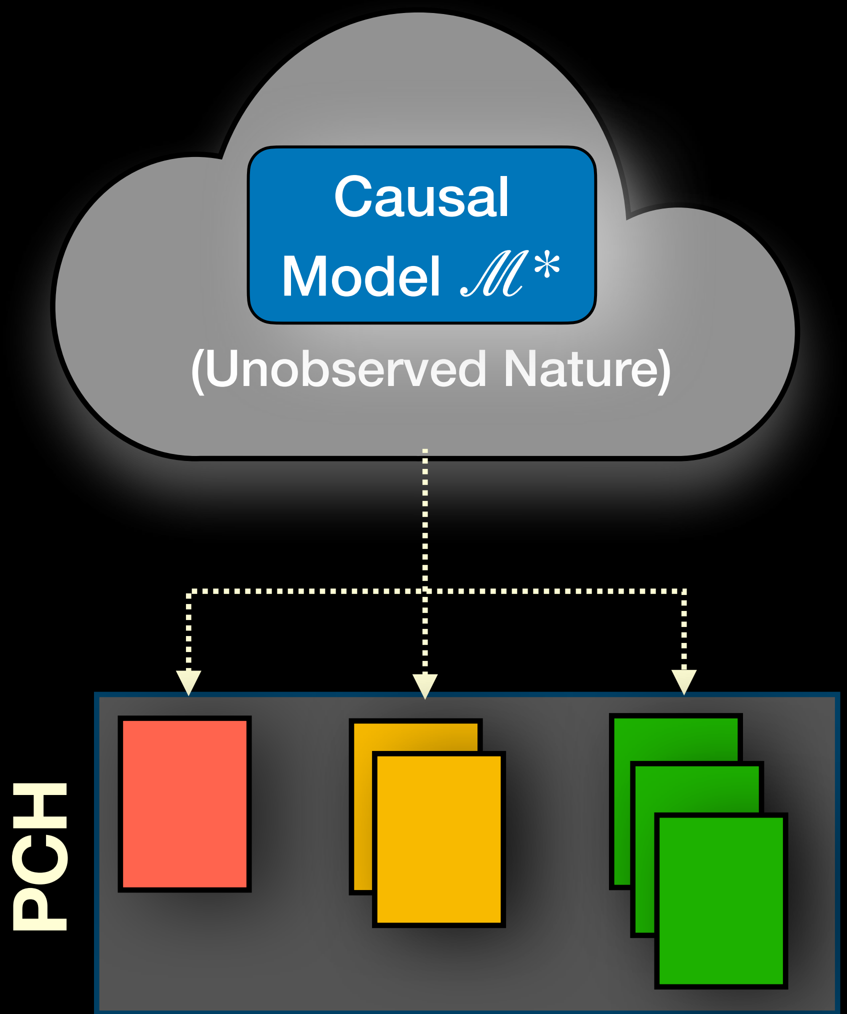
[Pearl & Mackenzie, 2018; Bareinboim et al., 2022]

Cross-layer inferences:

1	Input: (data)	Seeing	most of the available data is observational, passively collected
2	Output: (query)	Doing	most of the inferences are about causal effects (policies, treatments, decisions)

Research Question.
How to use the data collected from observations (layer 1) to answer questions about interventions (layer 2)?

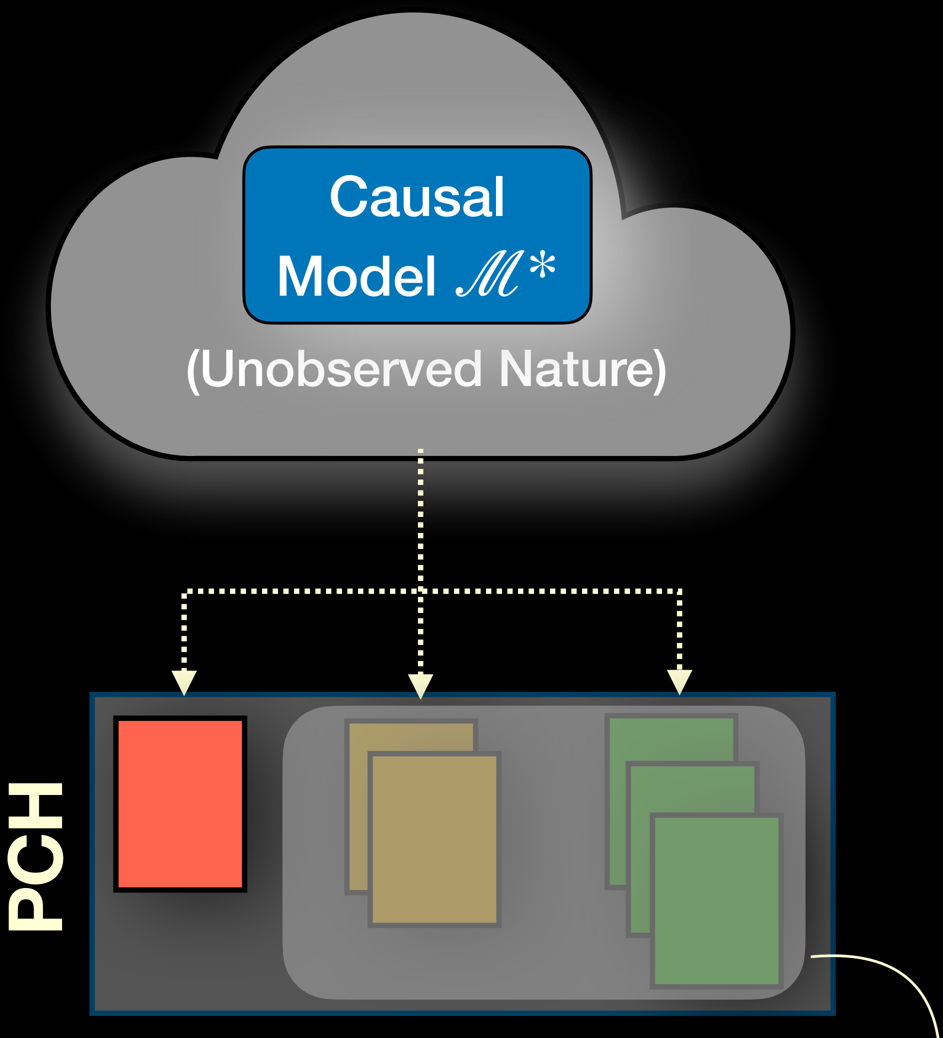
Challenge: Causal Generative Models



$L_1 : P(\mathbf{V})$ $L_2 : P(\mathbf{V} \mid do(\mathbf{x}))$ $L_3 : P(\mathbf{V}_{\mathbf{x}} \mid \mathbf{x}')$

Observations Interventions Counterfactuals

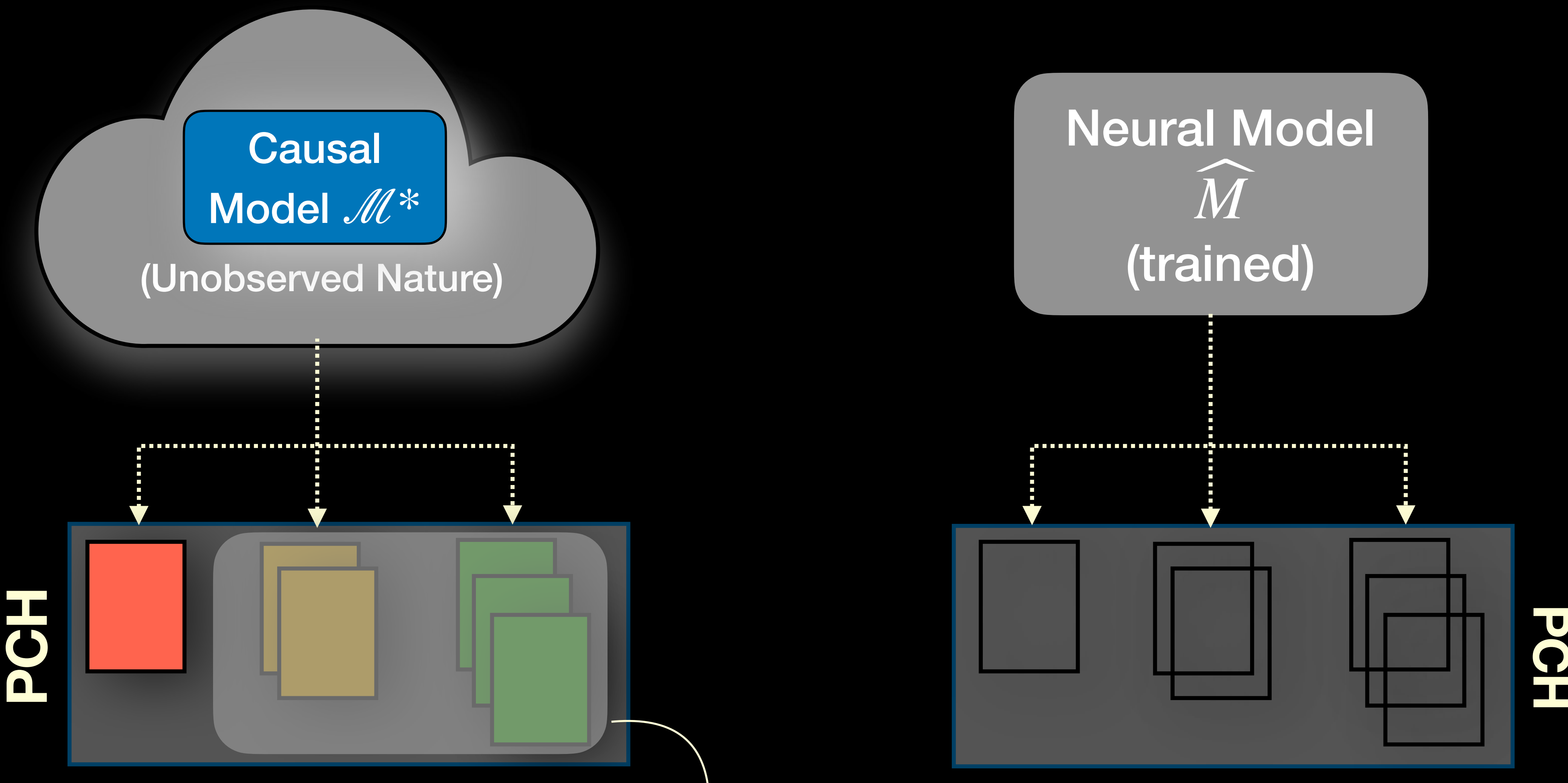
Challenge: Causal Generative Models



$L_1 : P(\mathbf{V})$ $L_2 : P(\mathbf{V} \mid do(\mathbf{x}))$ $L_3 : P(\mathbf{V}_\mathbf{x} \mid \mathbf{x}')$ Unobserved

Observations Interventions Counterfactuals

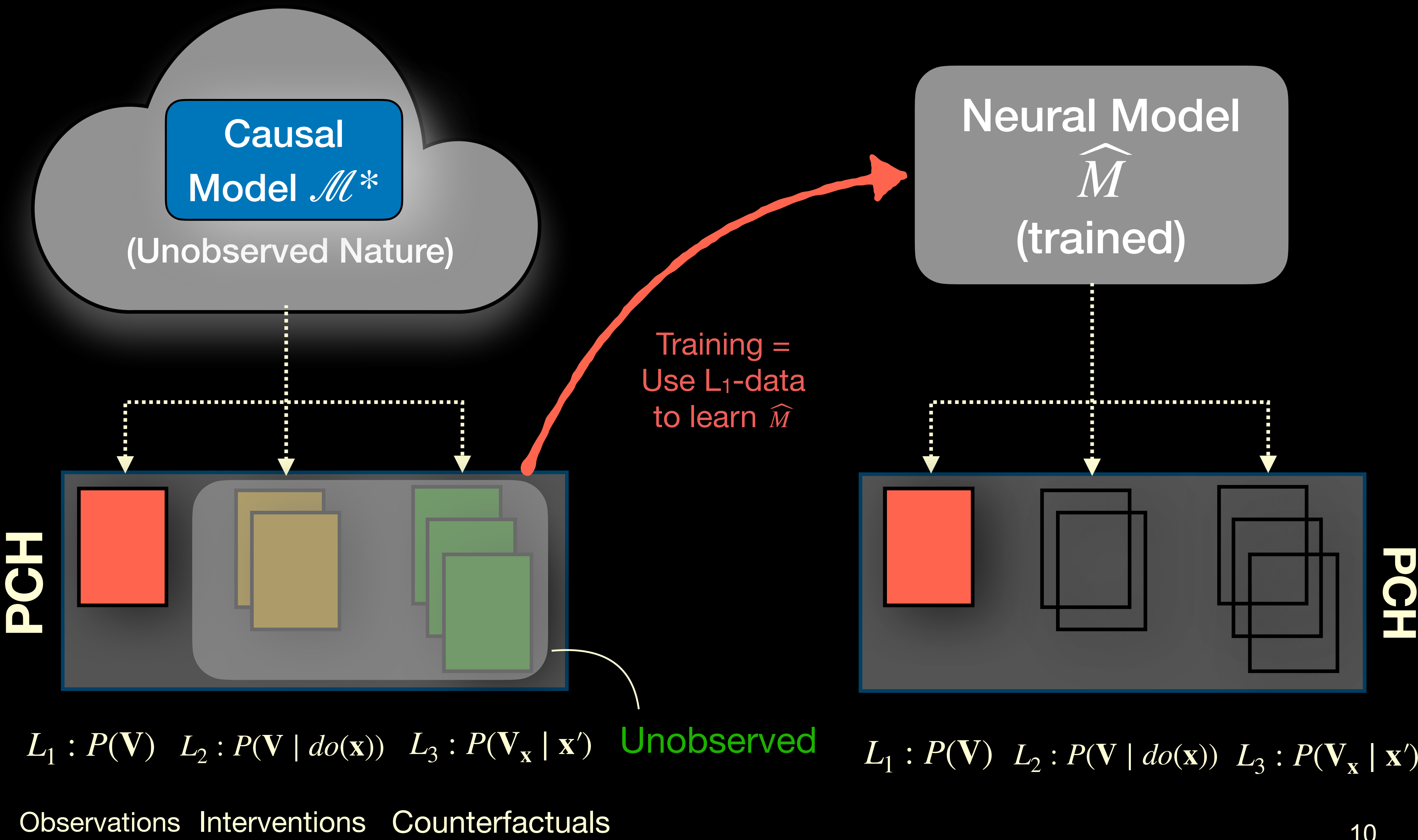
Challenge: Causal Generative Models



$L_1 : P(\mathbf{V})$ $L_2 : P(\mathbf{V} \mid do(\mathbf{x}))$ $L_3 : P(\mathbf{V}_{\mathbf{x}} \mid \mathbf{x}')$ Unobserved

Observations Interventions Counterfactuals

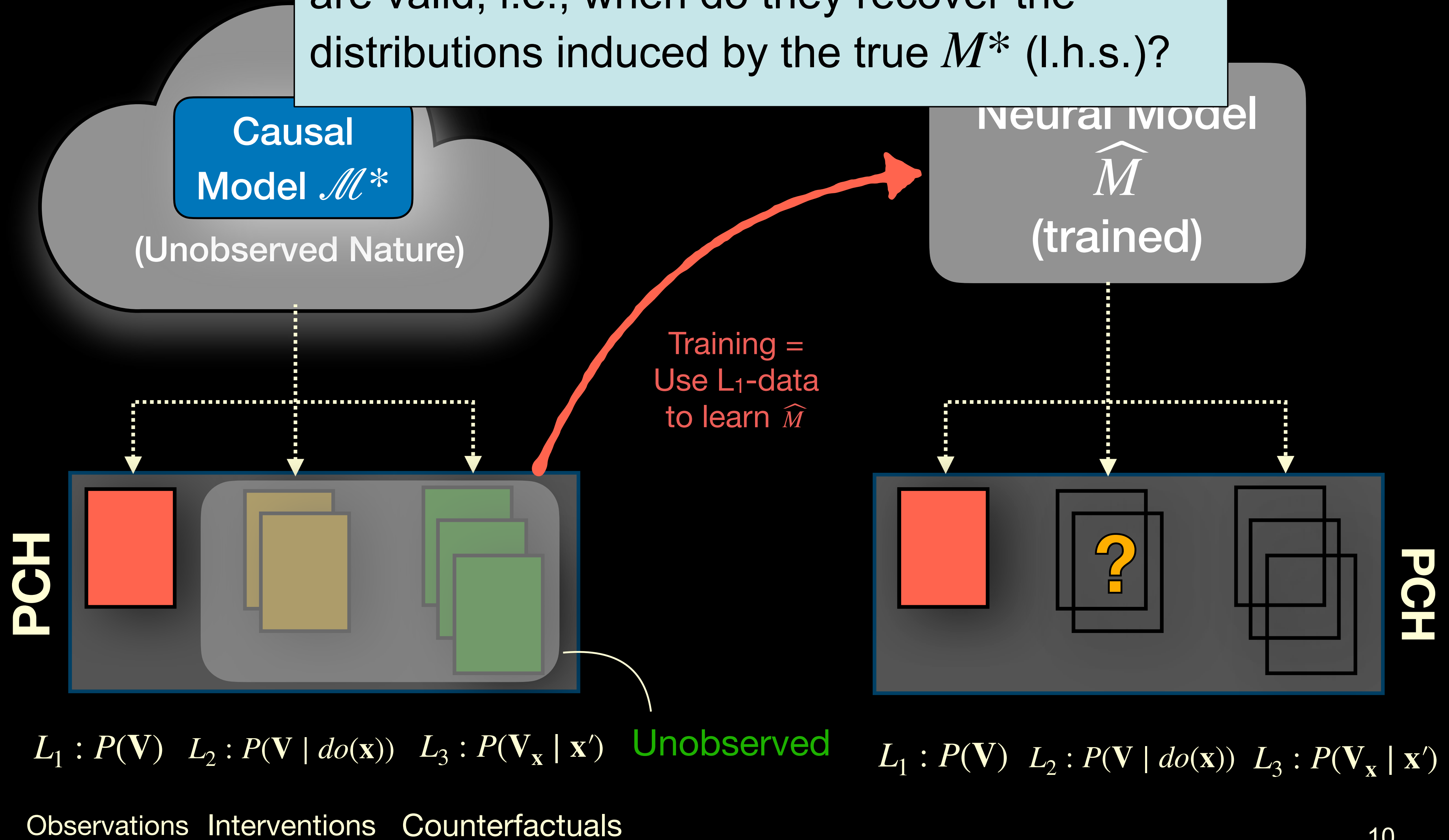
Challenge: Causal Generative Models



Challe… dels

Fundamental problem of generative AI.

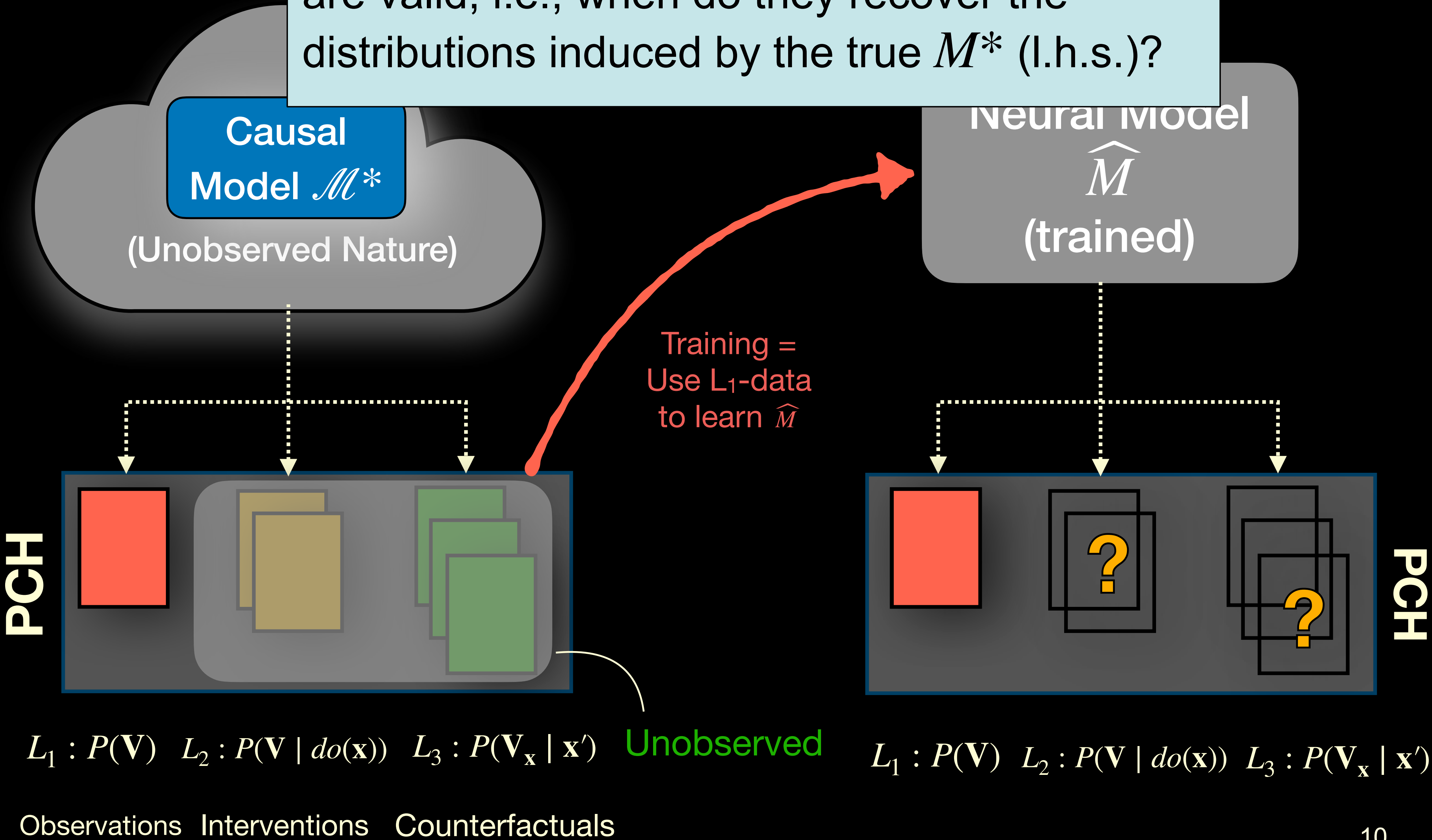
Under what conditions inferences in $\hat{M}$ (r.h.s.) are valid, i.e., when do they recover the distributions induced by the true M^* (l.h.s.)?



Challe… dels

Fundamental problem of generative AI.

Under what conditions inferences in $\hat{M}$ (r.h.s.) are valid, i.e., when do they recover the distributions induced by the true M^* (l.h.s.)?



Causal Artificial Intelligence

Goal: Develop more general AI systems endowed with the following capabilities:

1. Causal Understanding & Explanations
2. Efficient & Precise Decision-Making
3. Generalizable & Robust Inferences
4. Causal & Counterfactual Generation
5. Model Learning & Discovery

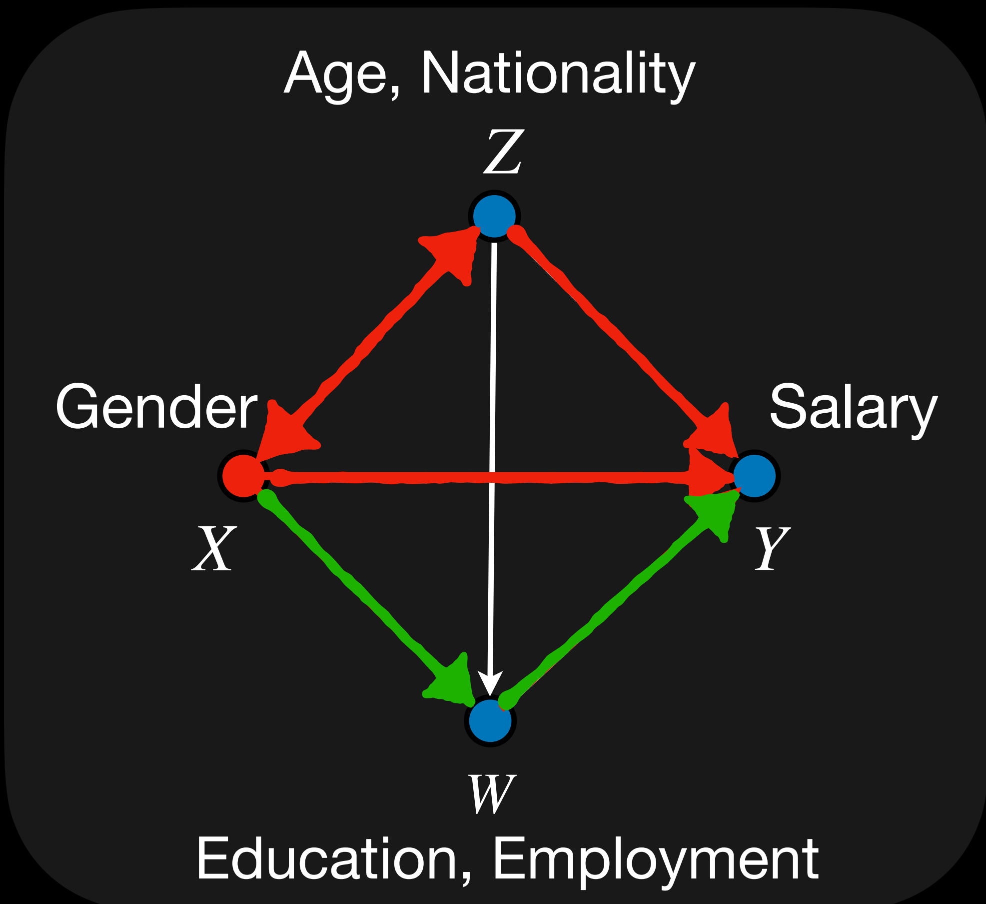
Causal Artificial Intelligence

Goal: Develop more general AI systems endowed with the following capabilities:

1. Causal Understanding & Explanations
2. Efficient & Precise Decision-Making
3. Generalizable & Robust Inferences
4. Causal & Counterfactual Generation
5. Model Learning & Discovery

Example 1: US Census 2018

[fairness.causalai.net]



- The data science team observes that $\text{TV} = E[Y \mid \text{male}] - E[Y \mid \text{female}]$

  This disparity could be explained in different ways, i.e.,

  (1) The salary decision is based on employee' gender: $X \rightarrow Y$.

  (2) Decisions were based on education or employment: $X \rightarrow W \rightarrow Y$.

  (3) Age or nationality are used to infer the person's gender: $X \leftrightarrow Z \rightarrow Y$.

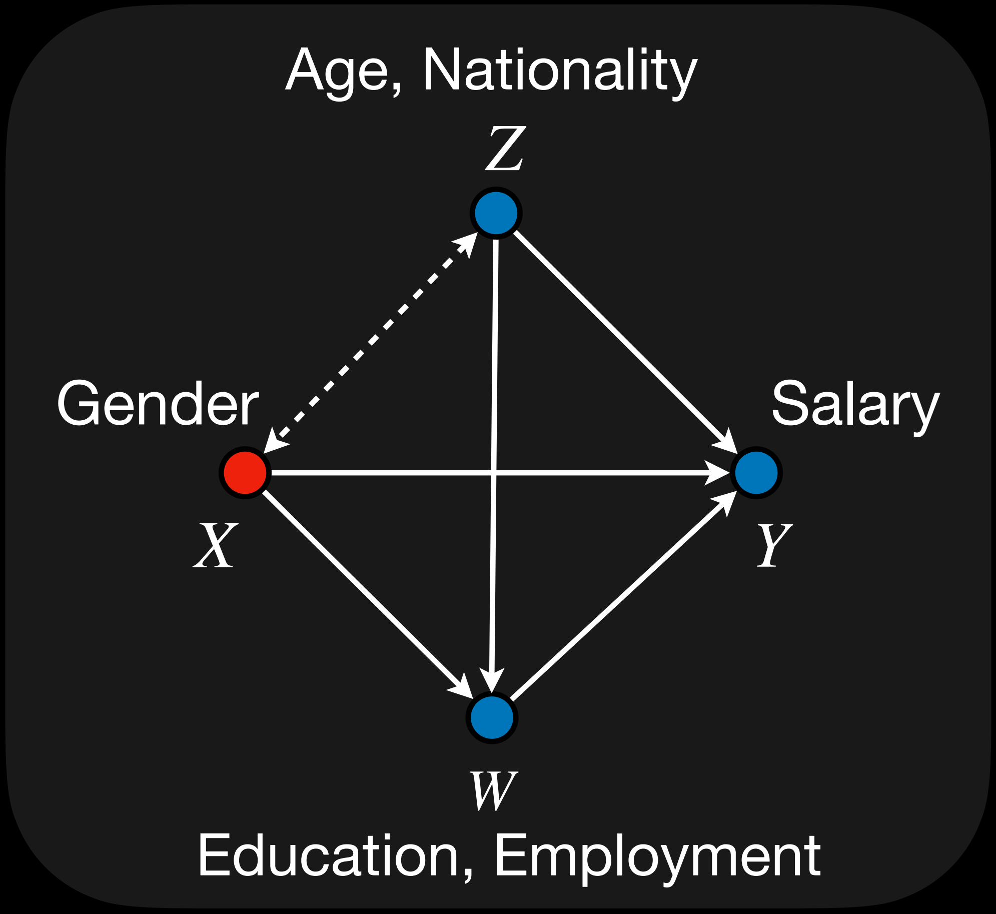
After a legal argument, the jury may be okay with *Y*'s variations due to education, but not okay with the variations due to gender or age.

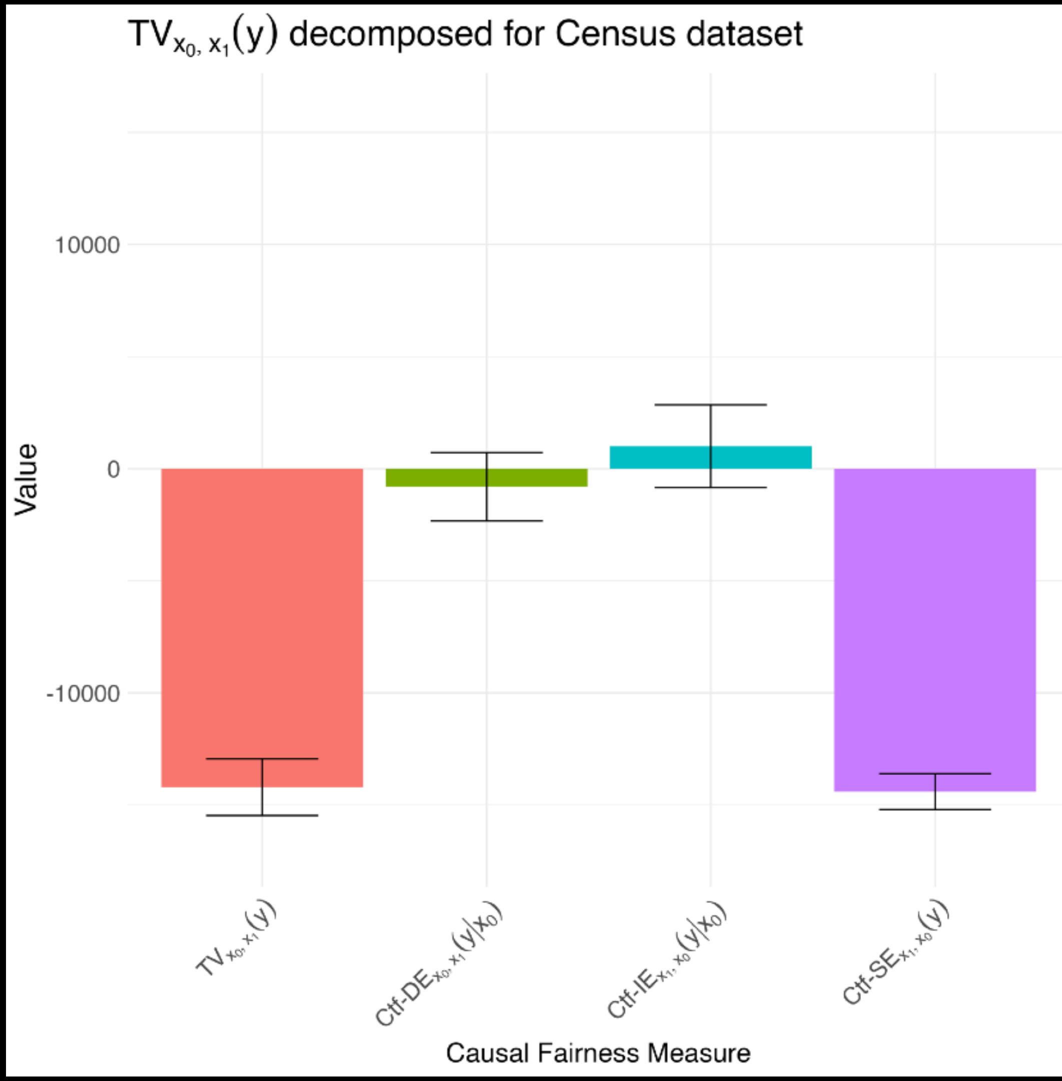
How to disentangle the variations within TV?

US Census 2018 — Causal Analysis

[fairness.causalai.net]

- Observed Disparity (data):
  - $TV_{x0, x1}(y)$ = \$14,000/year

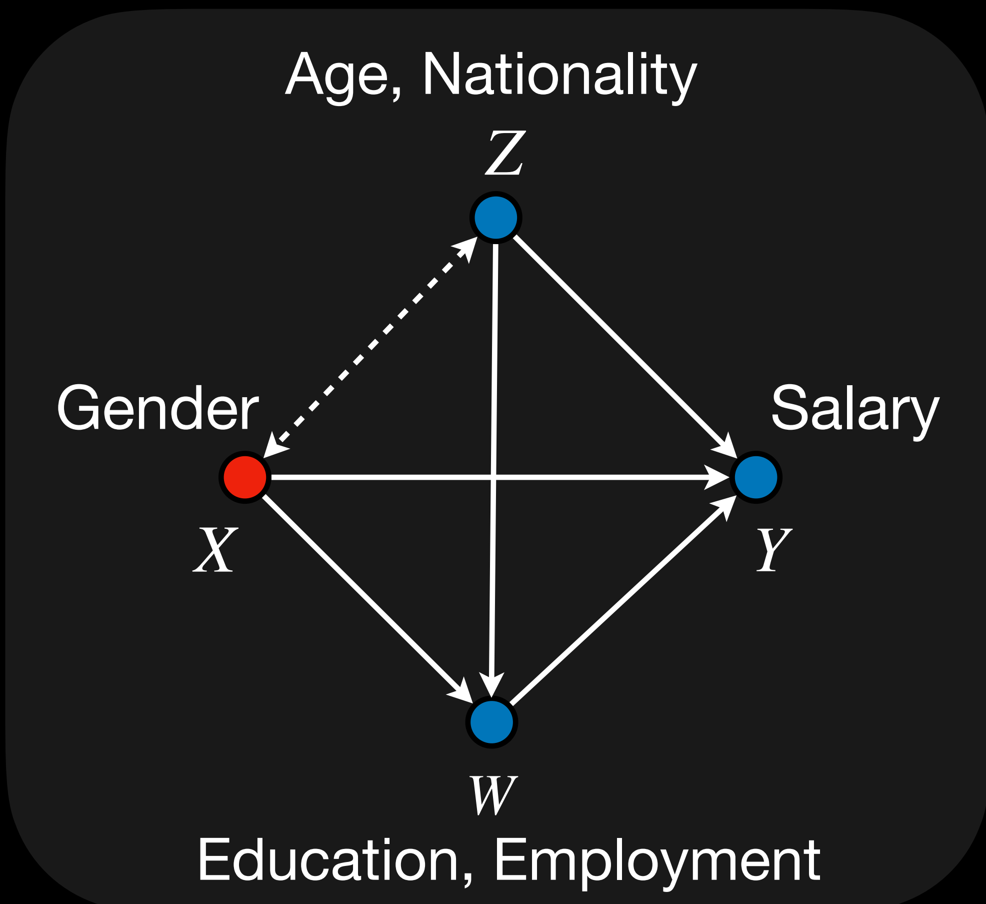


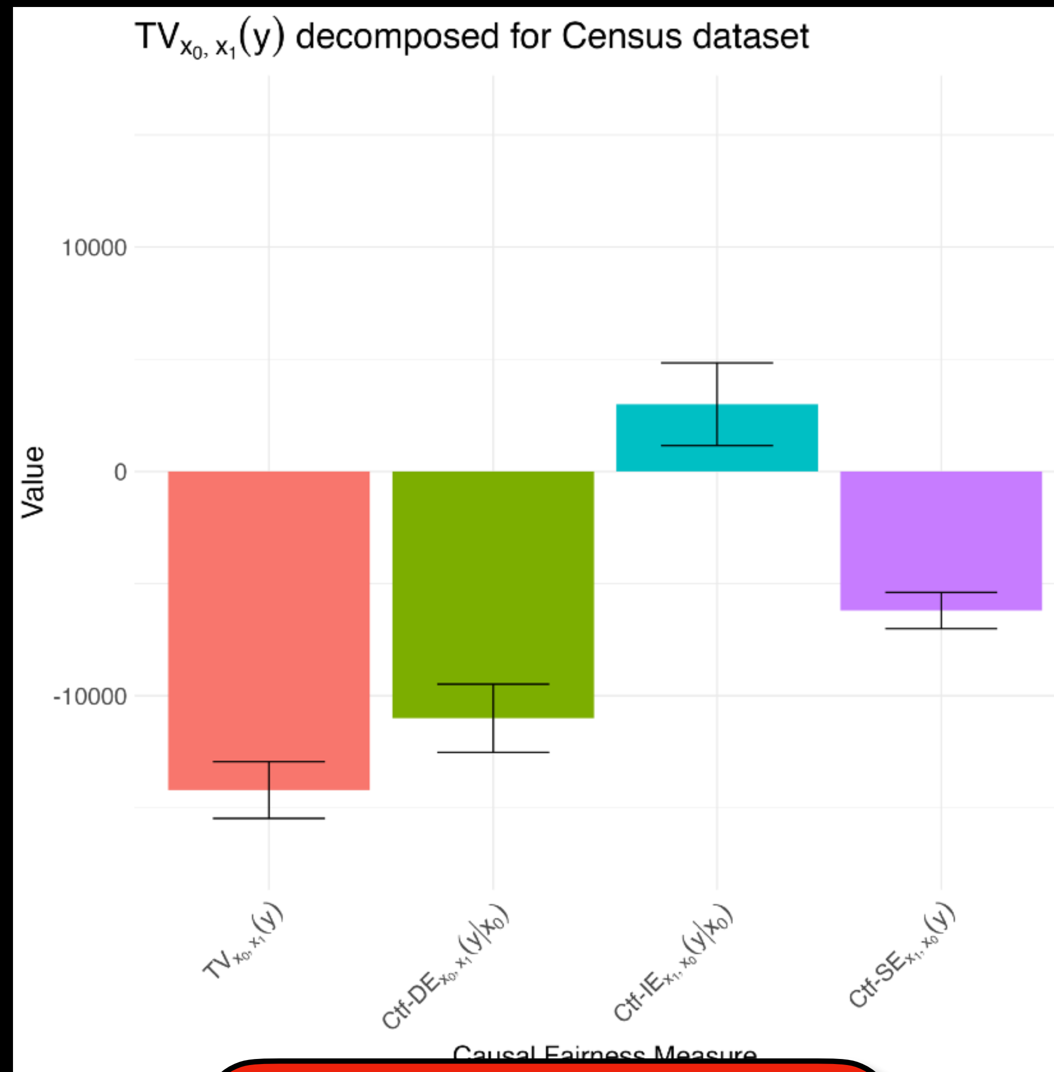


US Census 2018 — Causal Analysis

[fairness.causalai.net]

- Observed Disparity (data):
  - $TV_{x0,\,x1}(y)$ = \$14,000/year





TV cannot distinguish causally different explanations!

Example 2: College Admissions

- A university in the US admits applicants every year. The data science team is tasked with quantifying discrimination in the admission process and tracking it over time, between 2010 and 2020. The data-generating process changes over time and can be described as follows.
  Let X denote gender (x_0 female, x_1 male). Let Z be the age at the time of application ($Z = 0$ under 20 years, $Z = 1$ over 20 years), and let W denote the department of application ($W = 0$ for arts & humanities, $W = 1$ for sciences). Finally, let Y denote the admission decision.

SCM $M = \langle \mathcal{F}_t, P_t(U) \rangle$

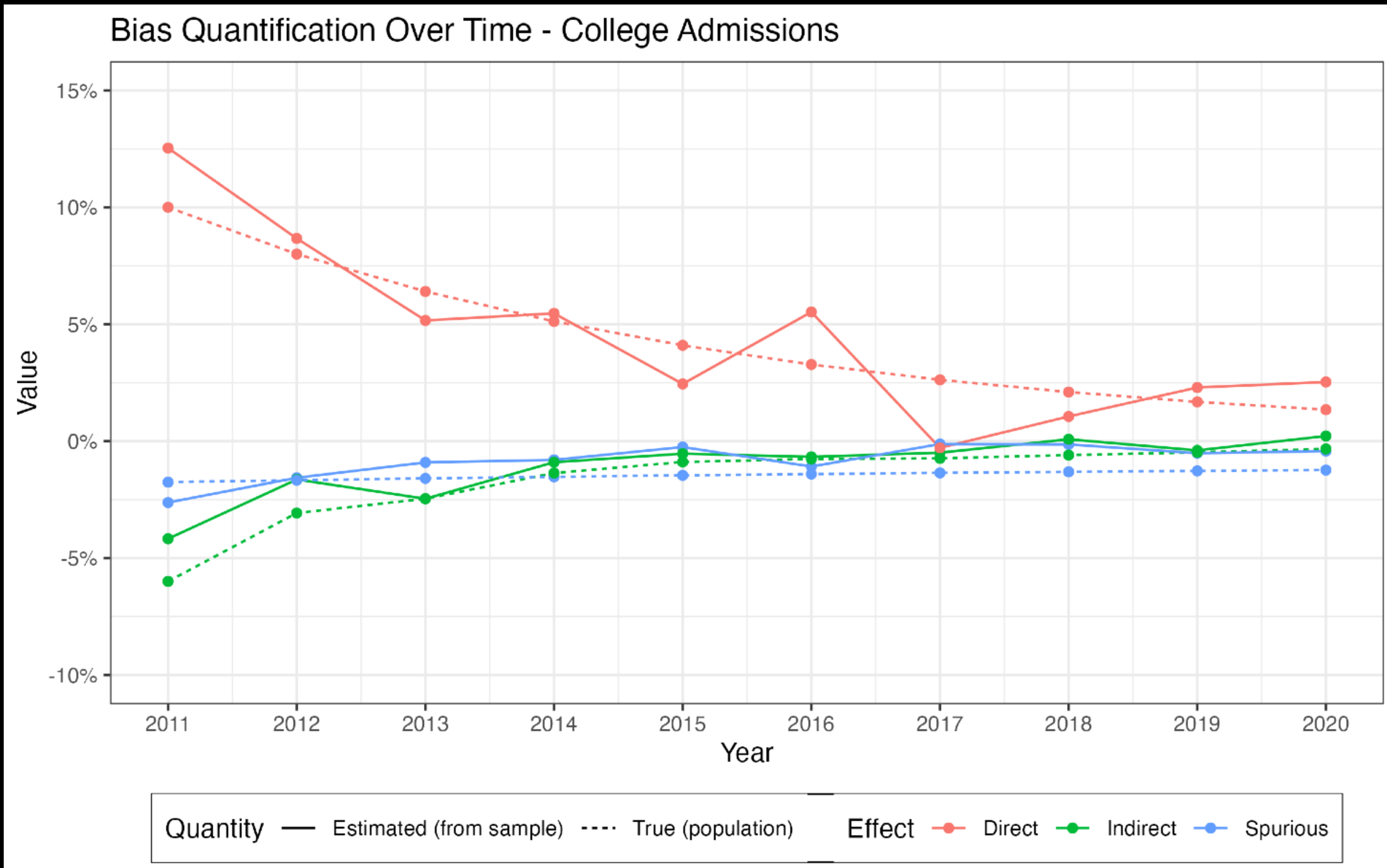
$$X \leftarrow 1(U_X < 0.5 + 0.1 U_{XZ})$$
$$Z \leftarrow 1(U_Z < 0.5 + \kappa(t) U_{XZ})$$
$$W \leftarrow 1(U_W < 0.5 + \lambda(t) X)$$
$$Y \leftarrow 1(U_Y < 0.1 + \alpha(t) X + \beta(t) W + 0.1 Z)$$

$$U_{XZ} \in \{0,1\}, P(U_{XZ} = 1) = 0.5,$$
$$U_X, U_Z, U_W, U_Y \sim \textbf{Unif}[0,1].$$

Time Evolution $\theta_{t \to t+1}$

$$\kappa(t+1) = 0.9\kappa(t)$$
$$\lambda(t+1) = \lambda(t)(1 - \beta(t))$$
$$\beta(t+1) = \beta(t)(1 - \lambda(t)) f(t),$$
$$f(t) \sim \textbf{Unif}[0.8, 1.2]$$
$$\alpha(t+1) = 0.8\alpha(t)$$

Bias Quantification over time



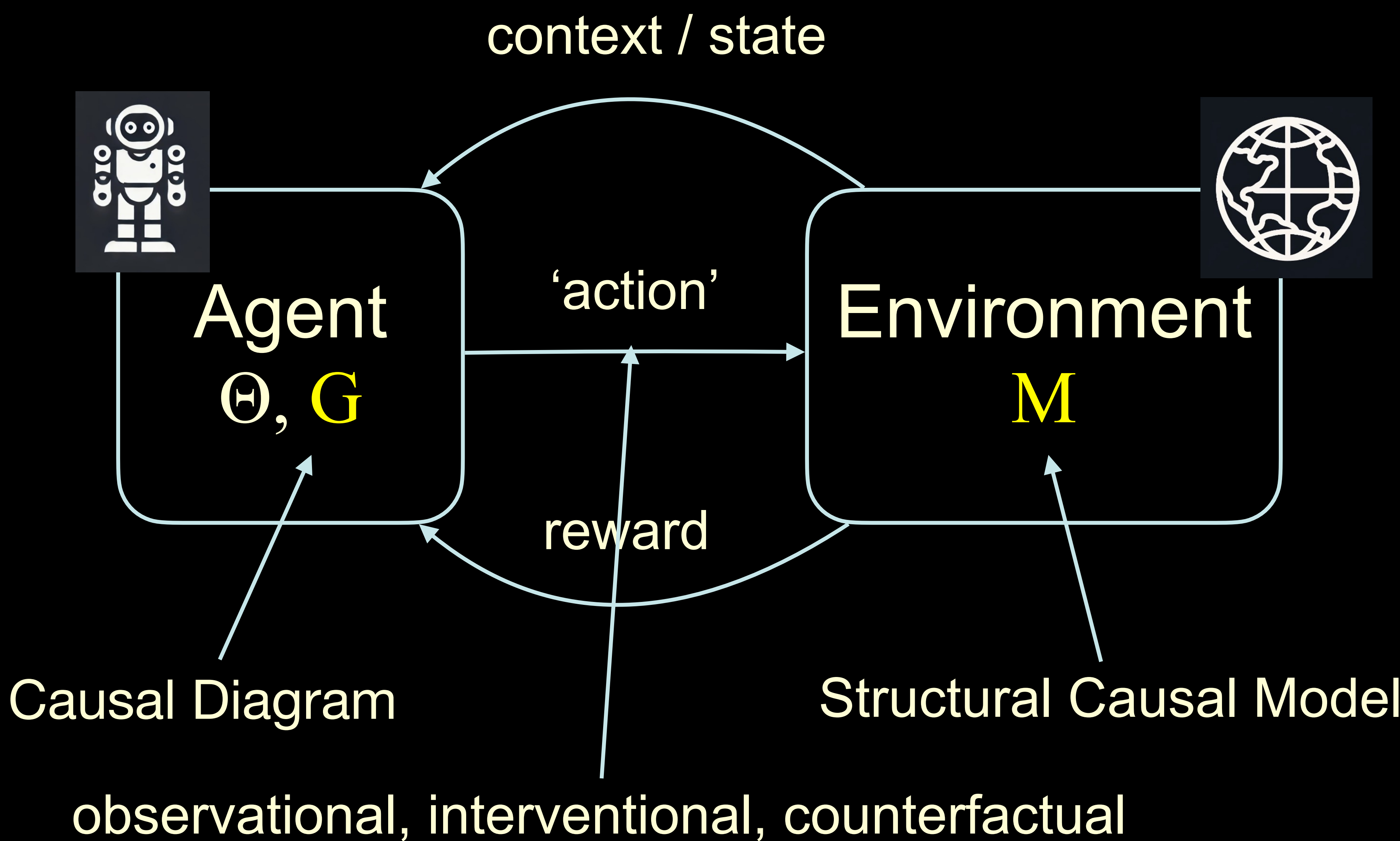
Causal AI — Desiderata

The new generation of AI systems is expected to provide the following capabilities:

1. Causal Understanding & Explanations
2. Efficient & Precise Decision-Making
3. Generalizable & Robust Inferences
4. Causal Generative Capabilities
5. Model Learning & Discovery

Causal RL - Big Picture

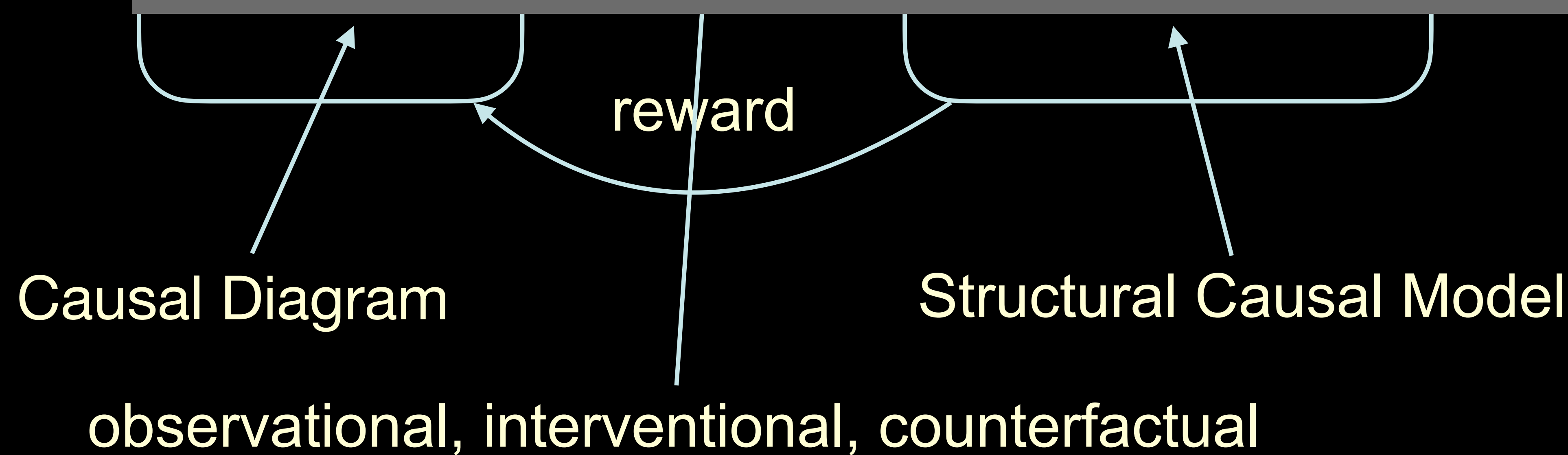
[crl.causalai.net]



Two key observations (RL → CRL):

1. **The environment and the agent are tied through the pair SCM M & causal graph G.**
2. **We defined different types of “actions”, or interactions, to avoid ambiguity (thr. PCH).**

As formally defined by
(1) the pair <M, G>, and (2) the PCH.



CRL NEW CHALLENGES & OPPORTUNITIES (I)

[crl.causalai.net]

Task 1 (NeurIPS’19, ICML’20, NeurIPS’21, ICML’22, UAI’25)

Causal Offline to Online Learning (COOL)

(generalized policy learning)

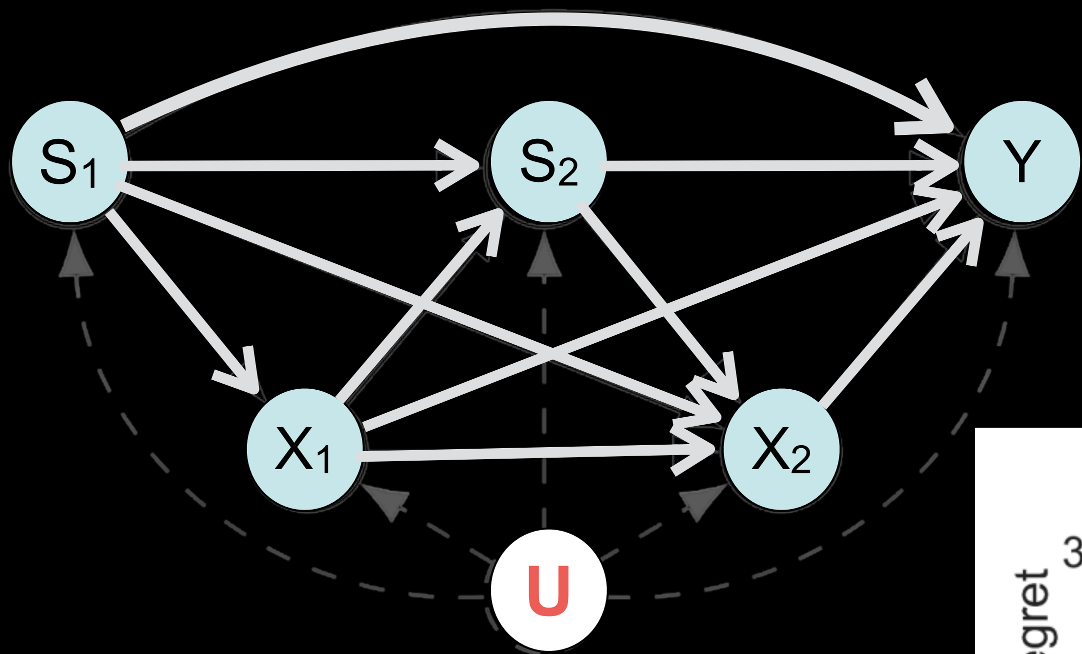
Task 1: COOL -- Cancer Dynamic Treatment Regime

[crl.causalai.net]

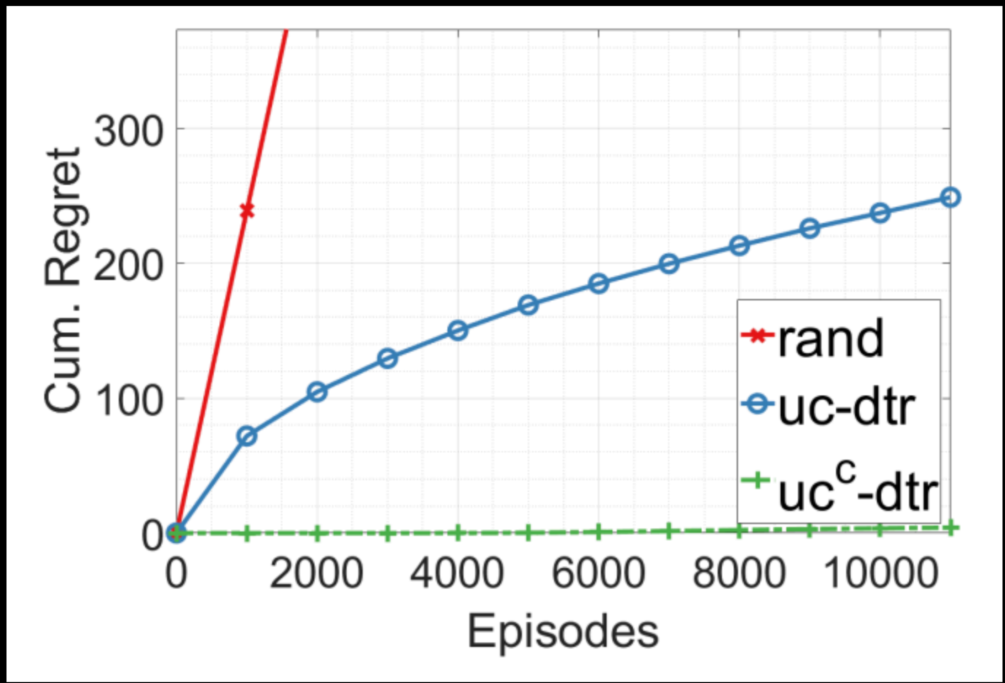
- We test the survival model of the two-stage clinical trial conducted by the Cancer and Leukemia Group B. Protocol 8923 was a double-blind, placebo controlled two-stage trial reported by (Stone et al. NEJM'95) examining the effects of infusions of granulocyte-macrophage colony-stimulating factor (GM-CSF) after initial chemotherapy in patients with acute myelogenous leukemia (AML).
- Standard chemotherapy could place patients at increased risk of death due to infection or bleeding-related complications. GM-CSF administered after chemo might assist patient recovery, thus reducing the number of deaths due to such complications.
- Patients were randomized initially to GM-CSF or placebo following standard chemo. Later, patients meeting the remission criteria and consenting to further participation were offered a second randomization to one of two intensification treatments.

Task 1: COOL -- Cancer Dynamic Treatment Regime

[crl.causalai.net]



- X_1, X_2: treatment
- S_1, S_2: state
- Y: outcome
- U: unobserved confounders



CRL NEW CHALLENGES & OPPORTUNITIES (I)

[crl.causalai.net]

Task 1 (NeurIPS’19, ICML’20, NeurIPS’21, ICML’22, UAI’25)

Causal Offline to Online Learning (COOL)

(generalized policy learning)

Task 2 (NeurIPS’18, AAAI’19, NeurIPS’20)

When and where to intervene?

(refining the policy space)

Task 3 (NeurIPS’15, ICML’17, AAAI’19, CleaR’22)

Counterfactual Decision-Making

(changing optimization function based on intentionality, free will, and autonomy)

CRL NEW CHALLENGES & OPPORTUNITIES (II)

[crl.causalai.net]

Task 4 (PNAS'16, UAI'19, AAAI'20, NeurIPS'22, AAAI'24)

Generalizability & robustness of causal claims (transportability & structural invariances)

Task 5 (NeurIPS'17, ICML'18, NeurIPS'19, '20, '22,'23)

Learning causal model by combining observations (L_1) and experiments (L_2)

Task 6 (NeurIPS'20, '21, '24, ICLR'23)

Causal Imitation Learning

CRL NEW CHALLENGES & OPPORTUNITIES (III)

[crl.causalai.net]

Task 7 (ICLR'24)

Causally Aligned Curriculum Learning

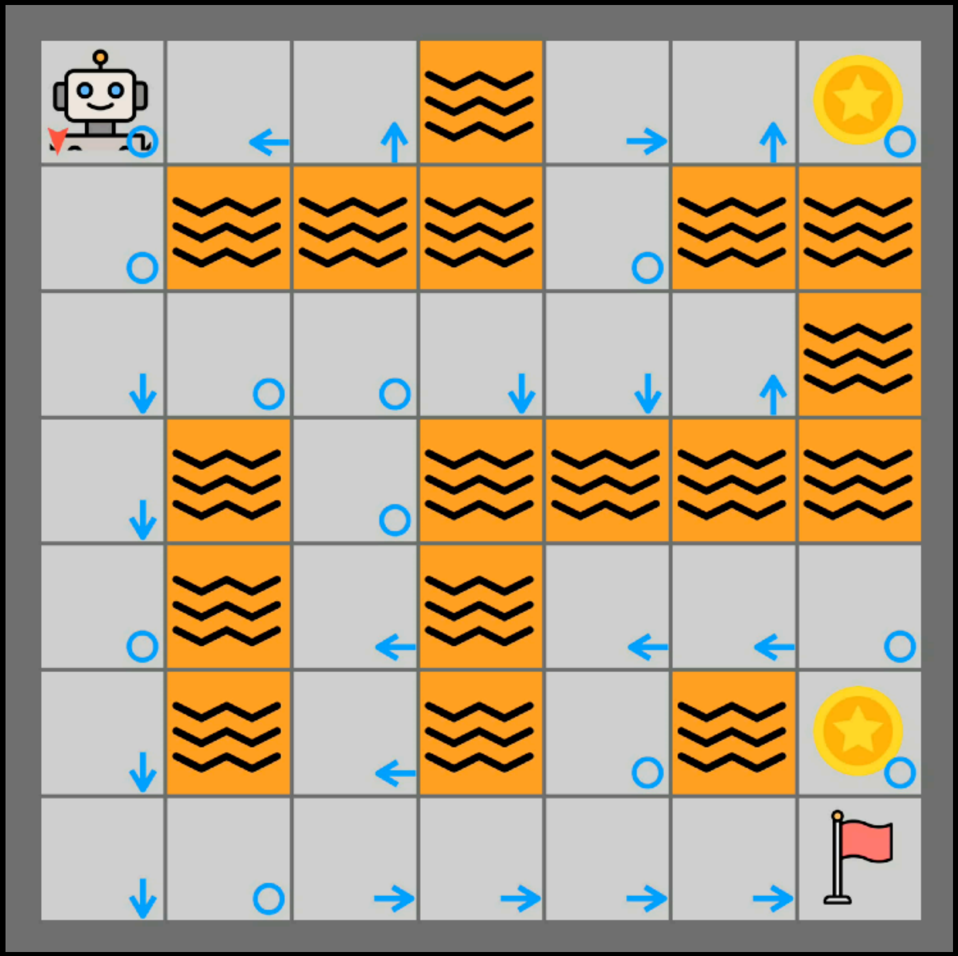
Task 8 (ICML'25)

Automatic Reward Shaping from Offline (confounded) data

Example 2: Reward Shaping

- A robot is in a maze where the walls are made of lava, which is highly lethal!
- The agent's movements are affected by the wind (indicated by blue marks in the plot), which it cannot perceive.
- How can the agent escape the maze without getting hurt?
- More broadly, how can we design a reward-shaping function that enables the agent to minimize online experimentation?

[crl.causalai.net]

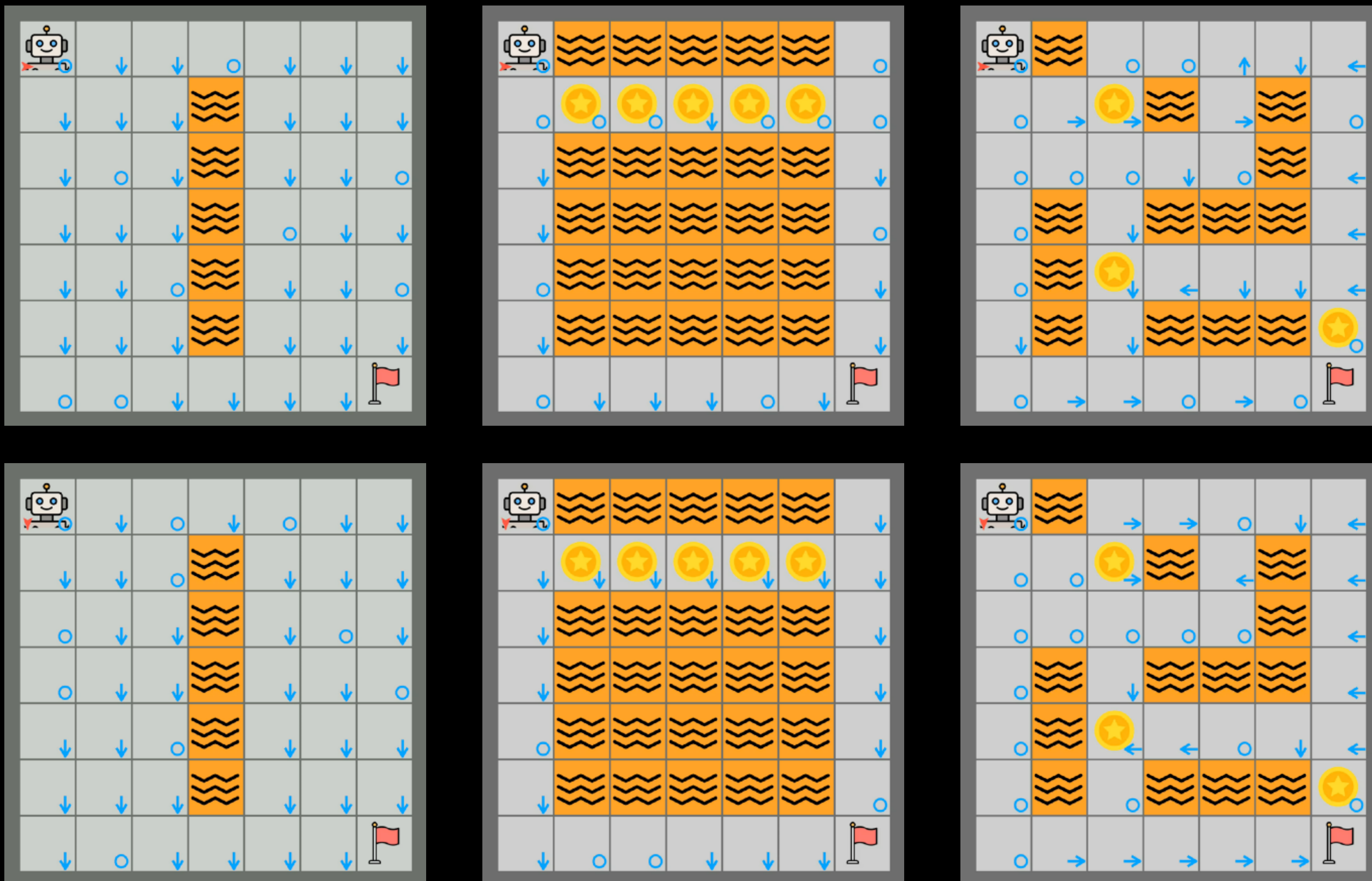


Example 2: Reward Shaping

[crl.causalai.net]

Baseline

Causal



CRL NEW CHALLENGES & OPPORTUNITIES (III)

[crl.causalai.net]

Task 7 (ICLR'24)

Causally Aligned Curriculum Learning

Task 8 (ICML'25)

Automatic Reward Shaping from Offline (confounded) data

Task 9 (TR-125)

Strategic (multi-agent) settings & Causal Game Theory

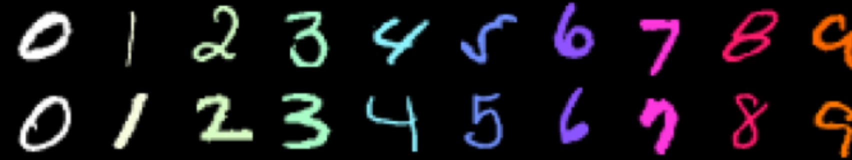
Causal AI — Desiderata

The new generation of AI systems is expected to provide the following capabilities:

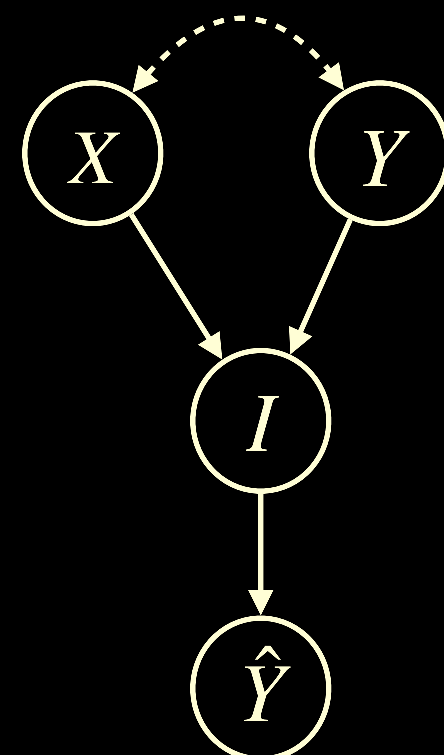
1. Causal Understanding & Explanations
2. Efficient & Precise Decision-Making
3. Generalizable & Robust Inferences
4. Causal Generative Capabilities
5. Model Learning & Discovery

Counterfactual Robustness

- Color X, Digit Y, Image I.



- Every digit has a different color, and saturation increases with the digit number.



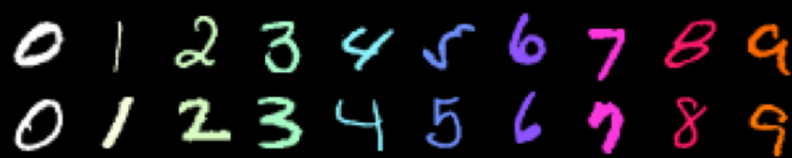
- A classifier outputs a prediction $\hat{Y} = \hat{y}$ given a colored image.

- **Why query**: given a sample $(x, y, i, \hat{y})$, why did the classifier predict $\hat{Y} = \hat{y}$?

Counterfactual Robustness

Consider **standard** vs. **robust** (greyscale) classifiers $\hat{Y}$.

Training distribution



Shifted distribution

Standard	100%	18.1%
Robust	99.4%	99.4%

A good explanation (answer to the why query) should be able to:

Task 1: Why the classifier predicted the digit the way it did.
An explanation should distinguish **standard** and **robust** classifiers.

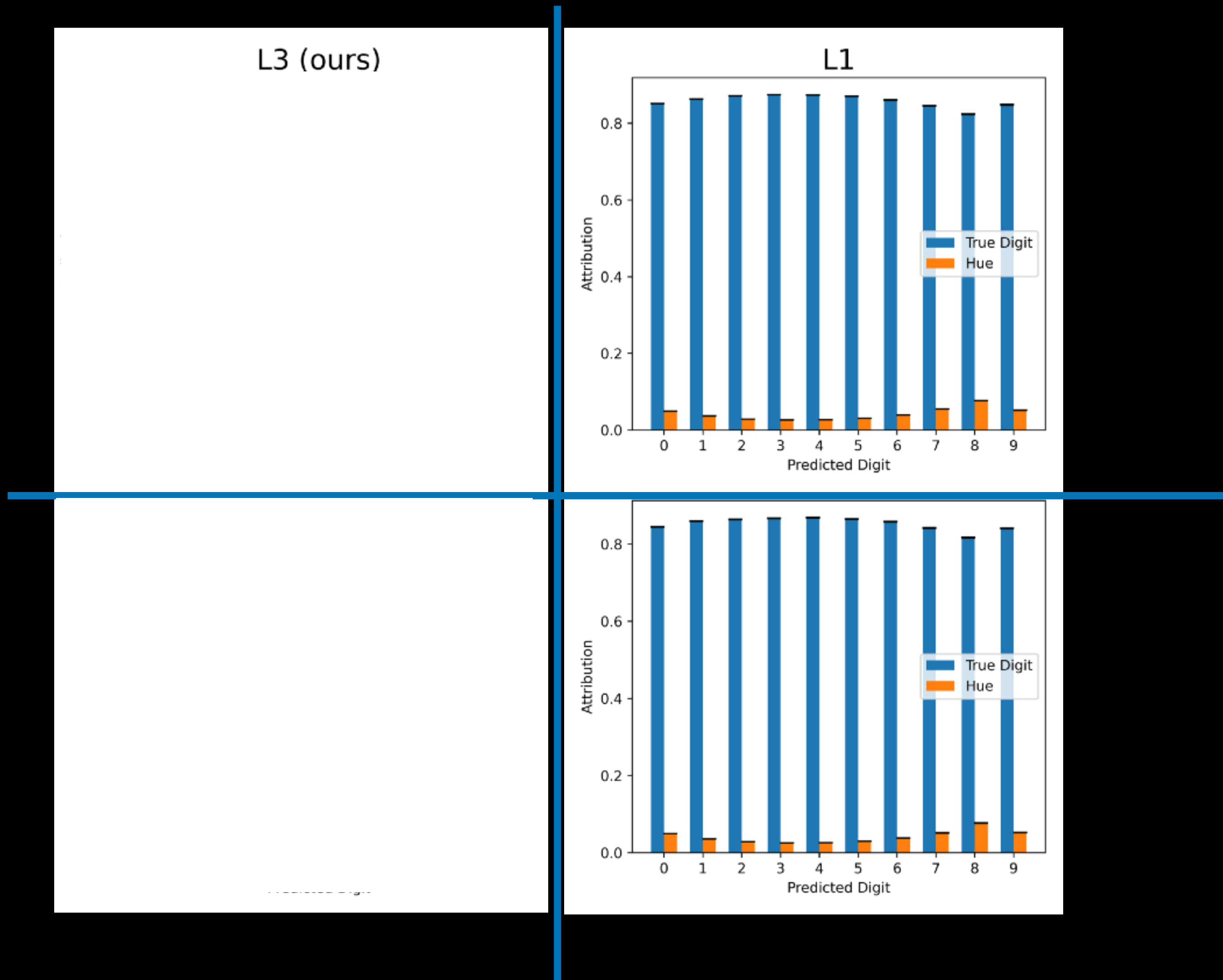
Task 2: Determine when variables (e.g., color X when digit $Y = 0$) have no effect on prediction $\hat{Y}$.

Counterfactual Robustness

We compare the explanation methods: SHAP (L1, Lundberg et al., 2017) and counterfactual Shapley values (L3, **ours**).

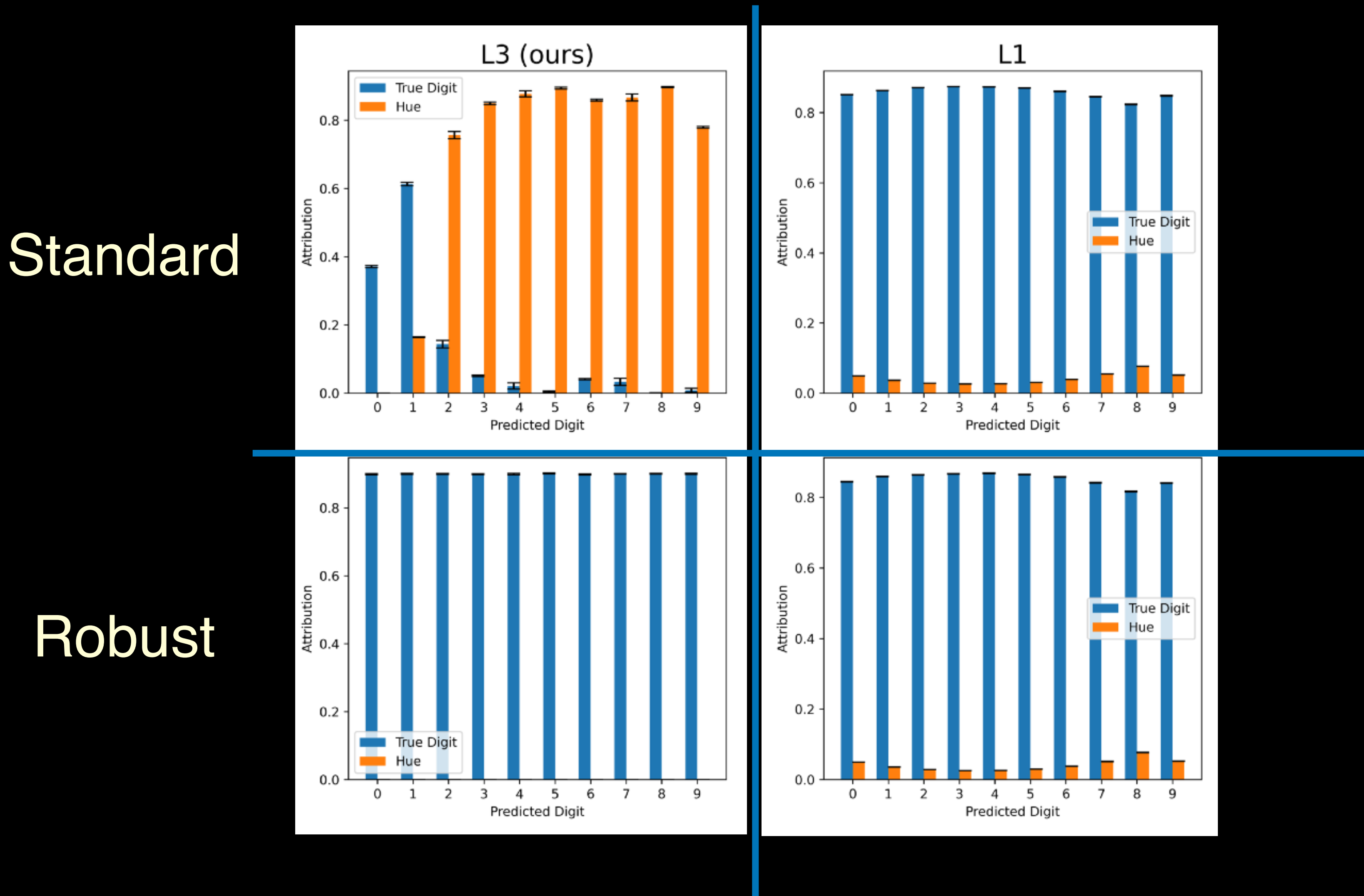
Standard

Robust



Counterfactual Robustness

We compare the explanation methods: SHAP (L1, Lundberg et al., 2017) and counterfactual Shapley values (L3, **ours**).

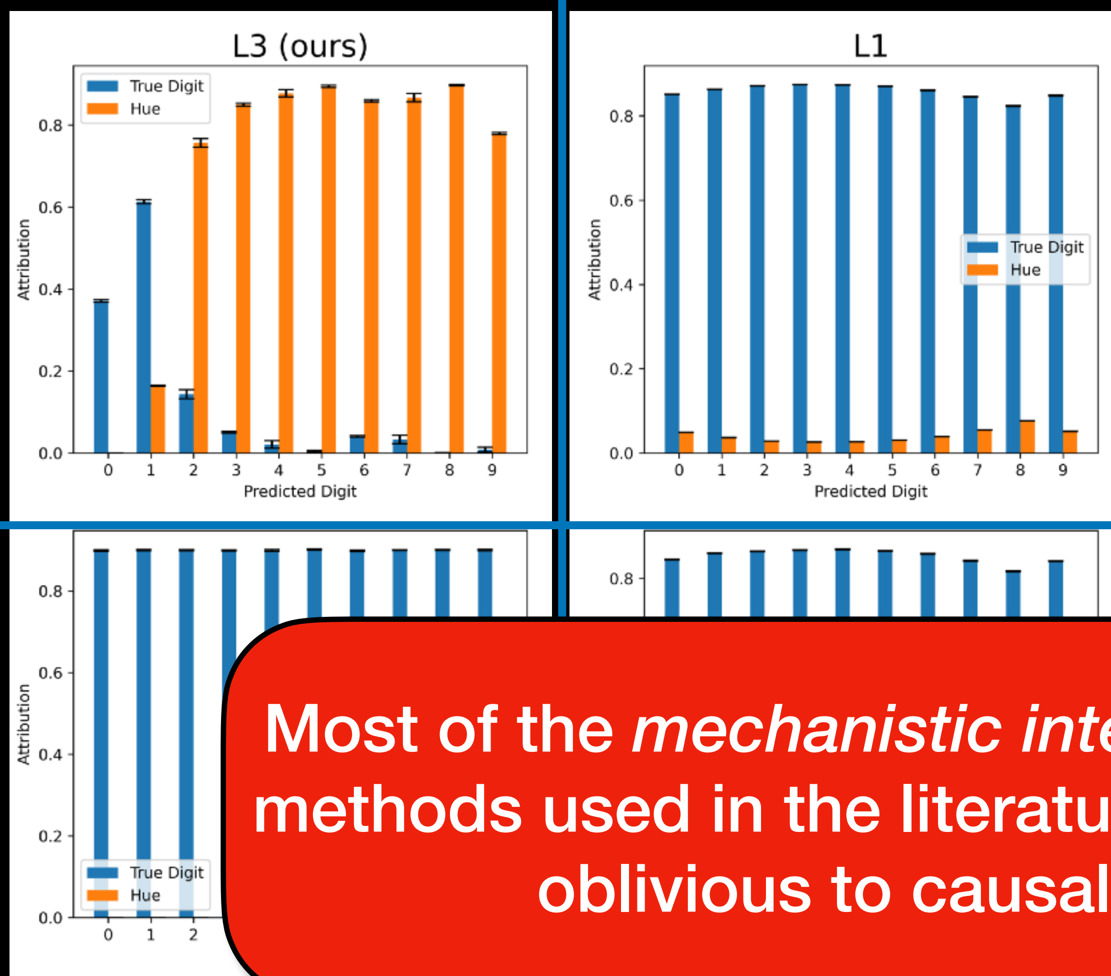


Counterfactual Robustness

We compare the explanation methods: SHAP (L1, Lundberg et al., 2017) and counterfactual Shapley values (L3, **ours**).

Standard

Robust



Most of the *mechanistic interpretability* methods used in the literature today are oblivious to causality.

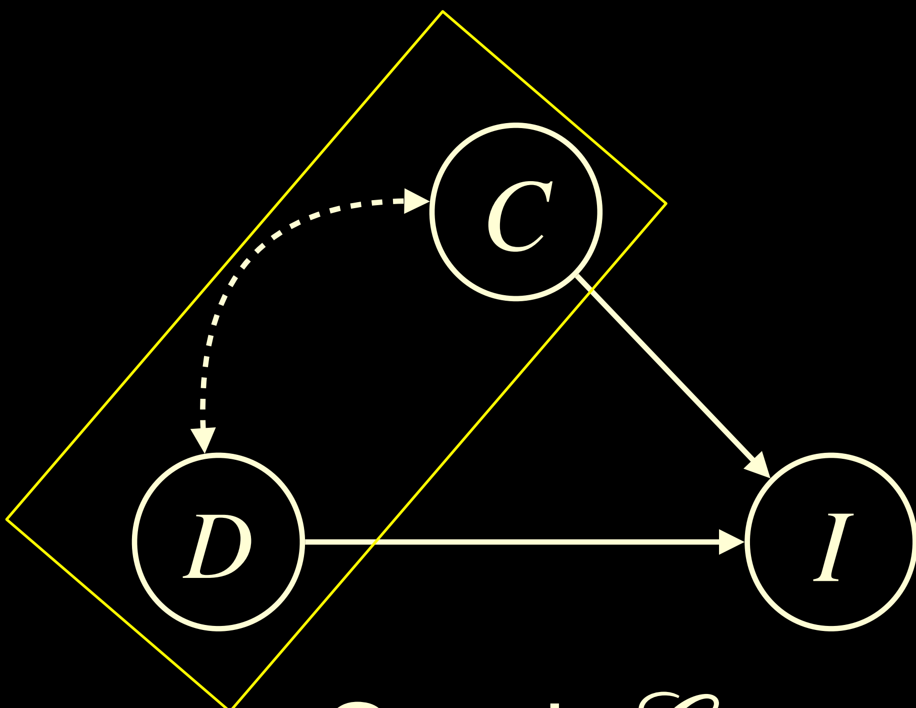
Causal AI — Desiderata

The new generation of AI systems is expected to provide the following capabilities:

1. Causal Understanding & Explanations
2. Efficient & Precise Decision-Making
3. Generalizable & Robust Inferences
4. Causal Generative Capabilities
5. Model Learning & Discovery

Example 1: Generative Modeling

Colored MNIST dataset



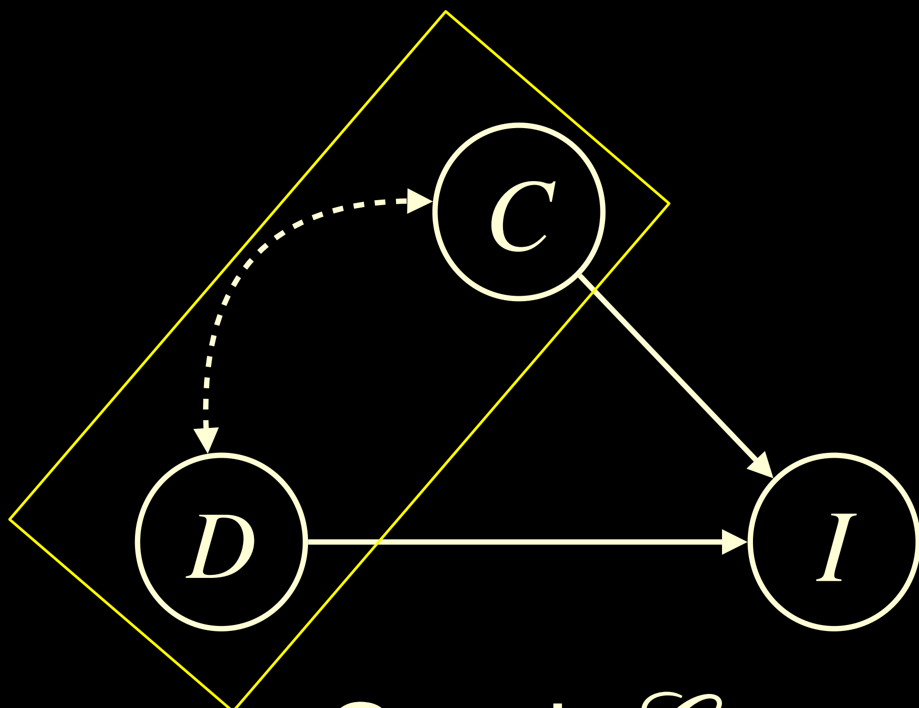
Graph $\mathcal{G}$

$\mathbf{V} = \{C, D, I\}$

- C: color, 10d one-hot
- D: digit, 10d one-hot
- I: image, $\mathbb{R}^{32\times32\times3}$

Example 1: Generative Modeling

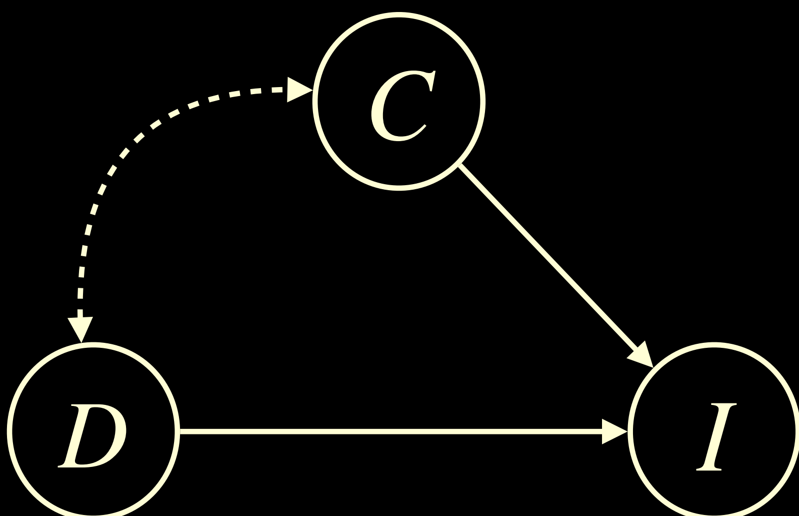
Colored MNIST dataset



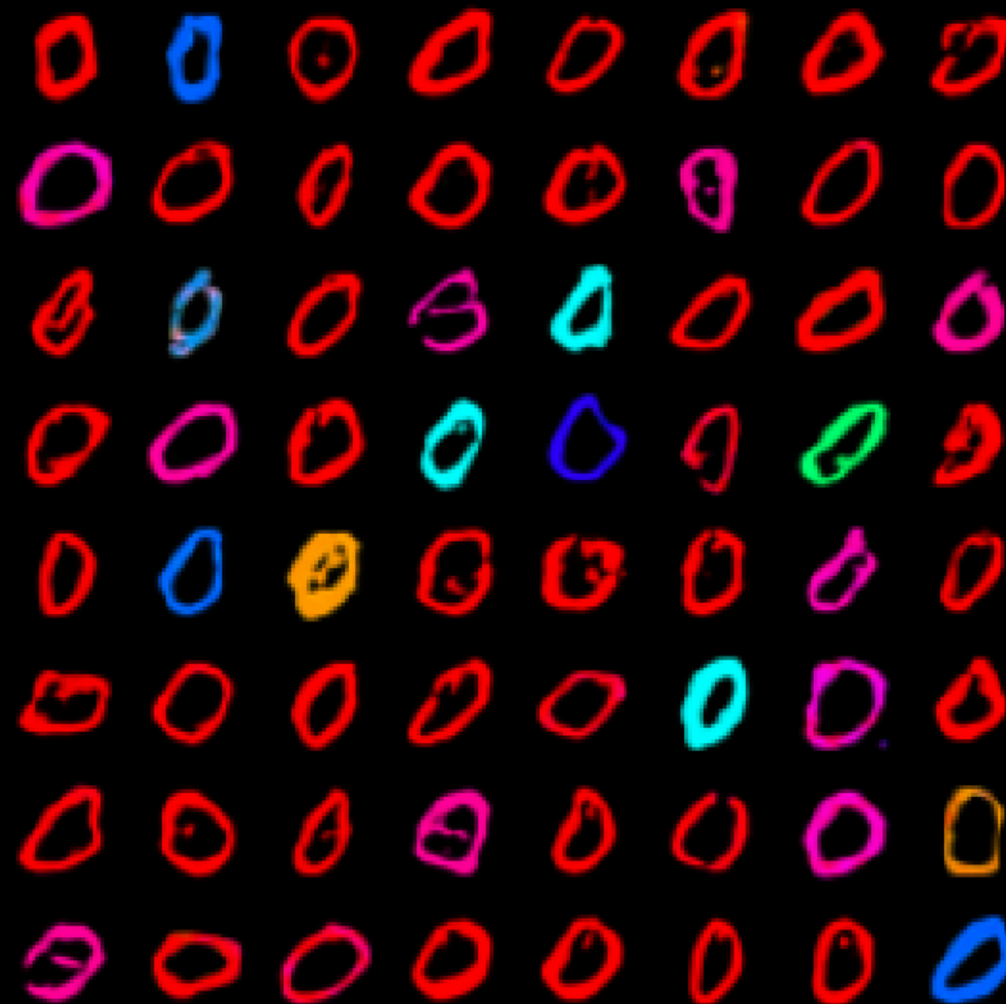
Graph $\mathscr{G}$

Conditional Query (L_1)

$Q = P(I \mid D=0)$

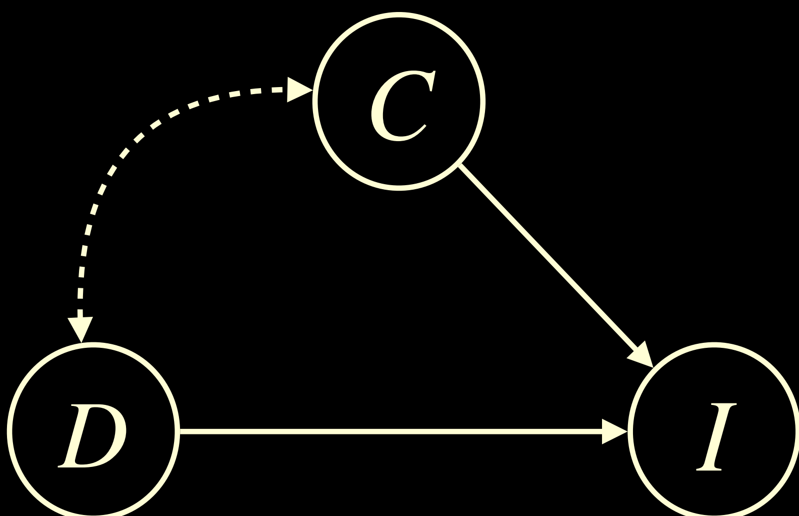


Graph $\mathcal{G}$

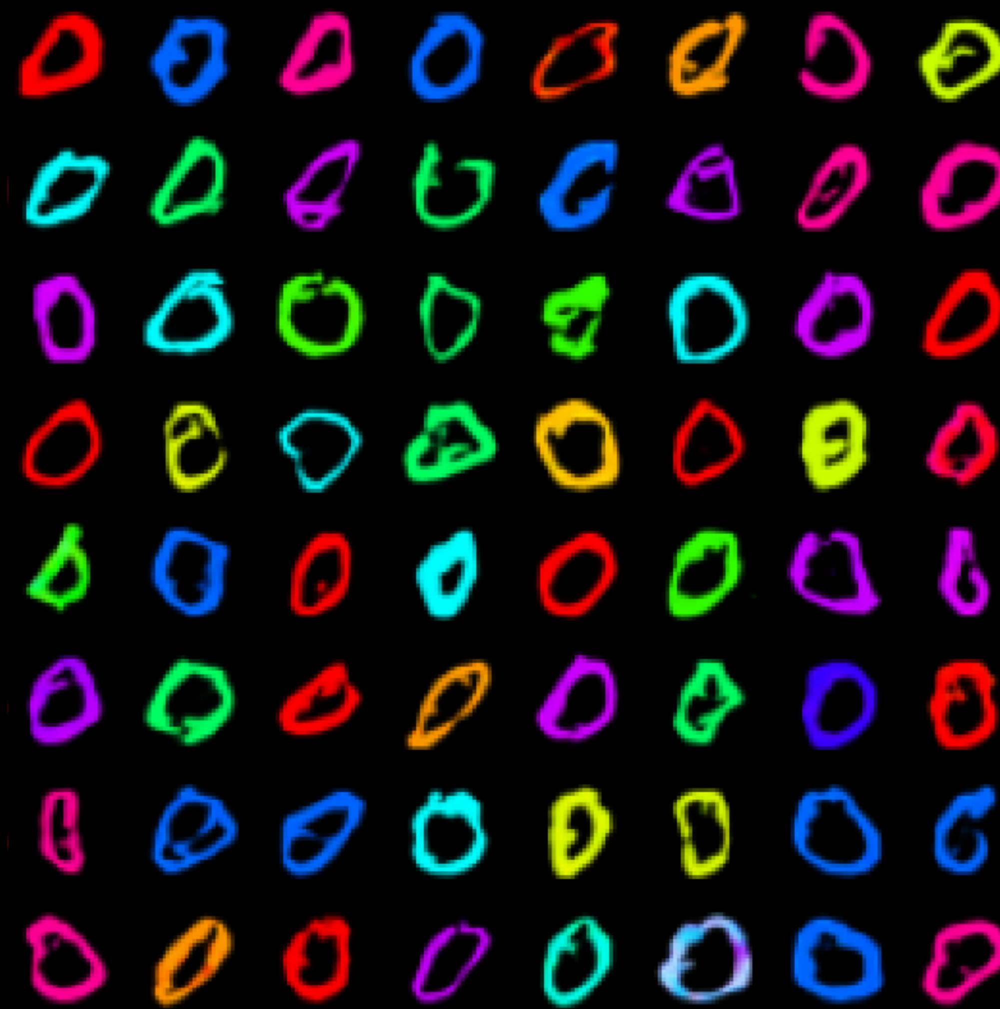


Interventional Query (L_2)

$$Q = P(I \mid do(D=0))$$

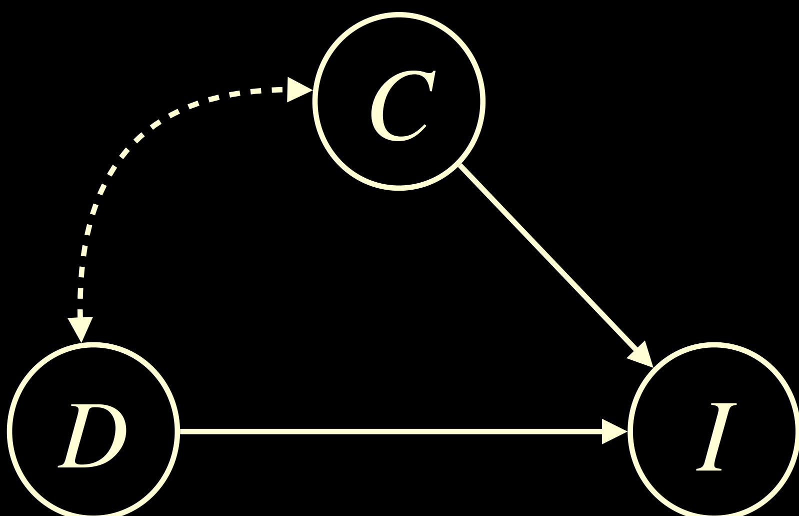


Graph $\mathcal{G}$

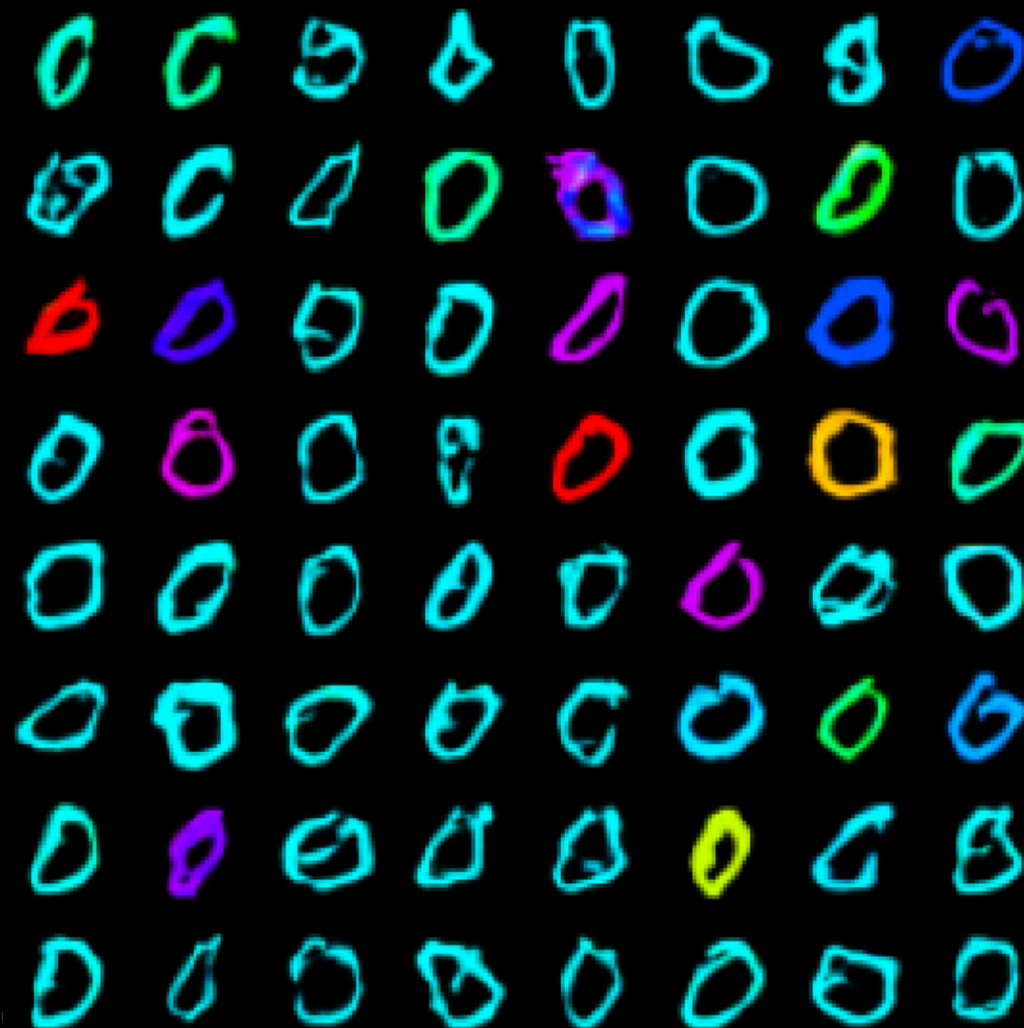


Counterfactual Query (L_3)

$$Q = P(I_{D=0} \mid D=5)$$



Graph $\mathcal{G}$



Example 2. Counterfactual Generation

- What would a person look like had they been … ?

Counterfactual Image Editing, Pan, Bareinboim, ICML-24.
Counterfactual Image Editing with Disentangled Causal Latent Space, Pan, Bareinboim, 2025.

Example 2. Counterfactual Generation

- What would a person look like had they been … ?

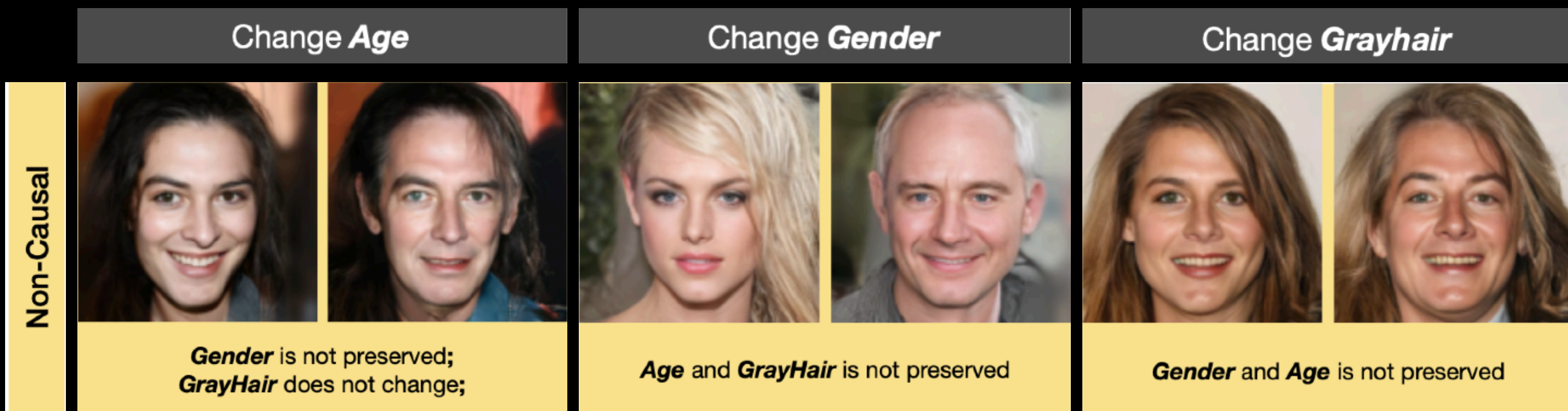
Change ***Age*** | Change ***Gender*** | Change ***Grayhair***

Non-Causal

Counterfactual Image Editing, Pan, Bareinboim, ICML-24.
Counterfactual Image Editing with Disentangled Causal Latent Space, Pan, Bareinboim, 2025.

Example 2. Counterfactual Generation

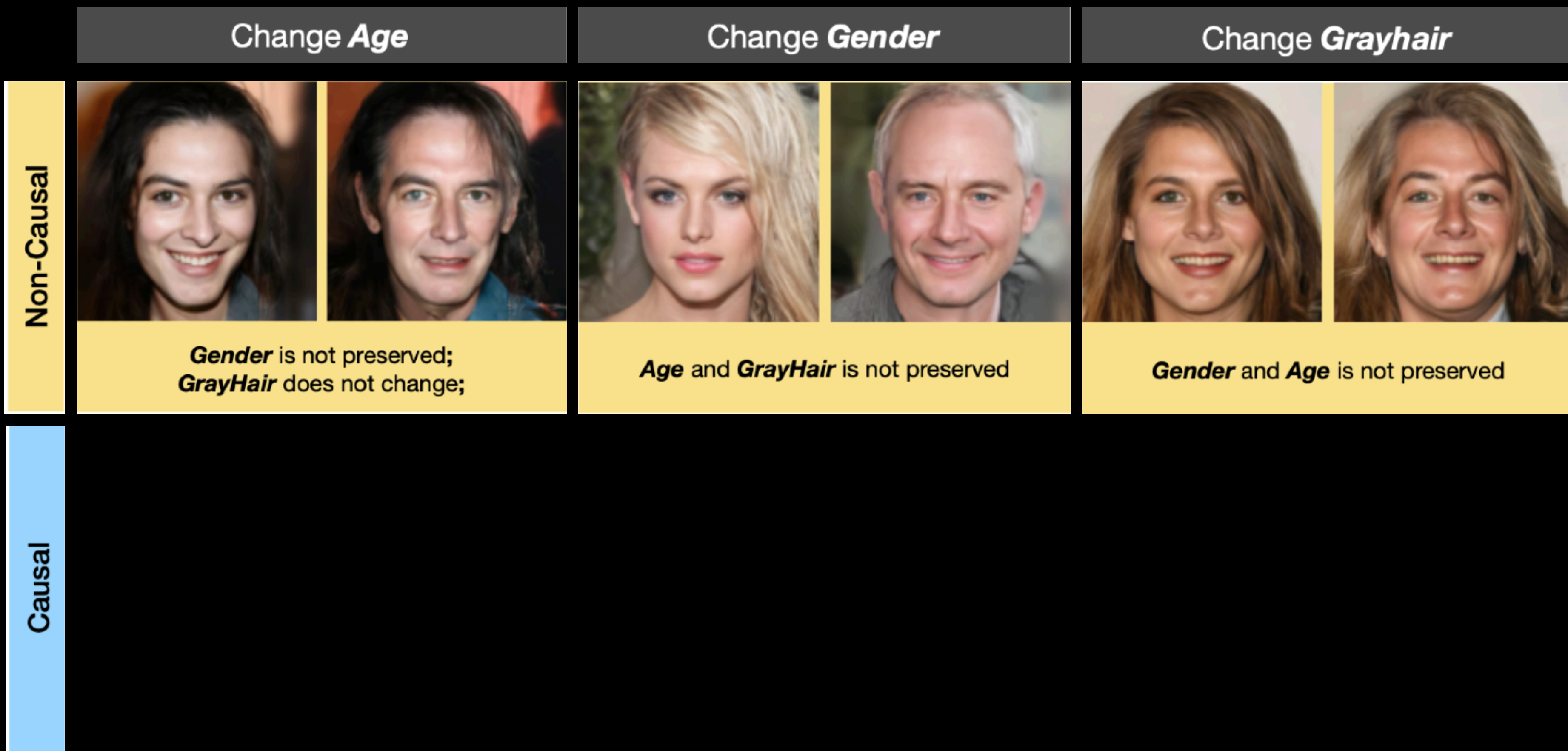
- What would a person look like had they been … ?



Counterfactual Image Editing, Pan, Bareinboim, ICML-24.
Counterfactual Image Editing with Disentangled Causal Latent Space, Pan, Bareinboim, 2025.

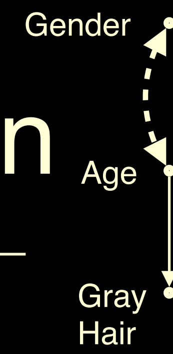
Example 2. Counterfactual Generation

- What would a person look like had they been … ?

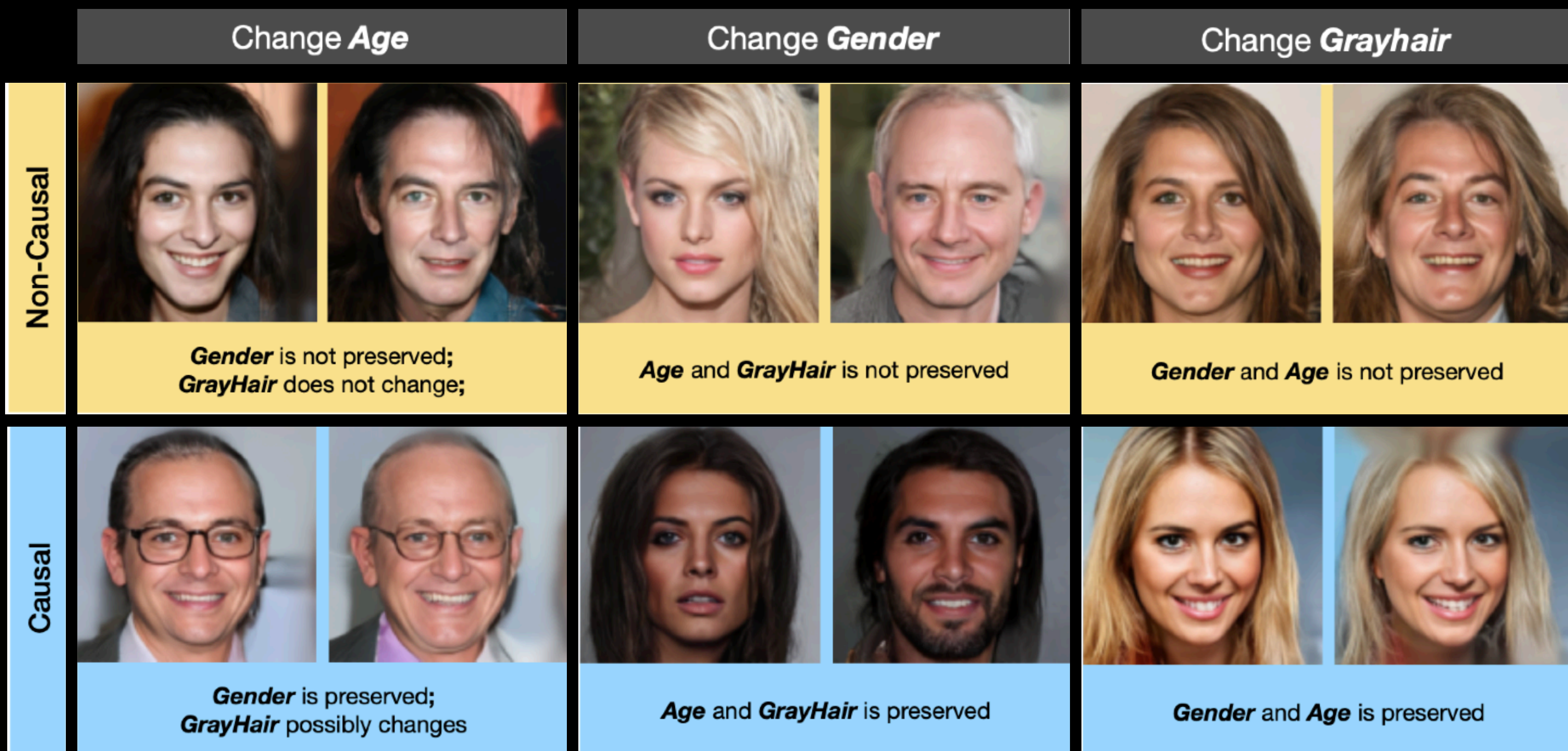


Counterfactual Image Editing, Pan, Bareinboim, ICML-24.
Counterfactual Image Editing with Disentangled Causal Latent Space, Pan, Bareinboim, 2025.

Example 2. Counterfactual Generation



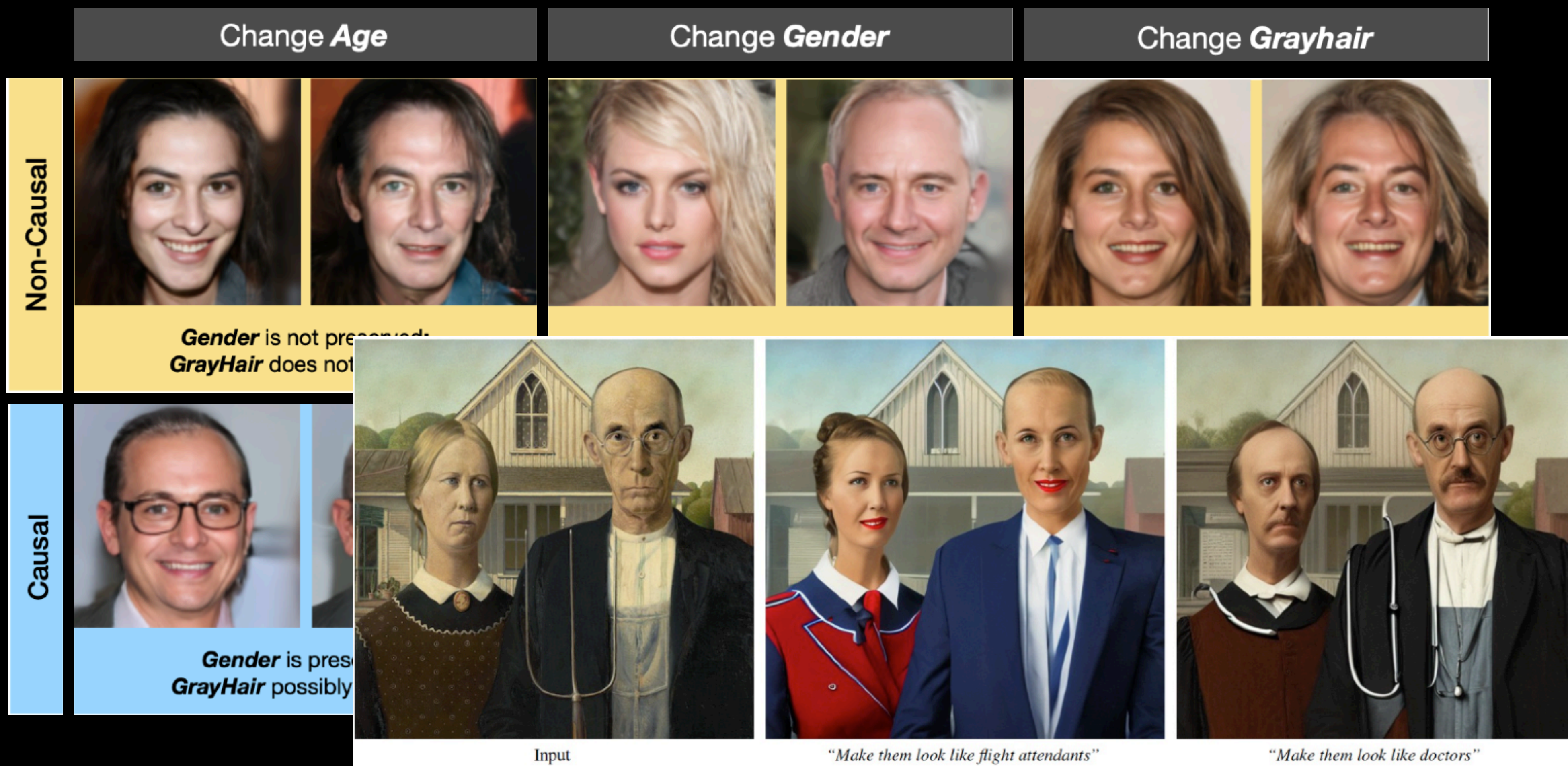
- What would a person look like had they been … ?



Counterfactual Image Editing, Pan, Bareinboim, ICML-24.
Counterfactual Image Editing with Disentangled Causal Latent Space, Pan, Bareinboim, 2025.

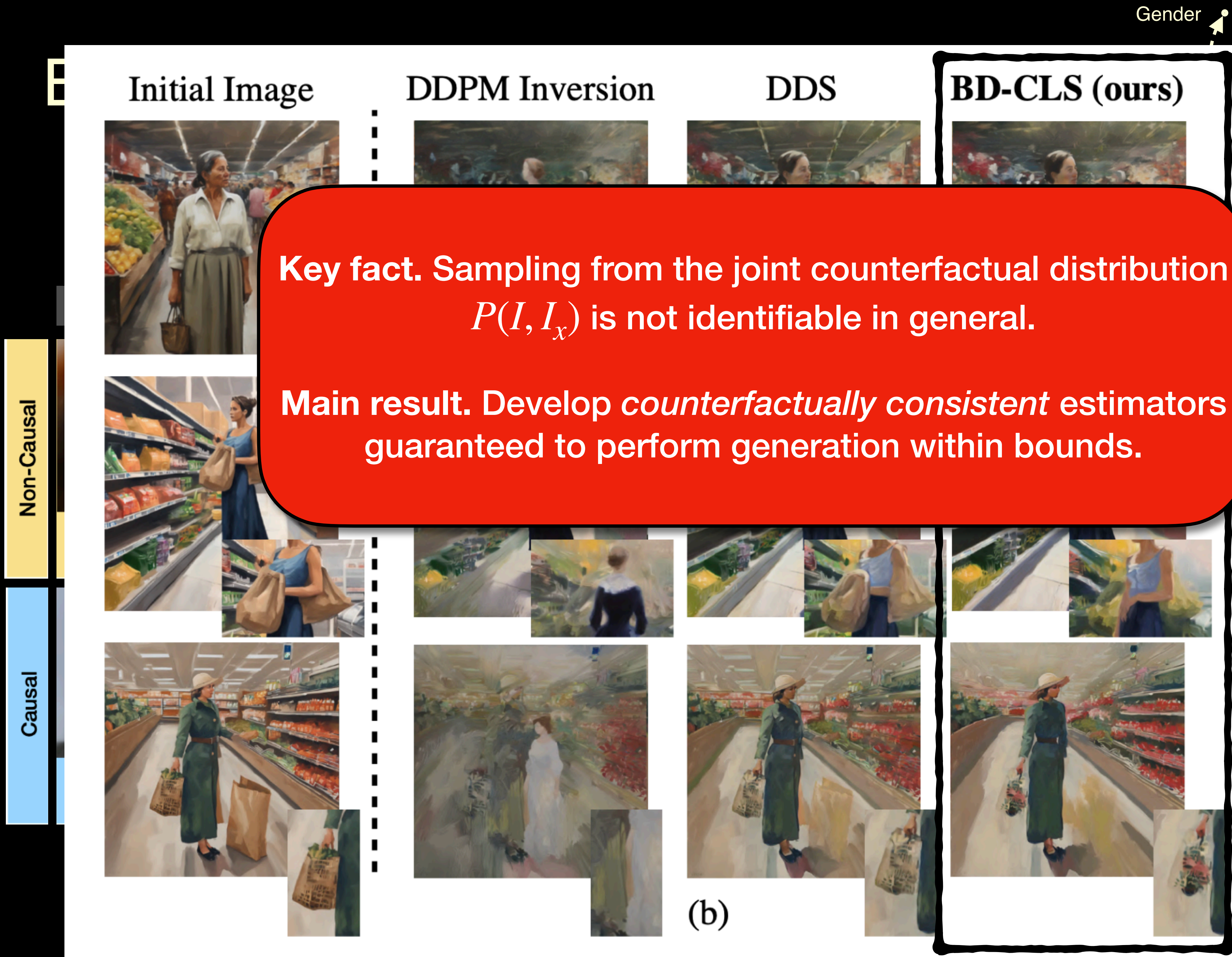
Example 2. Counterfactual Generation

- What would a person look like had they been … ?



Counterfactual Image Editing, Pan, Bareinboim, ICML-24.
Counterfactual Image Editing with Disentangled Causal Latent Space, Pan, Bareinboim, 2025.


Gender
Initial Image
DDPM Inversion
DDS
BD-CLS (ours)
Non-Causal
Causal
(b)



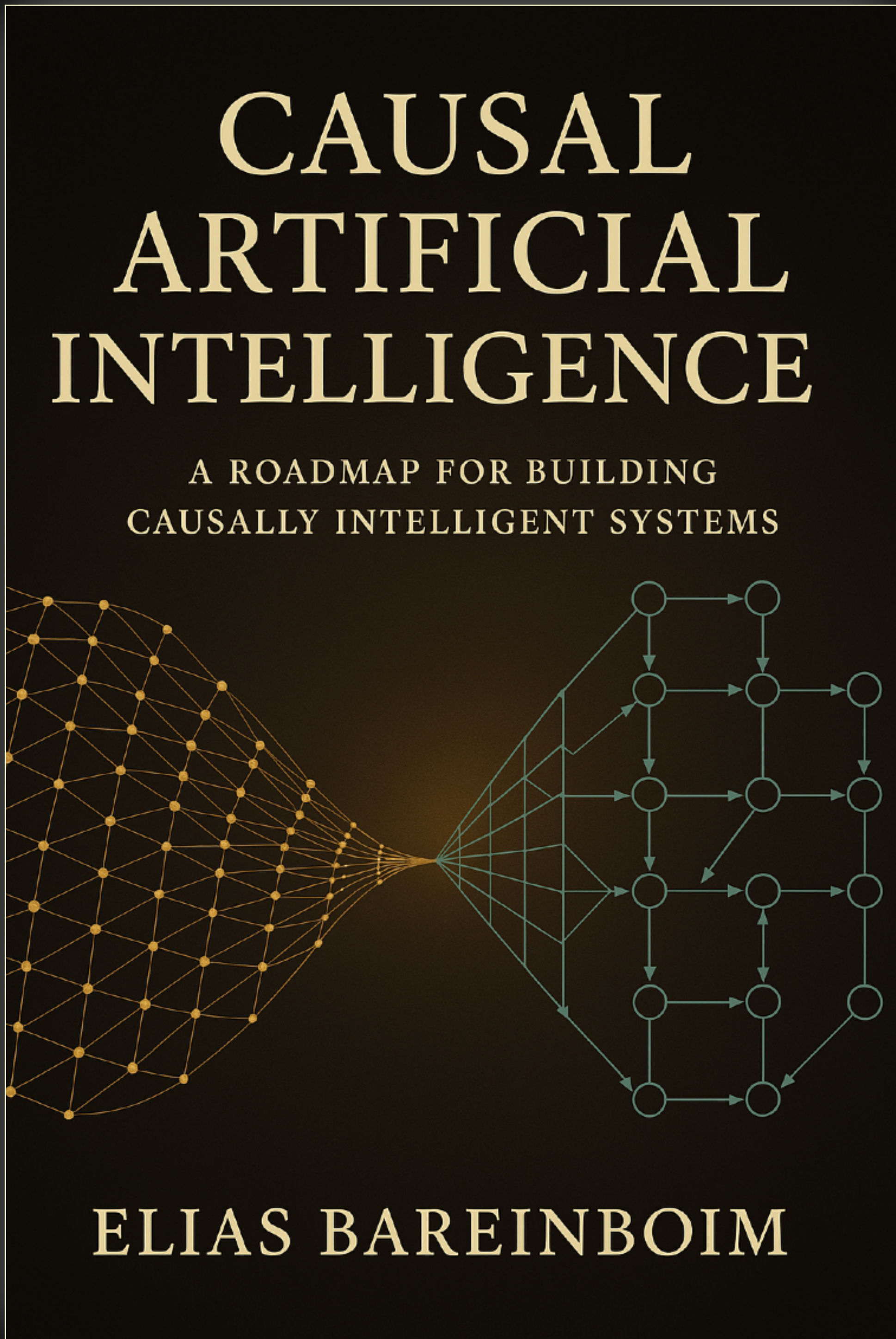
Key fact. Sampling from the joint counterfactual distribution $P(I, I_x)$ is not identifiable in general.

Main result. Develop *counterfactually consistent* estimators guaranteed to perform generation within bounds.

Causal AI Research Program

Develop more general & trustworthy AI systems endowed with the following capabilities:

1. Causal Understanding & Explanations
2. Efficient & Precise Decision-Making
3. Generalizable & Robust Inferences
4. Causal & Counterfactual Generation
5. Model Learning & Discovery

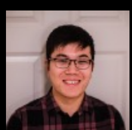


http://causalai-book.net

See also:
http://llm-observatory.org

THANK YOU, NSF, DARPA, AFOSR, ONR, NIH & CAUSAL AI LAB AND COLLABORATORS





Tara Anand (Columbia)
Adiba Ejaz (Columbia)
Inwoo Hwang (Columbia)
Kasra Jalaldoust (Columbia)
Drago Plecko (Columbia)
Kai-Zhan Lee (Columbia)
Mingxuan Li (Columbia)
Yushu Pan (Columbia)
Kevin Xia (Columbia)
Shreyas Havaldar (Columbia)
Jeffrey Wu (Columbia)
Arvind Raghavan (Columbia)
Hongshuo Yang (Columbia)
Yonghan Jung (Purdue)
Kevin Xia (Columbia)

Judea Pearl (UCLA)
Yoshua Bengio (MILA)
Carlos Cinelli (UW)
Andrew Forney (UCLA)
Duligur Ibeling (Stanford)
Thomas Icard (Stanford)
Murat Kocaoglu (Purdue)
Jin Tian (MBZUAI)
Ivan Diaz (NYU)
Sanghack Lee (SNU)
Karthikeyan Shanmugam (Google)
Juan Correa (UA Manizales)
Paul Hünermund (Copenhagen)
Junzhe Zhang (Syracuse U)