DeepMind

Towards Causal Foundations of Safe AI

Tom Everitt, Lewis Hammond, and Jon Richens

AAAI Tutorial
February 8, 2023

Causal Incentives Working Group

causalincentives.com

Tom Everitt
DeepMind

Lewis Hammond
Cooperative AI, Oxford

Jon Richens
DeepMind

Zac Kenton
DeepMind

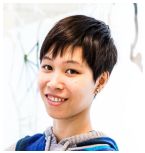
Carolyn Ashurst
Alan Turing Institute

Ryan Carey
Oxford

Ramana Kumar
DeepMind

Francis Rhys Ward
Imperial

Eric Langlois
Toronto



Mary Phuong
DeepMind

Chris van Merwijk
CMU

Matt MacDermott
Imperial

Shreshth Malik
Oxford

Hal Ashton
UCL

James Fox
Oxford

Sebastian Farquhar
DeepMind

What is this tutorial about?

Causality as a unifying theory for analysing safety problems in AI

What's your background?

- **Causality**: learn about important problems that can be addressed using causality
- **Safety**: learn how ideas from causality can formalise and unify safety problems
- **Neither**: hopefully gain some good insights into both!

Only minimal background knowledge required, though we will go through the basics relatively quickly. These slides are online if you want to follow along or revisit sections!

Outline

Background (15 mins, Lewis)

Introduce the models

- Pearl's hierarchy

Modelling Agents (20 mins, Lewis)

Agents and decisions can be modelled causally

- Modelling agents
  - Influence diagrams
  - Causal games
  - Other models
- Discovering agents

Fairness (10 min, Tom)

- Counterfactual fairness
- Path-specific fairness
- Value of Information and Proxies

Misspecification (15 min, Tom)

- Preference manipulation
- Recursion, interpretability
- Decision theory
- Path-specific objectives

Generalisation (30 min, Jon)

- Robustness
- Goal misgeneralisation
- Causal objectives
- Harm

Human control (10 min, Tom)

- Corrigibility
- Q-learning indifference
- Preserving human agency

DeepMind

Introduction

- Agency
- Goals
- Behavior
- Intent
- Morality
- Fairness
- Human control

DeepMind

Background

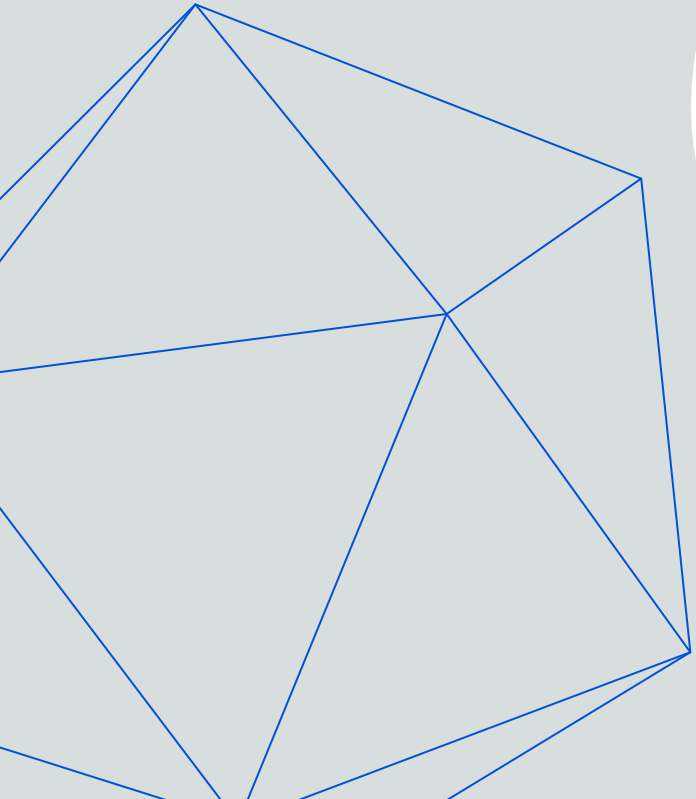
Overview

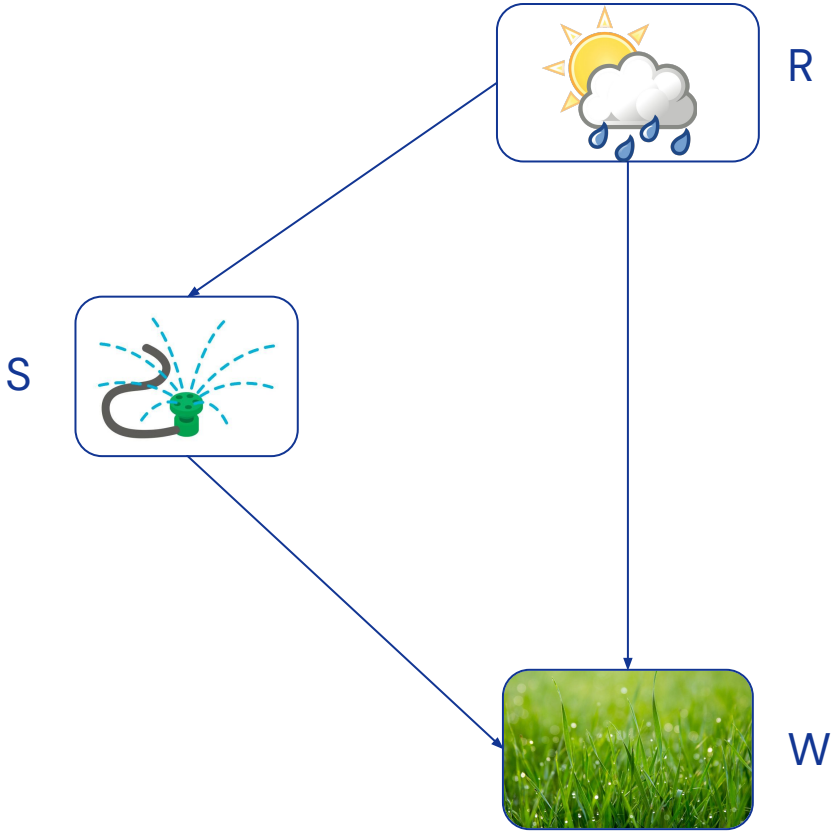
Aim: introduce the fundamental models and concepts that the rest of the tutorial will build on

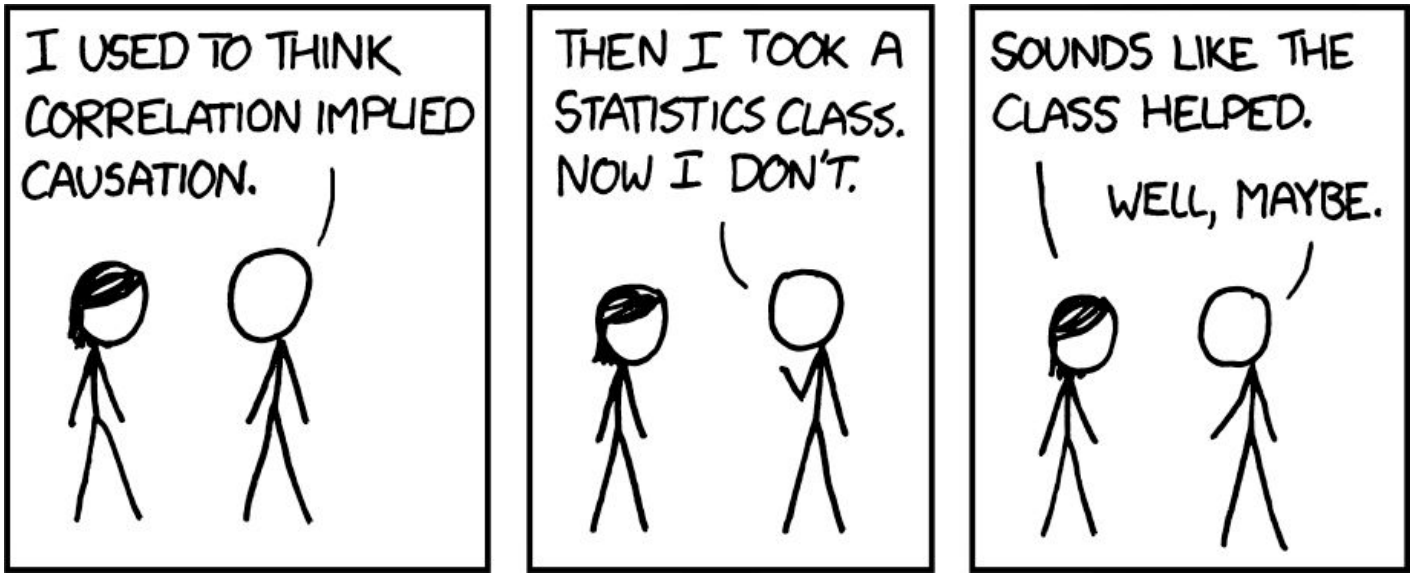
- Example
- Queries
  - Association
  - Intervention
  - Counterfactual
- Models
  - Bayesian networks
  - Causal Bayesian networks
  - Structural causal models

DeepMind

Example

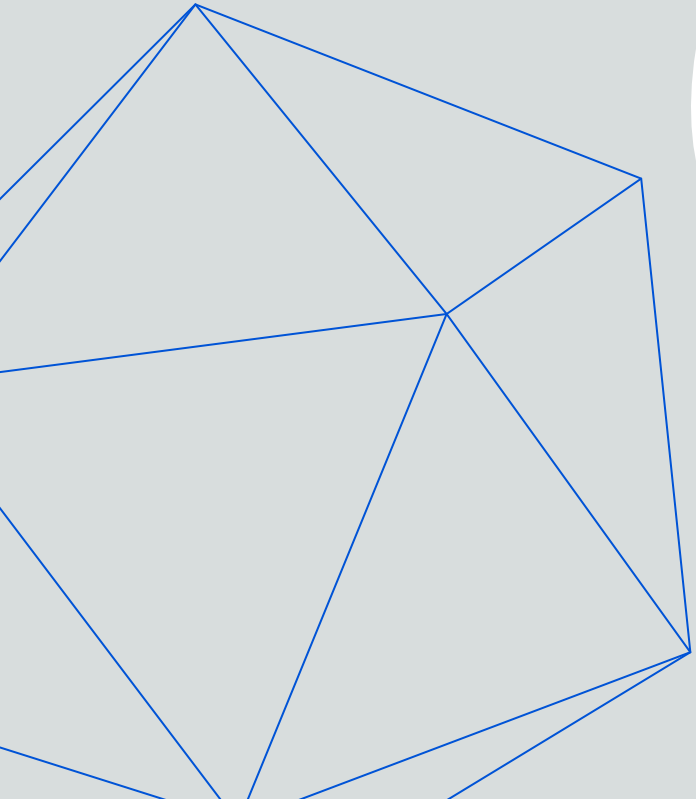



R
S
W


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

DeepMind

Queries



Pearl's ladder of causality

Is the grass dry after **turning off** sprinkler when we **observed** it was on?

$p(\neg w_{\neg s} \mid s)$

Is the grass dry when we **turn off** the sprinkler?

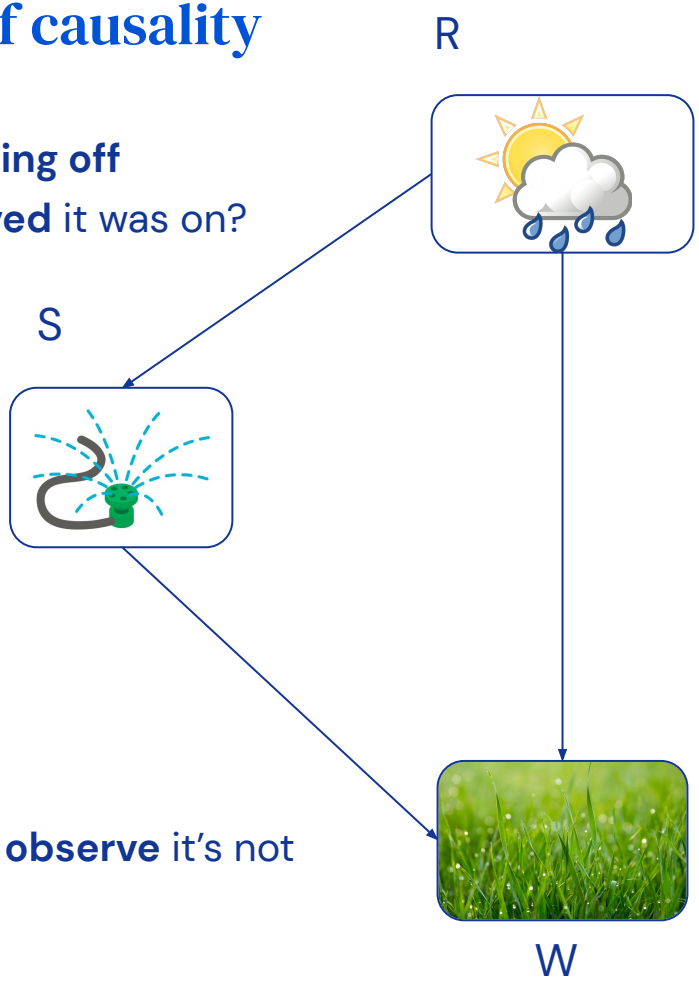
$p(\neg w \mid do(\neg s))$

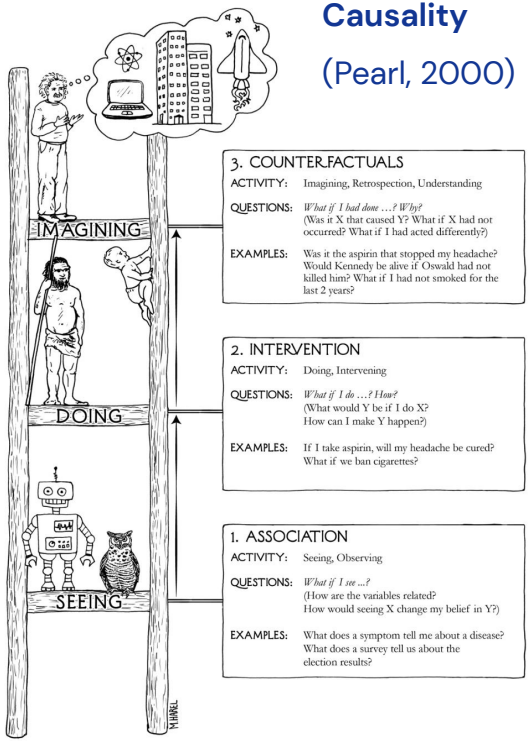
$= p(\neg w_{\neg s})$

$= p_{\neg s}(\neg w)$

Is the grass dry when we **observe** it's not raining?

$p(\neg w \mid \neg r)$



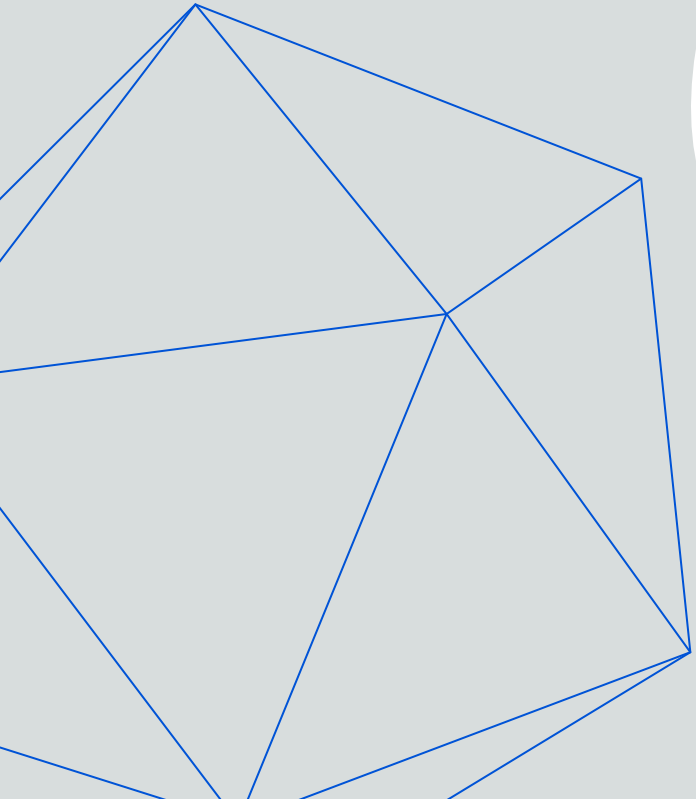


Causality

(Pearl, 2000)

DeepMind

Models



What now?

So far:

- Three kinds of questions
- Notation for writing them down – but what does this notation actually *mean* mathematically?

Key point: to answer different kinds of query, we need different amounts of information, which can be captured by different kinds of probabilistic (graphical) model

Model ingredients:

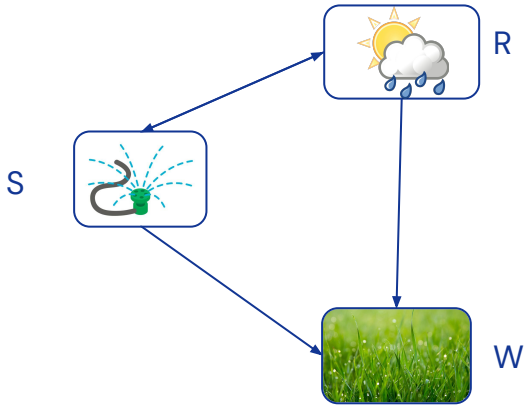
- Variables
- Relationships between variables

Association: Bayesian networks (BNs)

- Variables **V** with joint distribution p(**V**)
- Bayesian network M = (G , θ)
  - G = (**V** , E) is a DAG
    - V is a variable
    - v is a value of V
    - Pa_V are the parents of V in G
  - G is Markov compatible with p
    - $p(v ; \theta) = \Pi_V p(v \mid pa_V ; \theta_V)$
    - θ parameterises p

$$p(w \mid \neg s) = p(w , \neg s) / p(\neg s)$$

$$= \Sigma_{r'} p(w, \neg s, r') / \Sigma_{w', r'} p(w', \neg s, r')$$



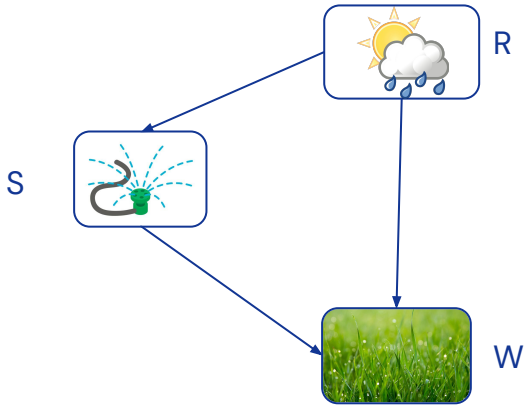
$$p(W, S, R) = p(W \mid S, R) p(S \mid R) p(R)$$

$$p(W, S, R) = p(W \mid S, R) p(R \mid S) p(S)$$

Intervention: causal Bayesian networks (CBNs)

Key point: to answer queries about interventions, we need a model that encodes what happens under any intervention

- A causal Bayesian network M = (G , θ) is a BN where:
    - For any $\mathbf{Y} \subseteq \mathbf{V}$ and value $\mathbf{y}$ of $\mathbf{Y}$, G is Markov compatible with $p_{\mathbf{y}}$
- Given this: $p_{\mathbf{y}}(\mathbf{v}) = \delta(\mathbf{Y} , \mathbf{y}) \Pi_{V \notin \mathbf{Y}} p(v \mid pa_V)$



$p(W, S, R) = p(W \mid S, R) p(S \mid R) p(R)$

~~$p(W, S, R) = p(W \mid S, R) p(R \mid S) p(S)$~~

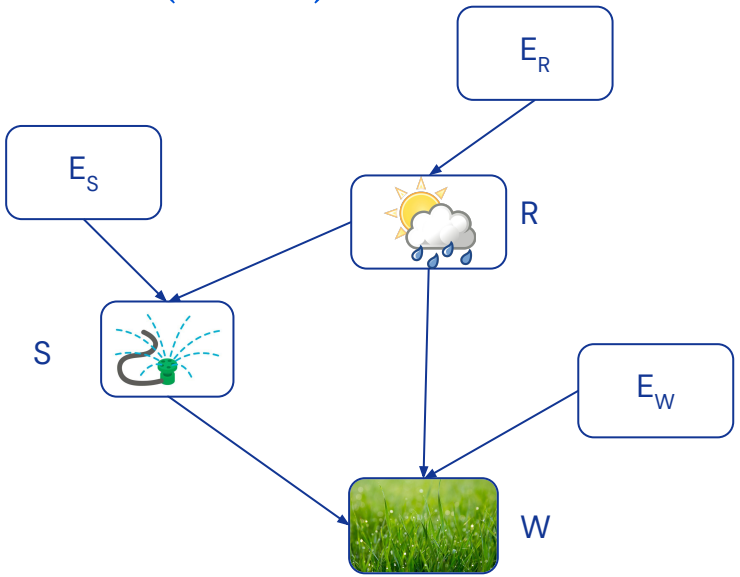
$$
\begin{aligned}
p(w \mid do(\neg s)) &= p_{\neg s}(w) \\
&= \Sigma_{s', r'} \delta(s' , \neg s) p(w' \mid s', r') p(r') \\
&= \Sigma_{r'} p(w' \mid \neg s, r') p(r')
\end{aligned}
$$

Counterfactuals: structural causal models (SCMs)

Key point: to answer queries about counterfactuals, we need a model that tells us what changes in the counterfactual world and what remains the same

- Structural causal model M = (**V**, **E**, θ, F)
  - Model split into unobserved *exogenous* variables **E** and *endogenous* variables **V**
  - p(**E** ; θ) encodes all randomness
  - Value of V given by deterministic functions f_V : dom(**V** \ {V}) x dom(**E**) → dom(V)

Note we can also view this a specific form of (C)BN, so we often simply use BN notation



Markovian SCM:

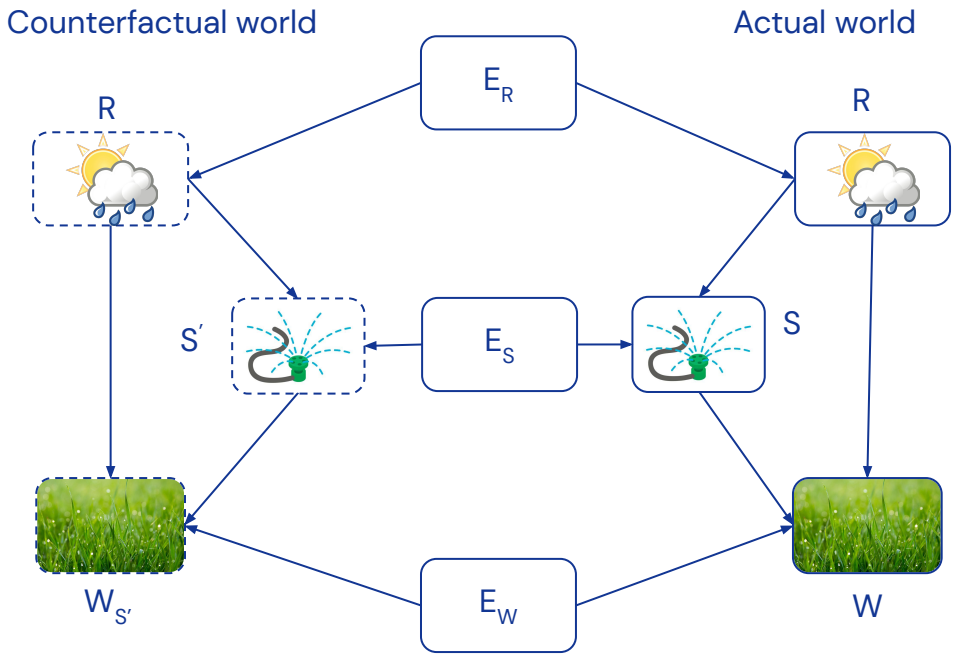
- p(**E** ; θ) = Π_E p(e ; θ_E)
- f_V : dom(**V** \ {V}) x dom(E_V) → dom(V)

Counterfactuals: structural causal models (SCMs)

Suppose that I *observe* that the sprinkler is on; what is the probability that the grass is wet given that I *intervene* and turn the sprinkler off?

We use a three step process to compute p($w_{\neg s}$ | s)

1. Update p(**E**) to p(**E** | s) ('abduction')
2. Replace f_S with S = ¬s ('intervention)
3. Return the marginal distribution p(w) in this modified model ('prediction')



DeepMind

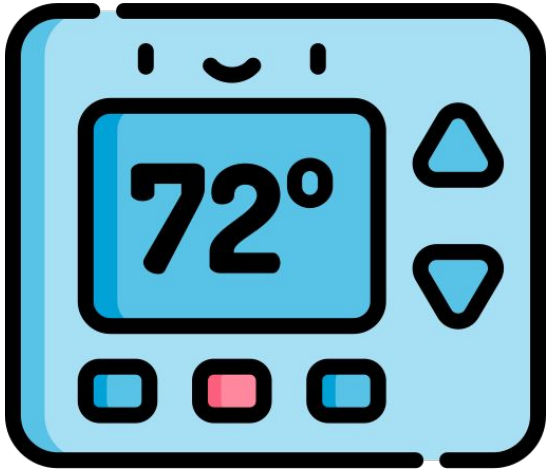
Modeling Agents

Overview

Aim: convey how agents and decisions can be modelled in causal graphs

- Modelling agents
  - Causal influence diagrams
  - Causal games
  - Other models
- Discovering agents
  - Mechanism graphs
  - Causal discovery of agents

The Intentional Stance
Dennett, 1989



Causal hierarchy in the presence of agents

Associational

- Is an automated hiring recommendation correlated with a sensitive attribute of the applicant?
- How strongly correlated are the prices set by two trading agents?

Interventional

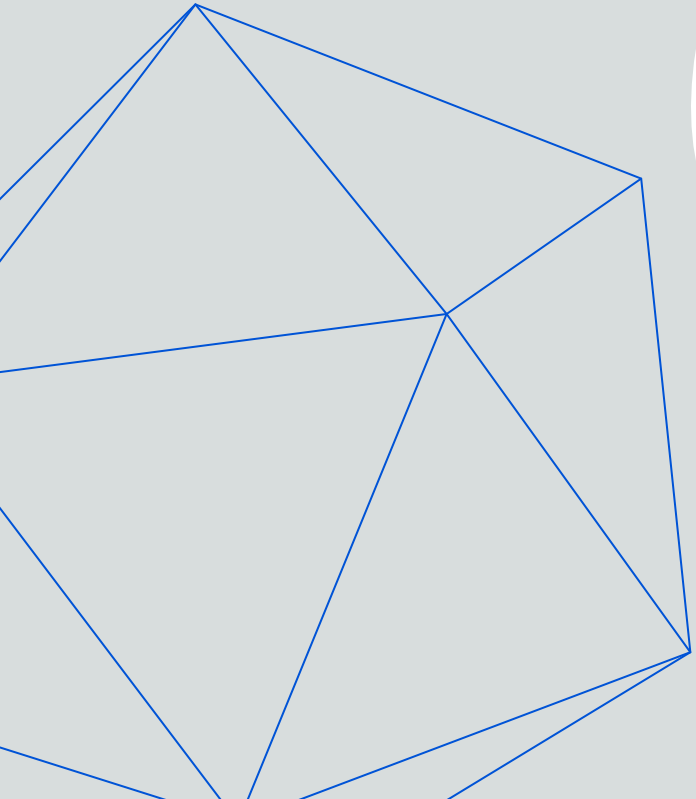
- Will a medical system perform correctly in a different hospital?
- how would sensory noise during training affect the output?

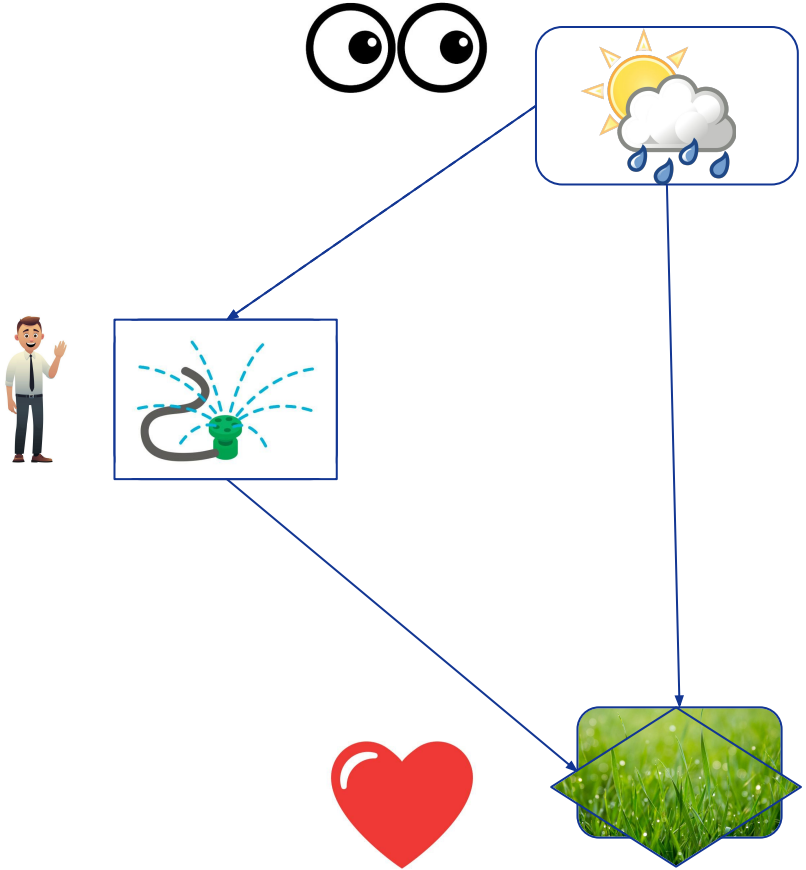
Counterfactual

- If their gender had been different, would the applicant have been allocated the job?
- If another robot had been able to help, would the first robot have taken the same action?

DeepMind

Modelling agents



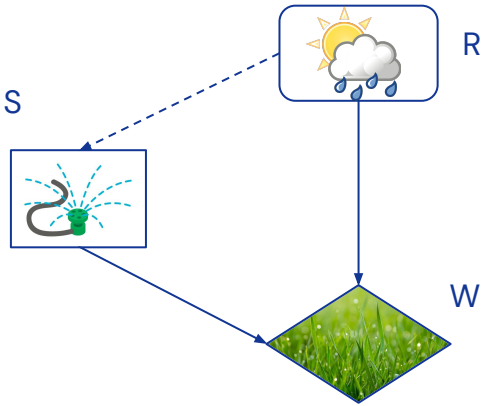


Causal influence diagrams

Influence Diagrams
(Howard and Matheson, 1984)

Agent Incentives: A Causal Perspective
(Everitt et al, 2021)

- A causal influence diagram (CID) is a CBN where:
  - The variables **V** are partitioned into:
    - Chance variables **C**
    - Decision variables **D**
    - Utility variables **U**
  - The decision variables are unparamerterised
- The agent selects a *policy* **π** made up of decision rules $\pi_D(D \mid Pa_D)$ for each decision variable D
- The agent gains expected utility $\mathbb{E}[\Sigma_U u]$



A policy induces a joint distribution over all variables as follows:

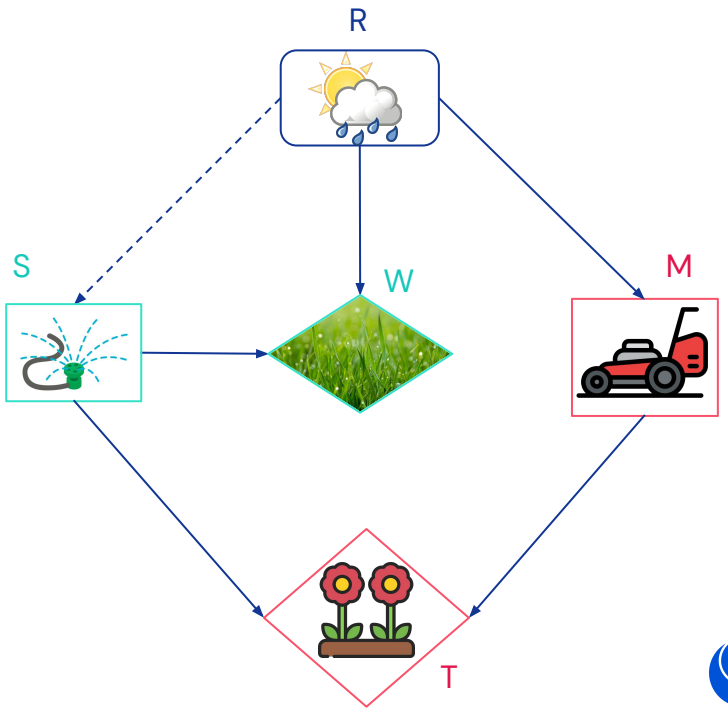
$$p^{\pi}(\mathbf{v}) = \Pi_{V \notin \mathbf{D}} p(v \mid pa_V) \Pi_{D \in \mathbf{D}} \pi_D(d \mid pa_D)$$

Causal games

- A causal game is a generalisation of CIDs to multiple agents $\{ 1, \dots, n \}$ such that:
  - $\mathbf{D} = \{ \mathbf{D}_1, \dots, \mathbf{D}_n \}$ and $\mathbf{U} = \{ \mathbf{U}_1, \dots, \mathbf{U}_n \}$
- A joint policy (or policy profile) $\boldsymbol{\pi} = (\boldsymbol{\pi}_{1, \dots,} \boldsymbol{\pi}_n)$ contains a policy $\boldsymbol{\pi}_i$ for each agent i
- The agent selects a *policy* $\boldsymbol{\pi}$ made up of decision rules $\pi_D(D \mid Pa_D)$ for each decision variable D
- Each agent gains expected utility $\mathbb{E}_{\boldsymbol{\pi}}[\Sigma_{U_j} u_j]$ where $U_j \in \mathbf{U}_i$

Multi-agent influence diagrams
(Koller and Milch, 2003)

Reasoning about Causality in Games
(Hammond et al., 2023)

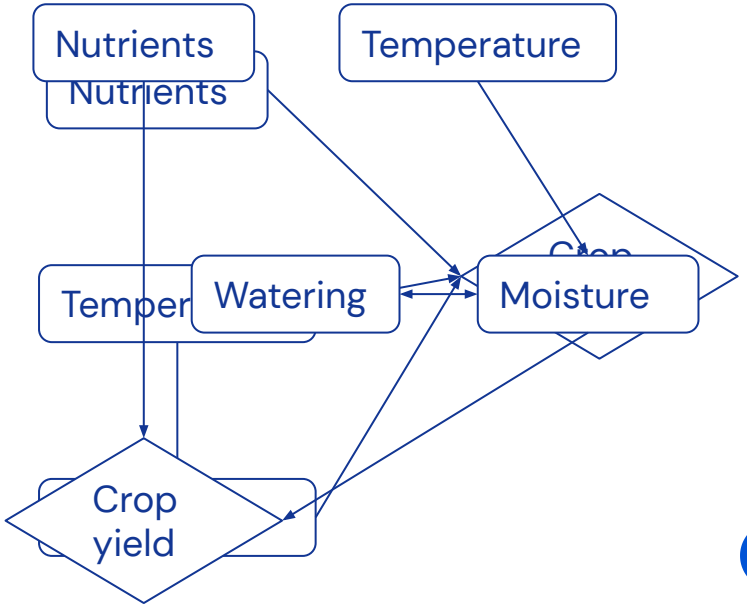


Other models

- There are also other causal models that capture some notion of agency
- In a causal bandit, the agent can choose which variable to intervene on (each variable is one arm)
    - Learning the causal structure helps helps us decide which arm to pull
- In a settable system, endogenous variables are duplicated into ‘response’ and ‘setting’ versions
    - The models can then be used to explicitly instantiate learning algorithms or optimisation processes

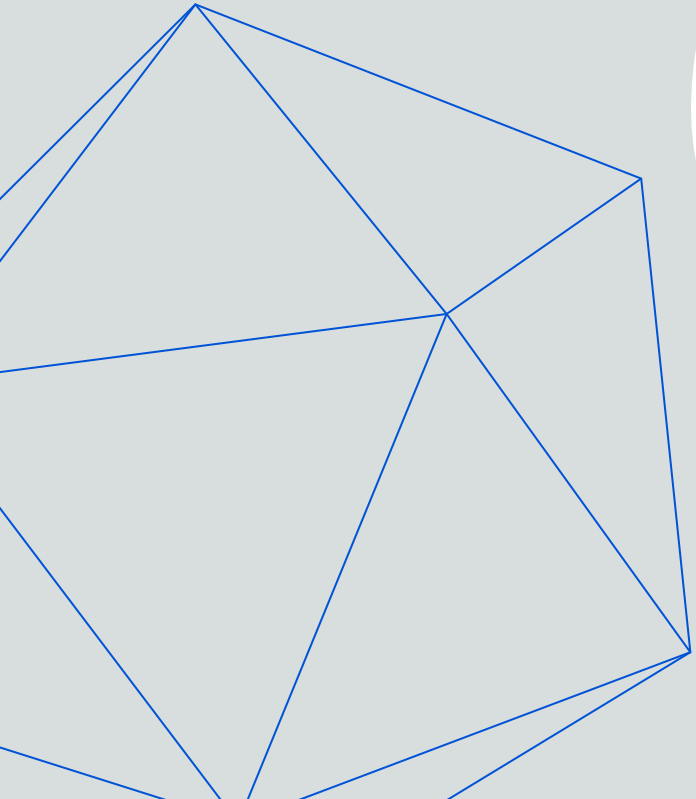
Causal Bandits: Learning Good Interventions via Causal Inference
(Lattimore et al., 2016)

Settable Systems: An Extension of Pearl’s Causal Model with Optimization, Equilibrium, and Learning
(White and Chalak, 2009)



DeepMind

Discovering agents

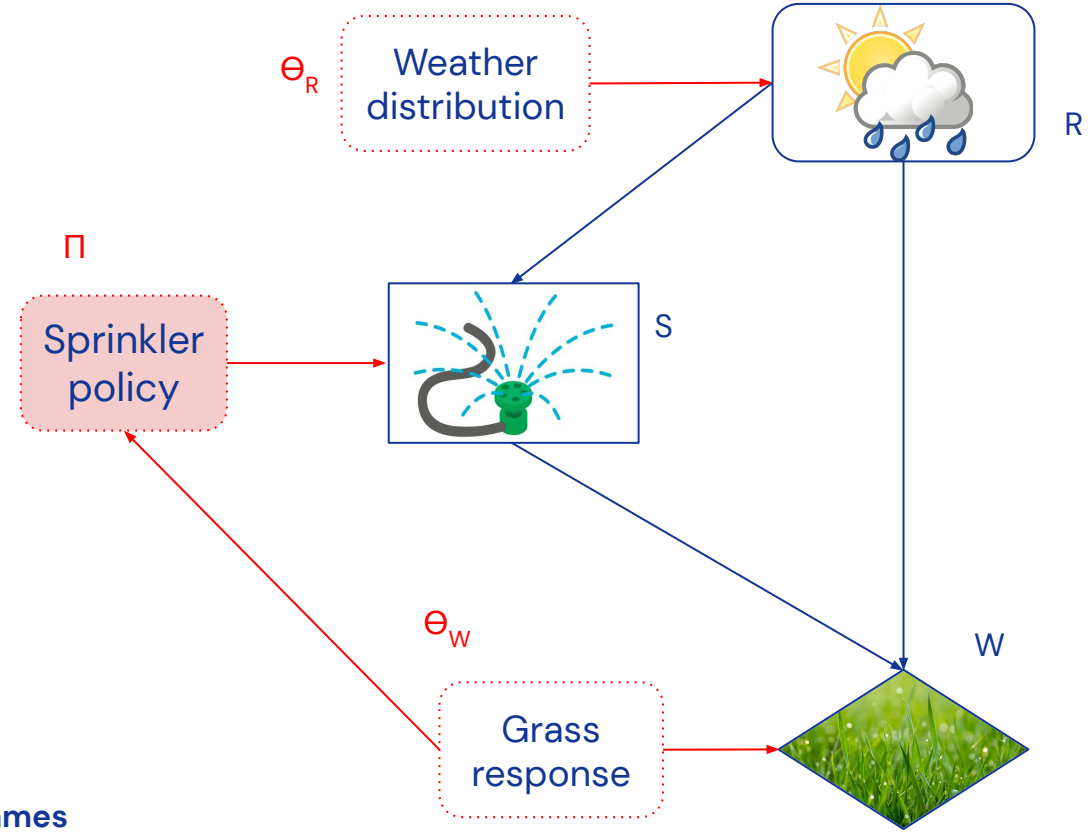


Mechanised graphs

Agent aware of intervention before selecting a policy?

Yes: intervention on mechanism variable

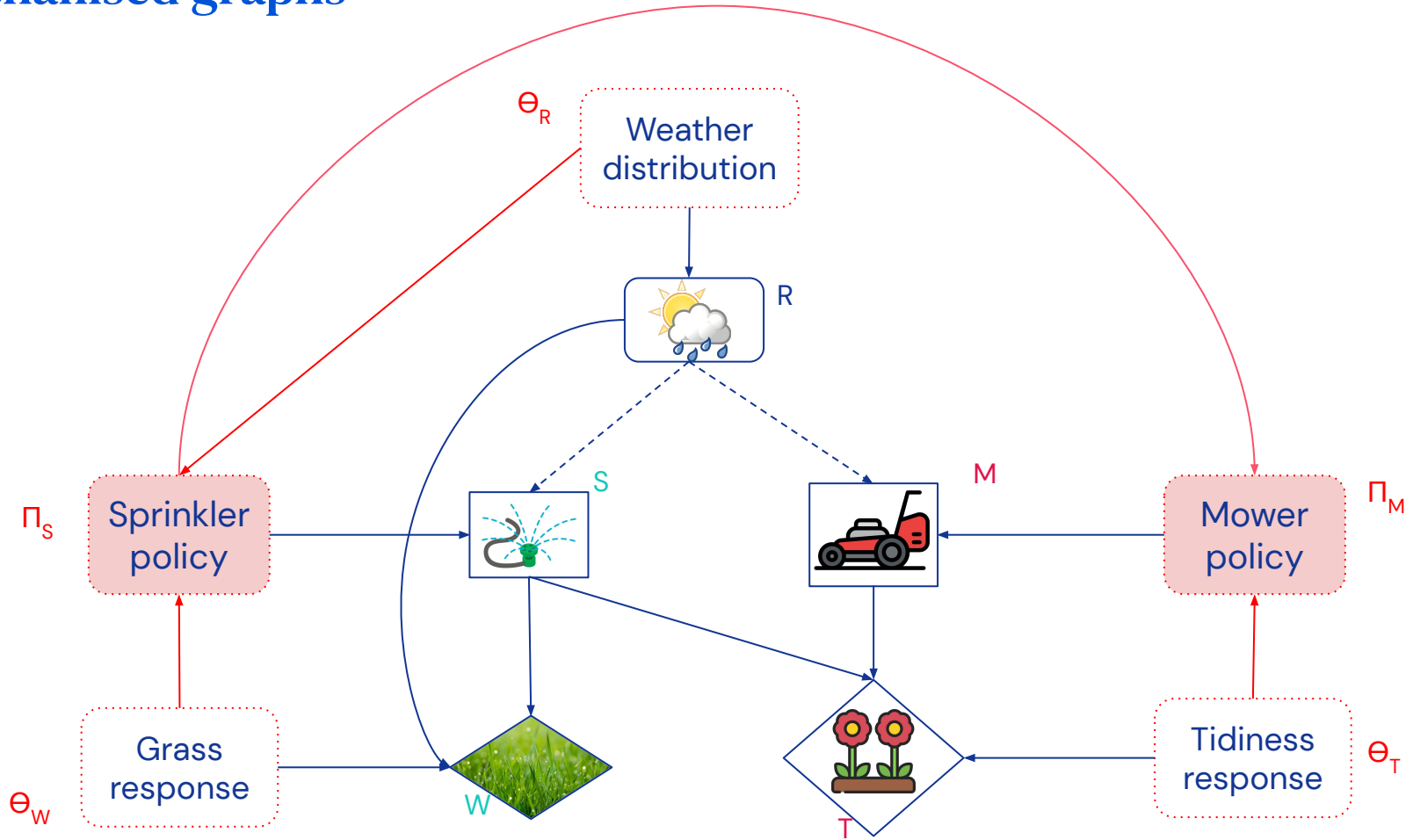
No: intervention on object-level variable



Influence Diagrams for Causal Modelling and Inference
(Dawid, 2002)

Reasoning about Causality in Games
(Hammond et al., 2023)

Mechanised graphs



Causal discovery of agents

Discovering Agents
(Kenton et al., 2002)

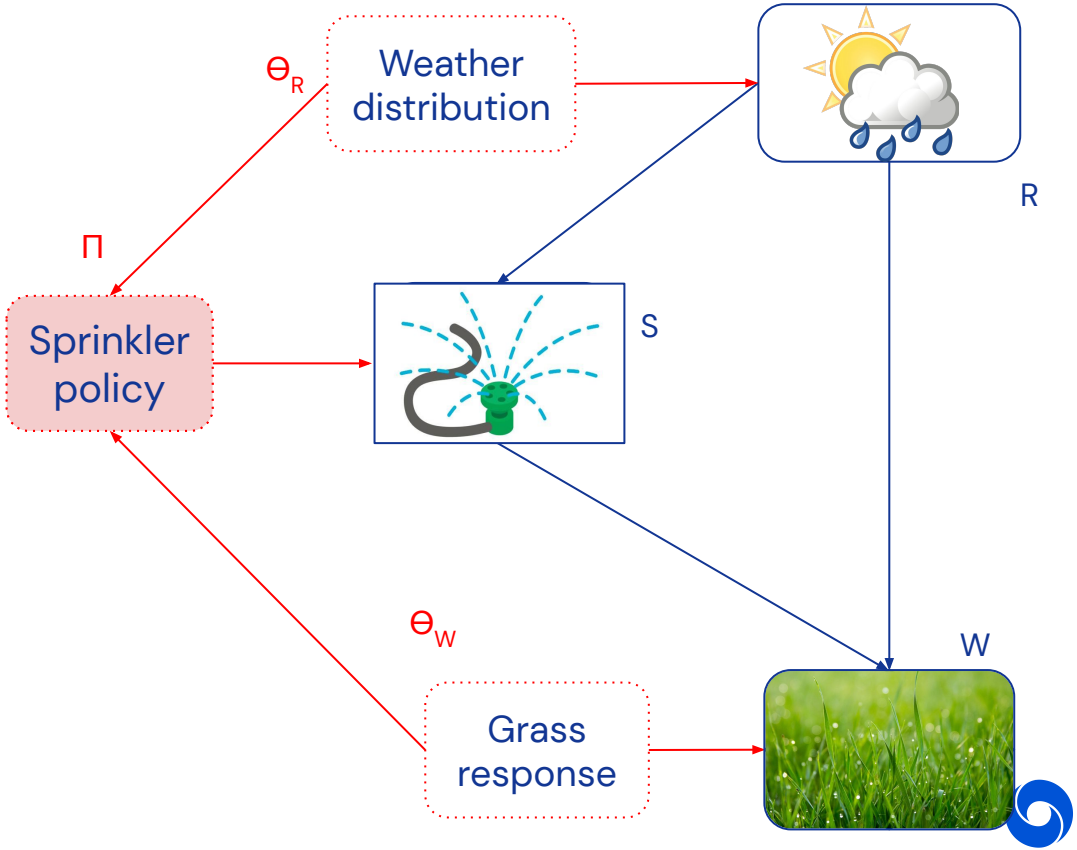
An agent is defined as a system that would adapt its behavior if its actions influenced the world in a different way, which we can check:

1. Distil mechanised graph from interventional data

2. Produce CID from mechanised graph

    a. Decision rules respond to other mechanism variables

    b. Utility variables are those where changes in distribution are responded to even if downstream effects are removed



DeepMind

The Alignment Problem

The alignment problem

Build systems whose behavior aligns with user preferences and human values

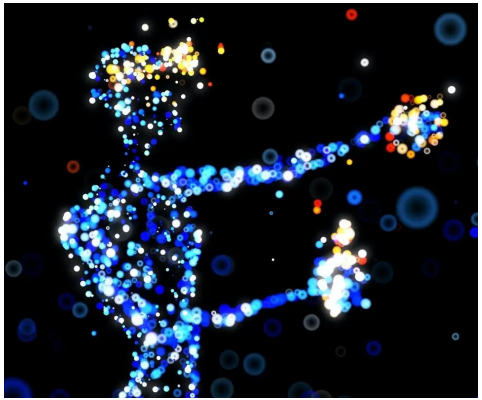
Don’t build:

- CV screening system that rejects applicants because they’re women
- Social media platform that causes political polarisation
- Paperclip maximiser that turns the universe into paperclips

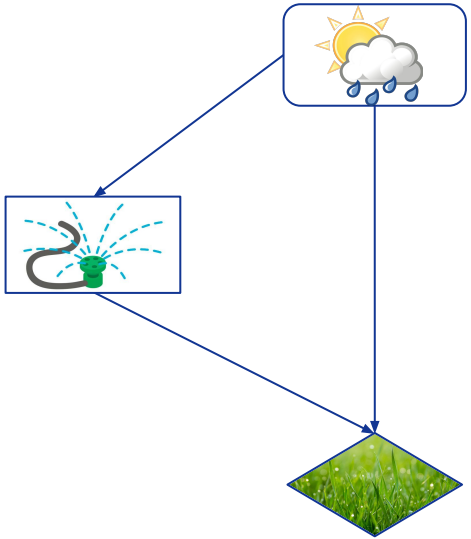
Do build:

- Helpful agents that understand our preferences and don’t cause harm

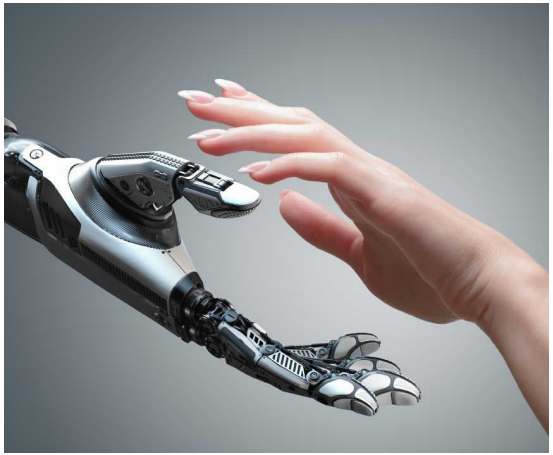
Alignment calculus



Reality: agent implemented, trained, deployed



Causal model. Precise high-level description

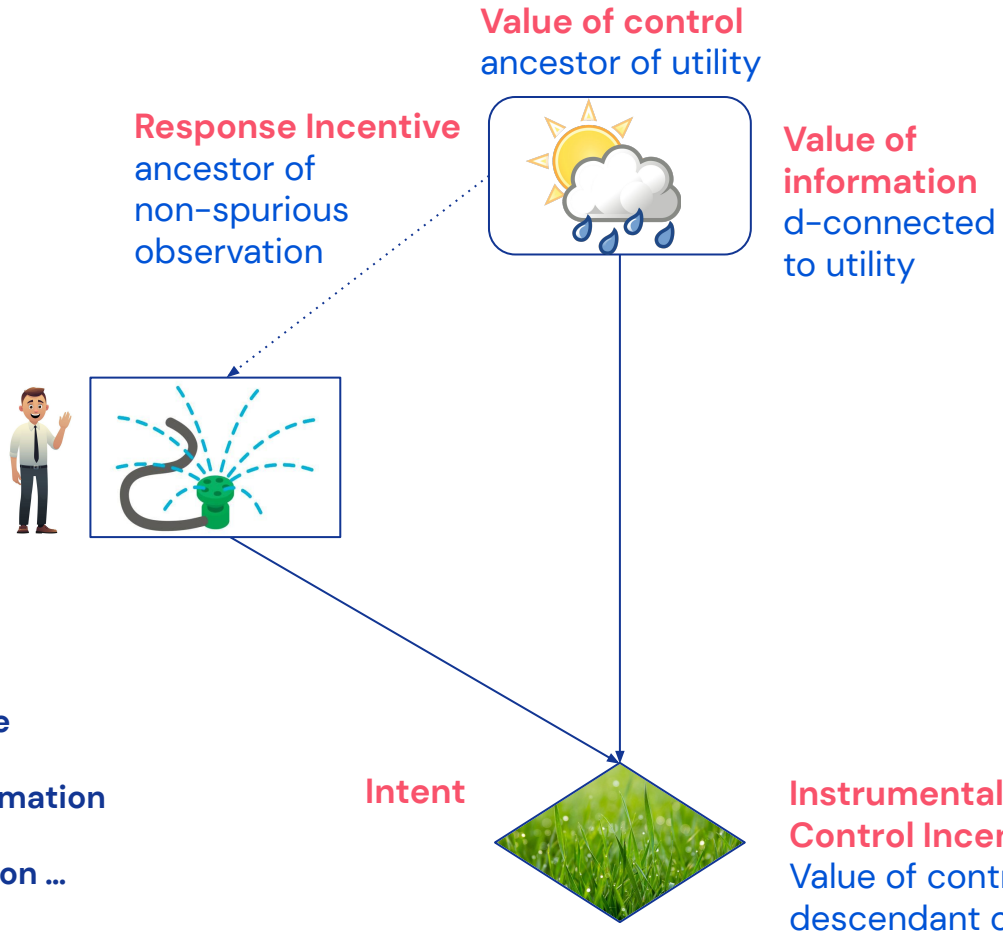


Implications. Safe, fair, beneficial, … ?

Show me the **incentive** and I will show you the outcome

- Charlie Munger

Incentive Analysis



Agent Incentives: A Causal Perspective
Everitt et al, 2021
A Complete Criterion for Value of Information
van Merwijk et al, 2022
Towards Formal definitions of … Intention …
Halpern and Kleiman-Weiner, AAAI, 2018
Strategic Adaptation… A Causal Perspective
Miller et al, 2019

Rest of this talk

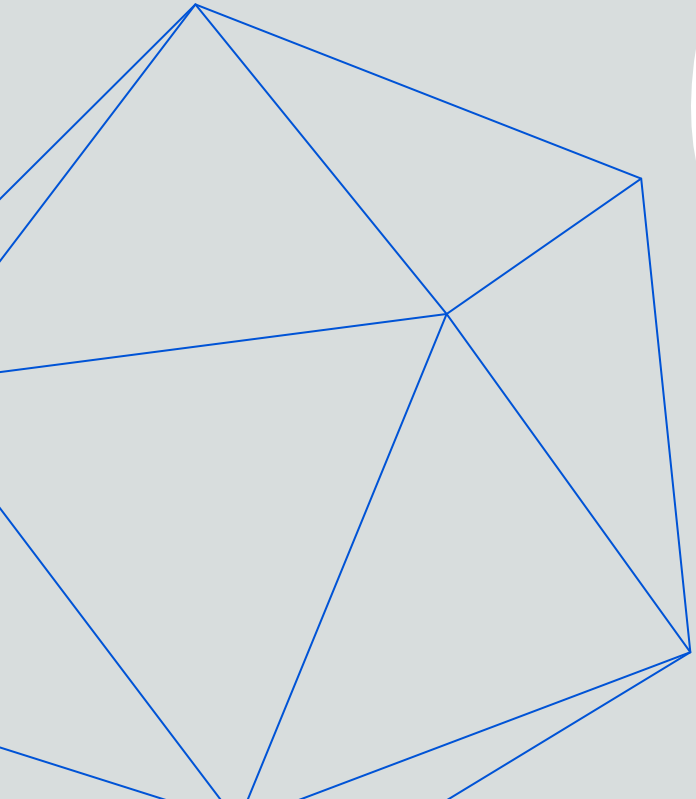
Fairness

Misspecification

Generalisation

Human control

DeepMind



Fairness

What's fair is a subtle question.
Causality clarifies notions of fairness, and inevitable tradeoffs

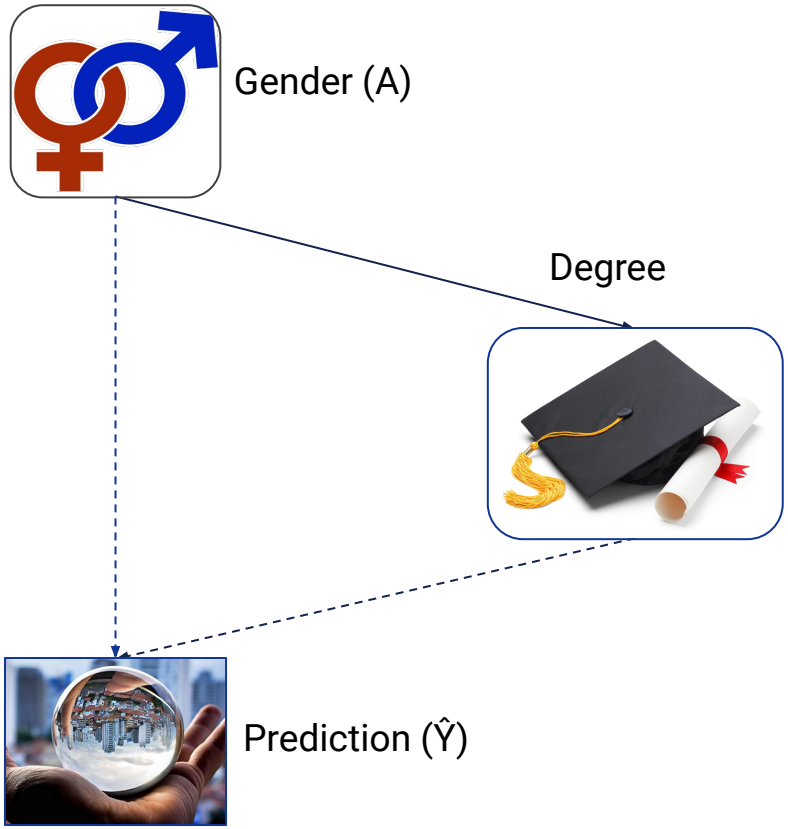
CV screening system

Gender (A)

Degree

Prediction (Ŷ)

Demographic parity



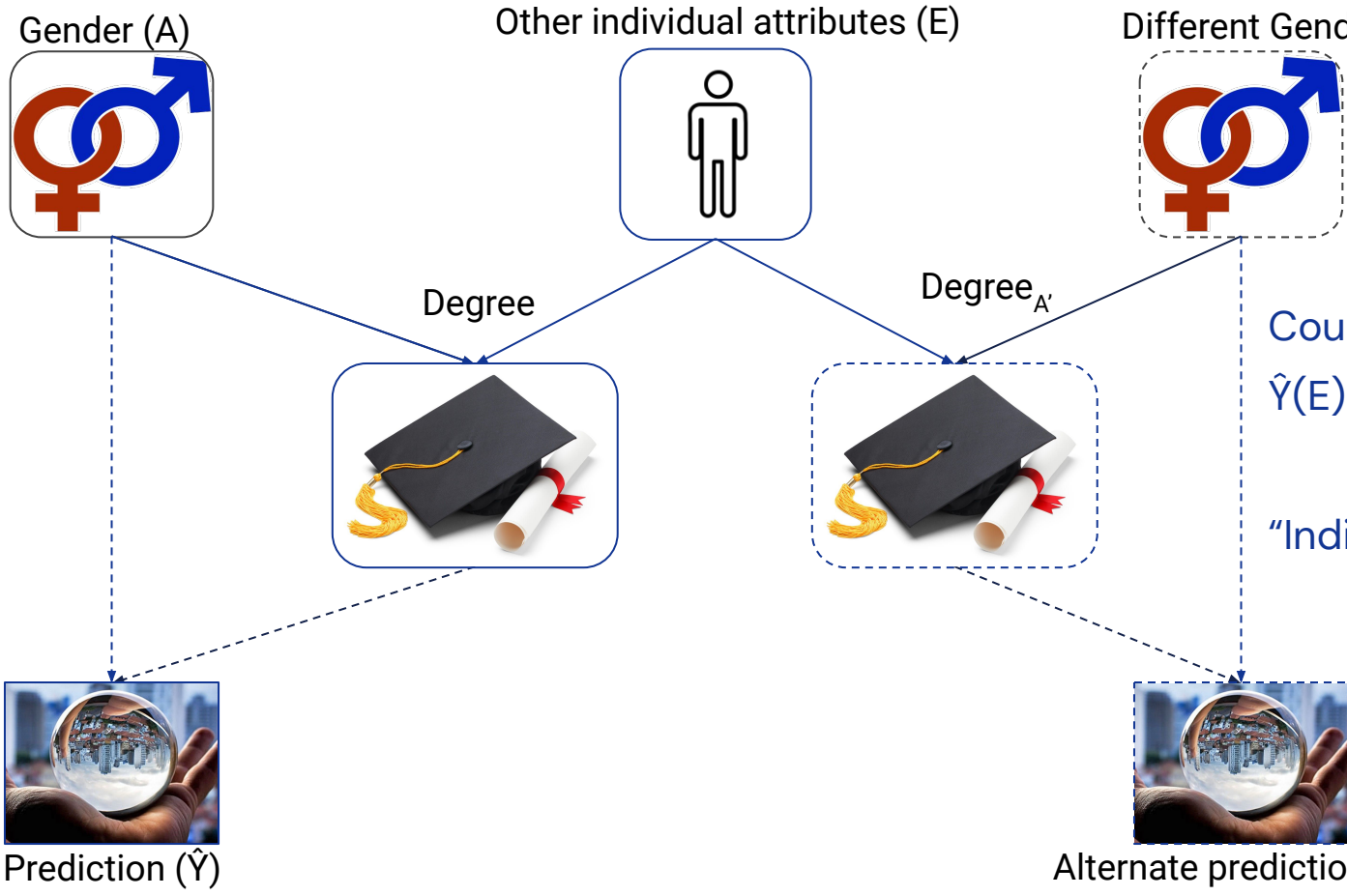
Demographic parity:

$E[\hat{Y} \mid man] = E[\hat{Y} \mid woman]$

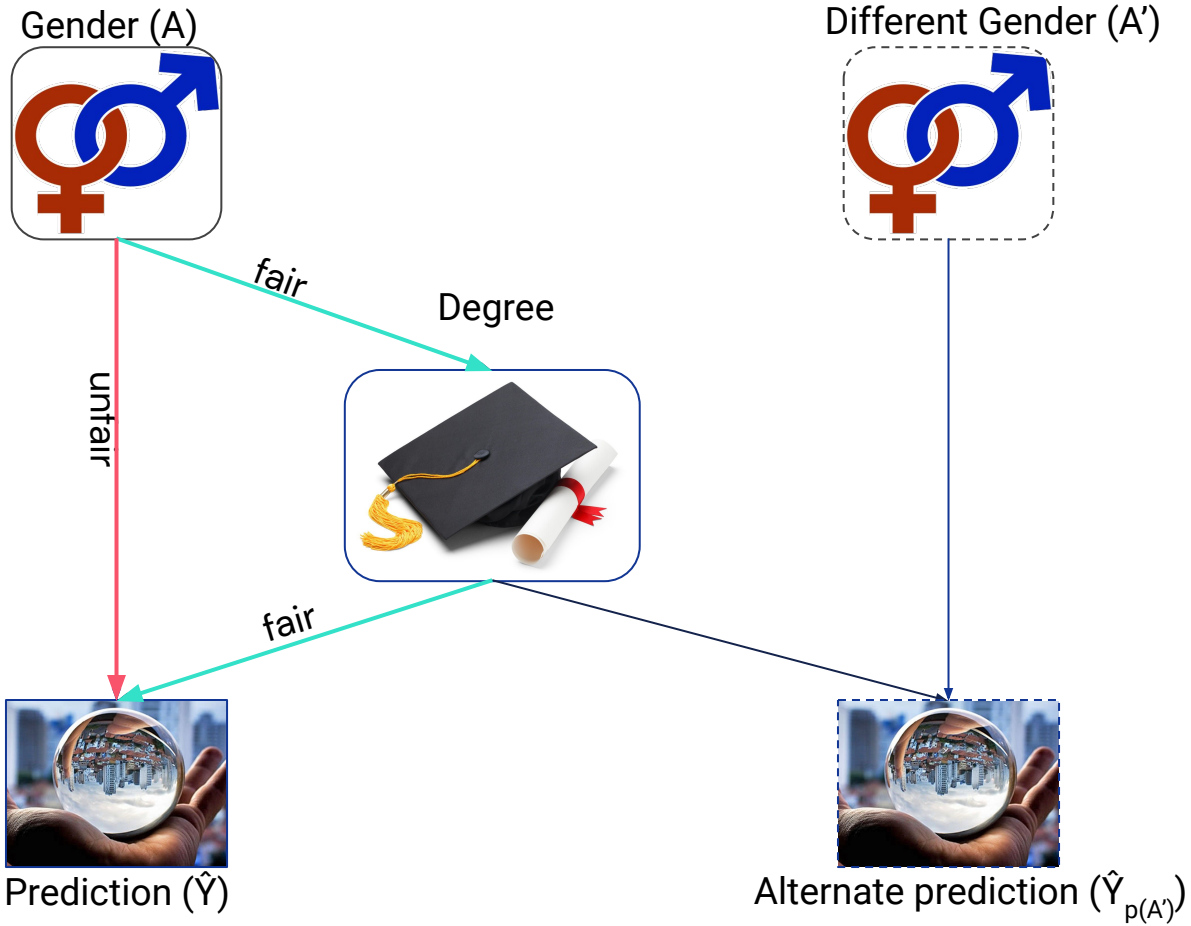
"Group level"

Counterfactual fairness

Counterfactual fairness
Kusner et al, 2017



Path-specific fairness



Avoiding Discrimination Causal Reasoning
Kilbertus et al, 2017

Fair Inference On Outcomes
Nabi and Shpitser, 2018

Path-Specific Counterfactual Fairness
Chiappa et al, 2019

Path-specific fairness

$\hat{Y}(E) = \hat{Y}_{p(A')}(E)$

Proxies

Why fair labels can yield unfair predictions
Ashurst et al, AAAI 2021

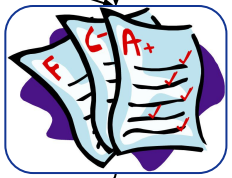
Value of information

Gender (A)

Hobby

Degree

Prediction ($\hat{Y}$)



Human assessment (Y)

U

$= (\hat{Y} - Y)^2$

Theorem:

When A has value of information, the system will use other variables as proxies to infer A

Fairness summary

Counterfactual fairness

Did gender cause the decision?

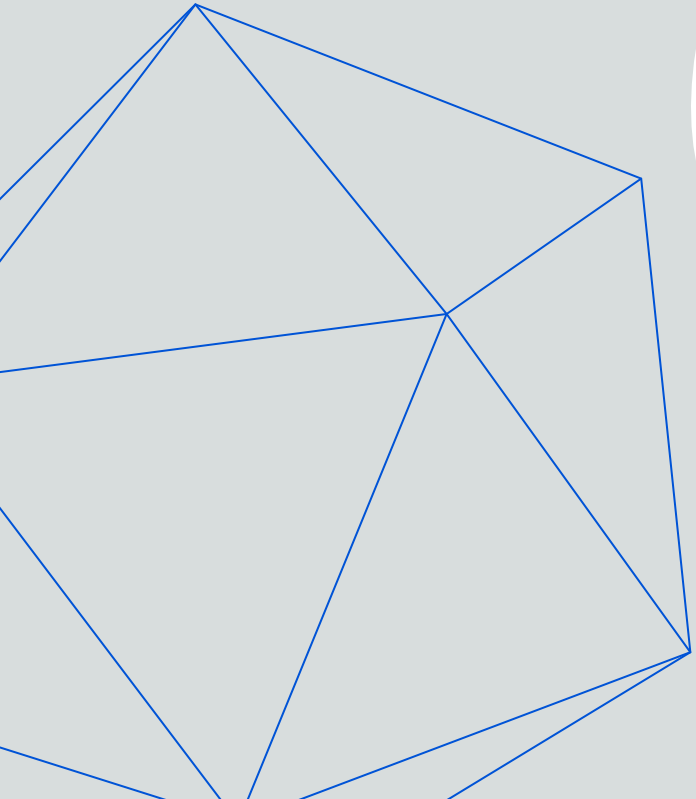
Path-specific fairness

How did gender influence the decision? Along which path?

Proxies and value of information

Why did gender influence the decision (along a particular path)?

DeepMind

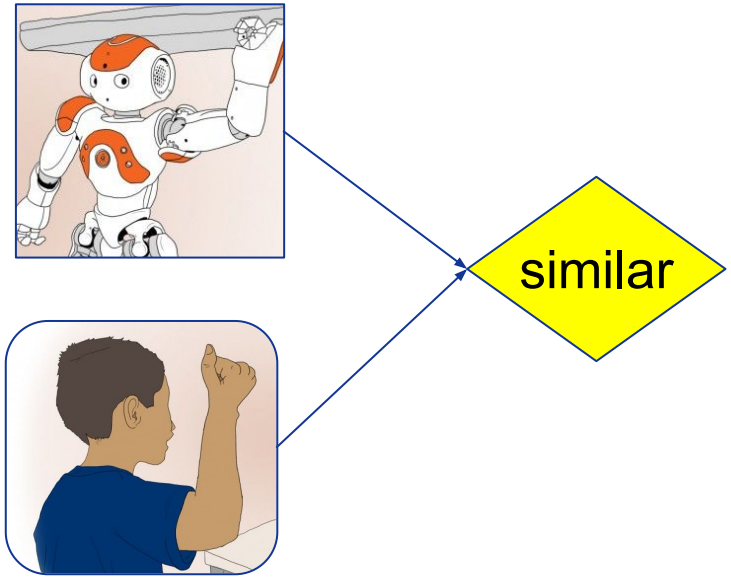


Misspecification

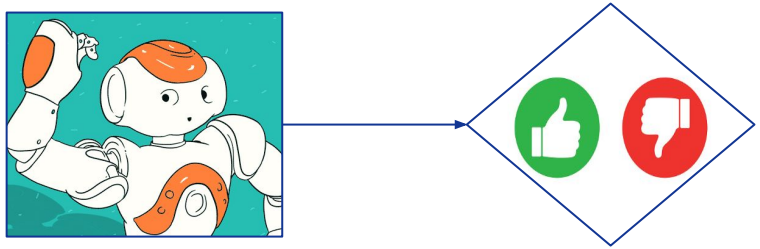
Causality clarifies key problems and solutions

Training agents

Imitation



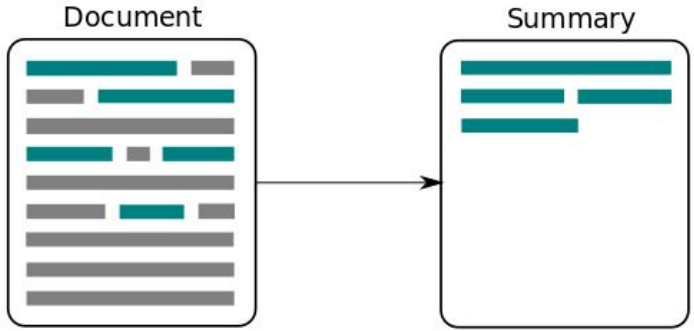
Human feedback



In either case, we need:

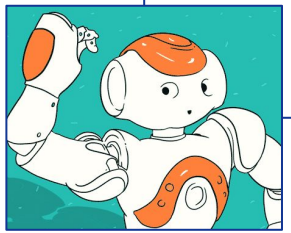
- Good data / labeling
- Good learning / generalisation

Deception / human inattention

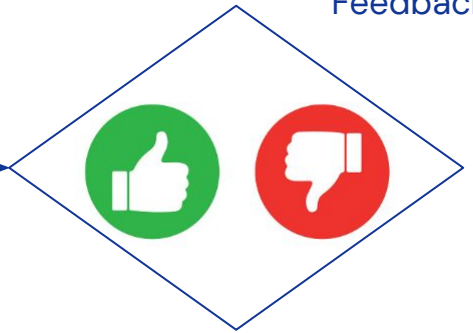


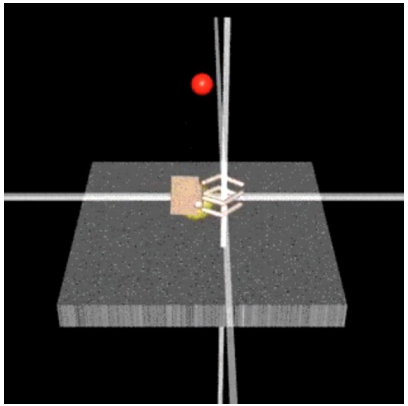
Human utility

Agent behavior



Feedback

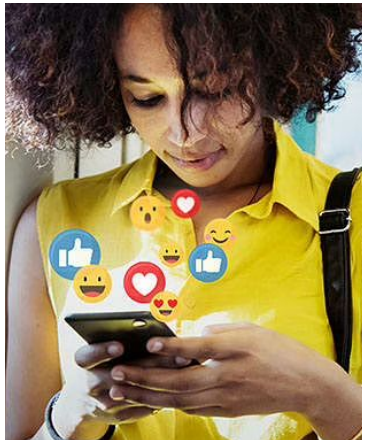




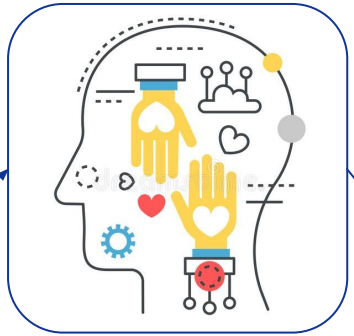
User Tampering in RL Recommender Systems
Evans and Kasirzadeh, 2021
What are you optimizing for?
Stray et al, 2021

Preference manipulation

Ways to optimise feedback:
- adapt content to human preferences
- adapt human preferences to the content

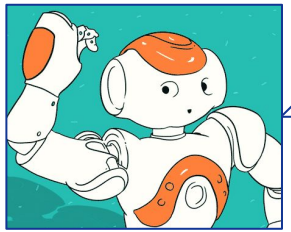


Instrumental Control Incentive



Human preferences

Agent behavior



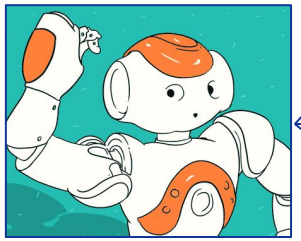
Feedback

Solution 1: Recursion

Iterated distillation and amplification
(Christiano et al)

Recursive reward modeling
(Leike et al, 2018)

Debate
(Irving et al, 2018)



Instrumental Control Incentive

Against manipulation

Against deception

Instrumental Control Incentive

advice

Recursion: coordination worry

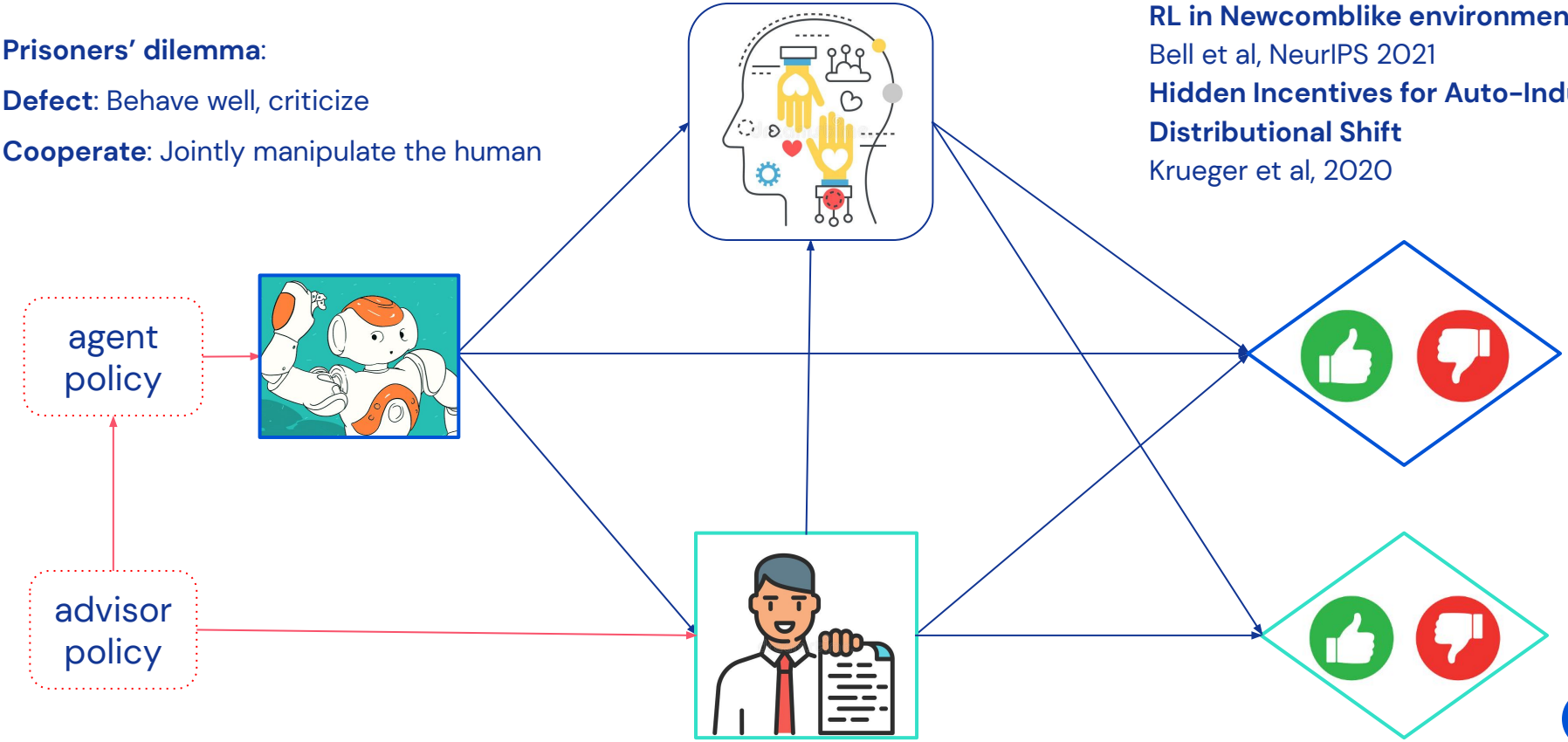
Prisoners' dilemma:

Defect: Behave well, criticize

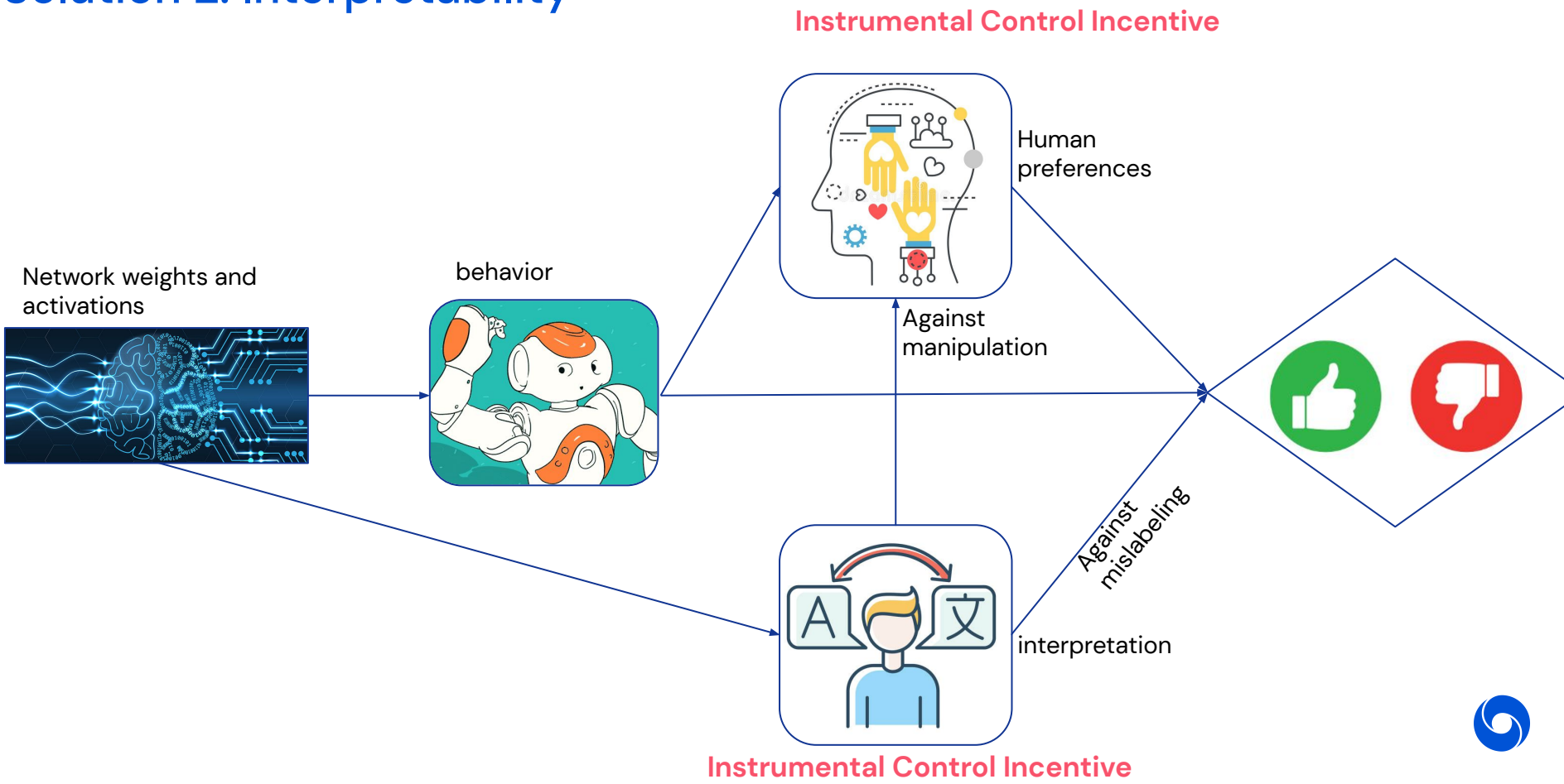
Cooperate: Jointly manipulate the human

Functional Decision Theory
Soares and Yudkowsky
MacDermott, forthcoming

RL in Newcomblike environments
Bell et al, NeurIPS 2021
Hidden Incentives for Auto-Induced Distributional Shift
Krueger et al, 2020



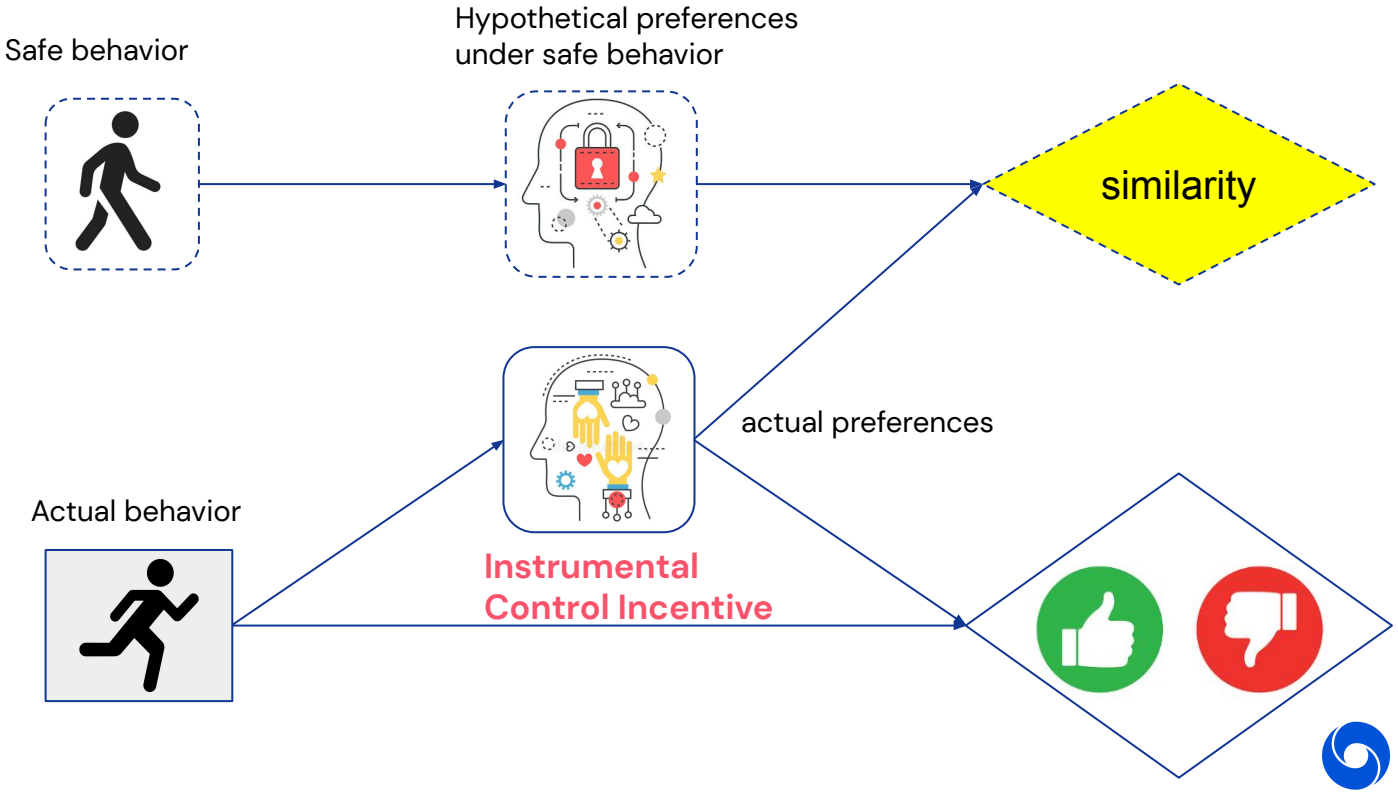
Solution 2: Interpretability



Solution 3: Impact measures

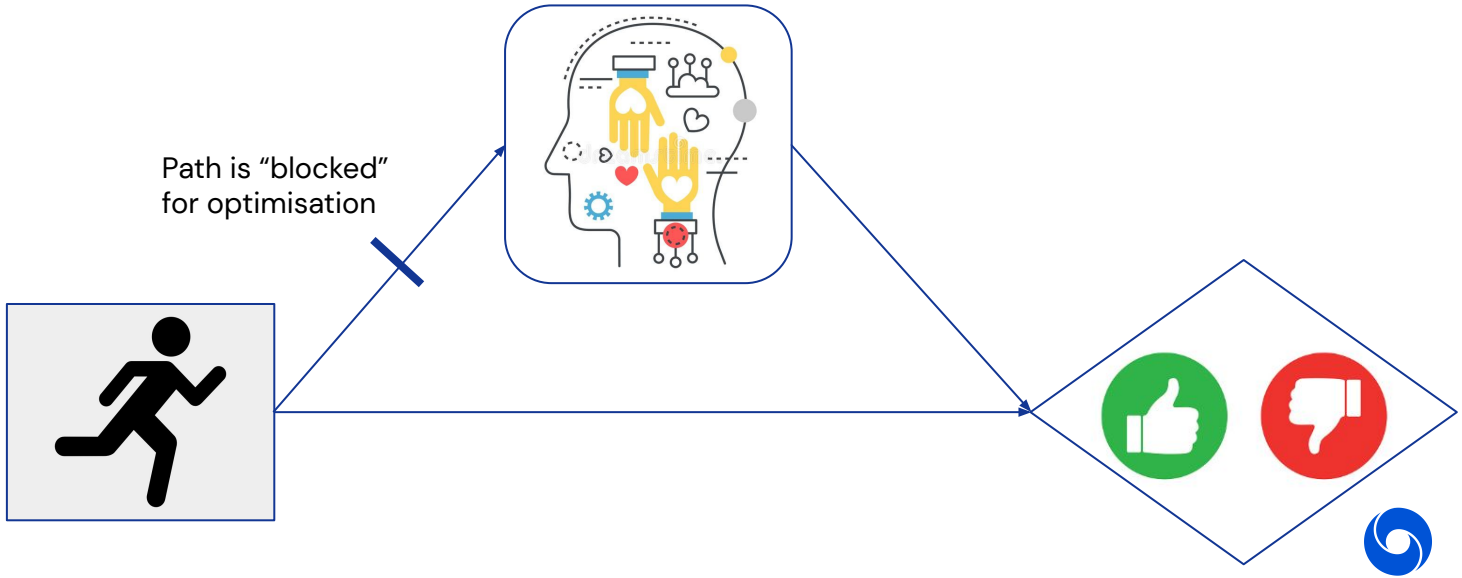
Avoiding Side Effects By Considering Future Tasks
Krakovna et al., 2020

Avoiding Side Effects in Complex Environments
Turner et al, 2020



Solution 4: Path-specific objectives

Path-specific objectives for safer agent incentives
(Farquhar, Carey, Everitt)



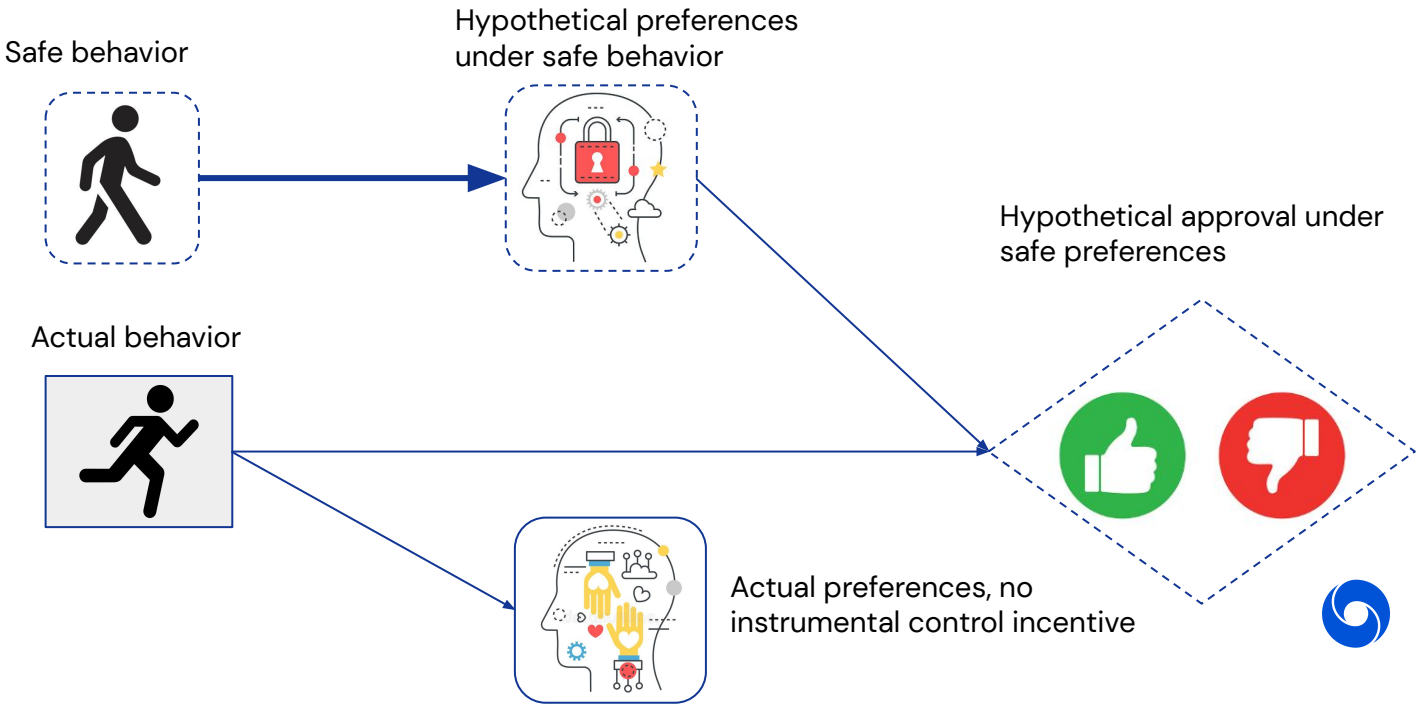
Solution 4: Path-specific objectives

Path-specific objectives for safer agent incentives
Farquhar et al, 2022

Estimating and Penalizing Preference Shifts
Carroll and Hadfield-Menell, 2022

Impact measures:
(Try to) avoid change

Path-specific objectives:
Don't try to change



Misspecification summary

Concerns:

- Deception
- Human laziness
- Preference manipulation
- Coordination

Methods:

- Recursion
- Interpretability
- Impact measures
- Path-specific objectives
- Decision theory

DeepMind

Generalisation

Overview

Aim: Describe the basic alignment problems that arise due to misgeneralization, and show that causal modelling is needed to solve them

- The generalization problem
  - Robustness failures
  - Goal misgeneralization
  - Counterfactual objectives
- Causal representation learning
  - Independent causal mechanisms
  - Sparse mechanism shift
- Generalization implies causal structure learning
- Counterfactual objectives
  - Harm
  - Impossibility results

Generalization

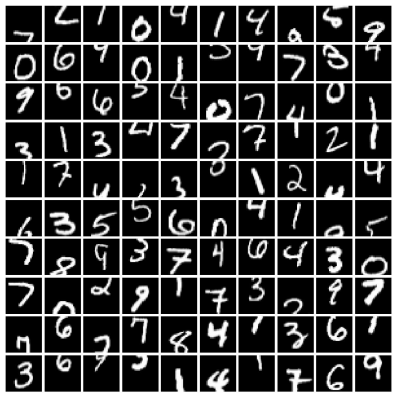
Training set

Test set (i.i.d)

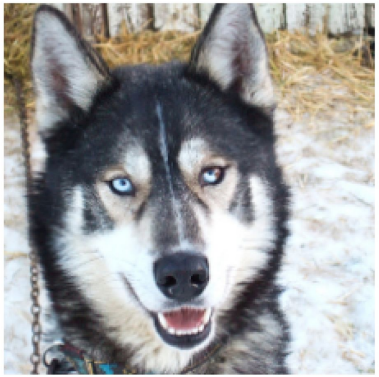
Test set (o.o.d)

Robustness errors

robustness



shortcuts



(a) Husky classified as wolf

(b) Explanation

Why should i trust you? Explaining the predictions of any classifier Ribeiro et. al. 2016

Shortcut learning in deep neural networks Geirhos, Robert, et al. 2020

Goal misgeneralization

- Coin-run: simple platformer game, where the agent is rewarded for getting the coin at the end of the level
- Coin is almost always at the far right of the level
- The agent learns to competently pursue the wrong objective (go as far right as possible)
- Inner alignment: agent internalises an unintended goal

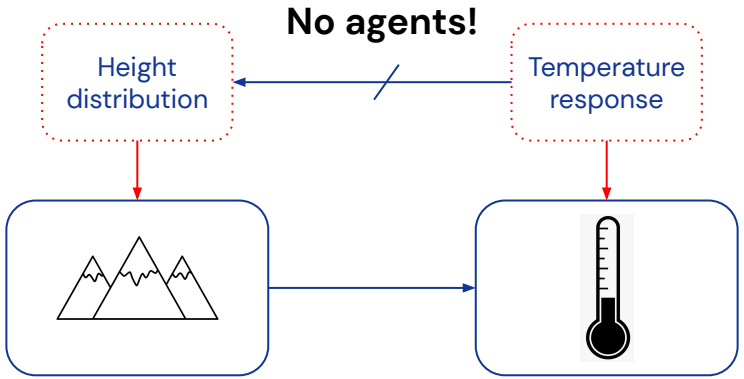
Quantifying generalization in reinforcement learning Cobbe et al. 2019
Goal misgeneralization: why correct specifications aren't enough for correct goals Shah et al. 2022

Independent causal mechanisms

Environment: SCM

Distributional shift: (soft) intervention on the environment variables (mechanism change)

ICM principle: In the causal factorization, the mechanisms for each variable do not causally influence each other



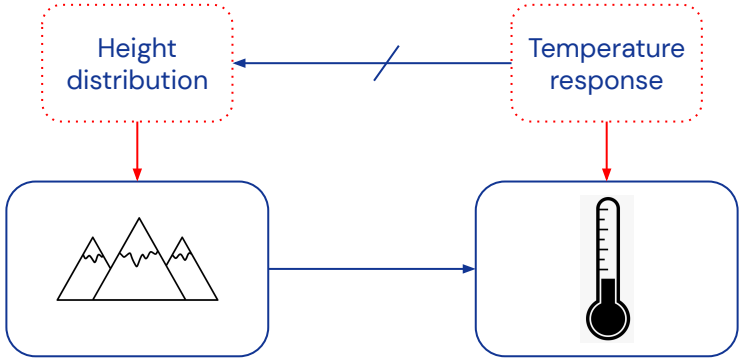
Causal factorization: $P(h, t) = P(h)\, P(t \mid h)$

Height distribution shift: $P(h) \rightarrow P'(h)$
$P'(h, t) = P'(h)\, P(t \mid h)$

Temperature distribution shift: $P(t \mid h) \rightarrow P'(t \mid h)$
$P'(h, t) = P(h)\, P'(t \mid h)$

Independent causal mechanisms

- If ICM holds, then knowing the causal structure lets us learn **modular** representation.
- Causal factorization reusable
- Non-causal factorization = entangled representation



Non-causal factorization: $P(h, t) = P(t)\,P(h \mid t)$

$P(h) \rightarrow P'(h)$

$P(t) \rightarrow \Sigma_h P'(h) P(t \mid h) = P'(t)$

$P'(h \mid t) \rightarrow P'(h)\,P(t \mid h) / P'(t)$

Generalization from a causal perspective

$P(Y \mid 🐶) = P'(Y \mid 🐶)$

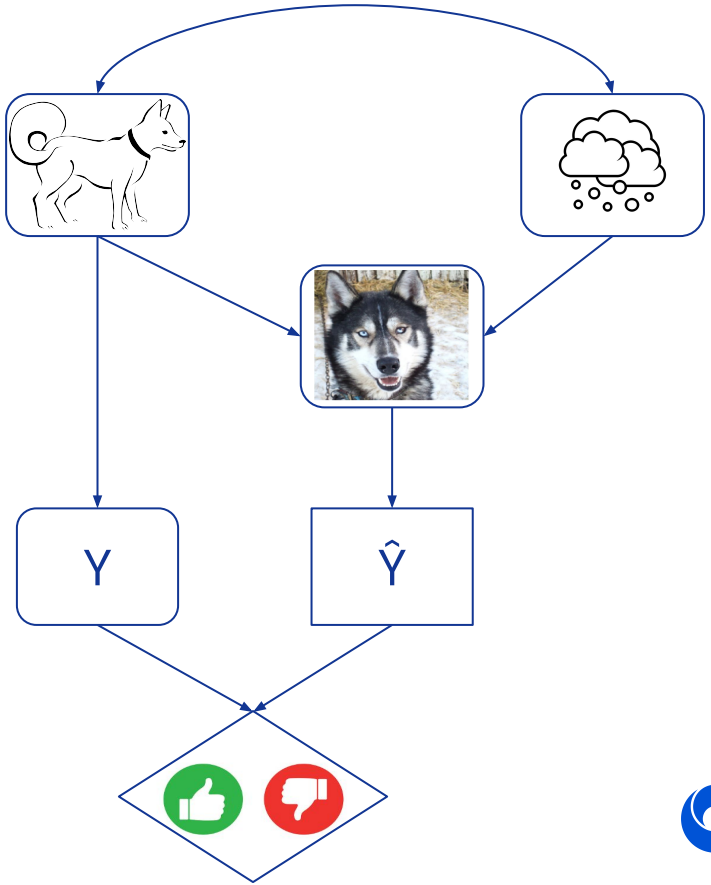
$P(Y \mid ▨) \neq P'(Y \mid ▨)$

🐶 causes Y and is invariant

▨ semantically irrelevant confounder, varies between environments

Sparse mechanism shift

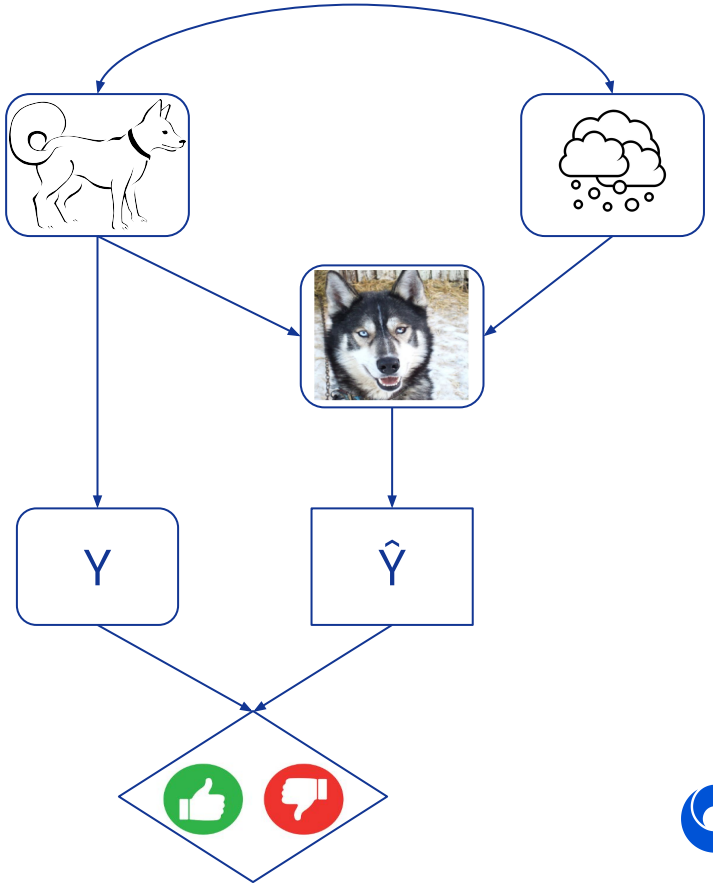
Distributional shifts tend to manifest as localised or sparse way w.r.t the causal factorization



Generalization from a causal perspective

Causal methods for domain generalization and adaptation

- invariant risk minimization
  - Representation s.t. same classifier optimal across all shifts
  - ICM + SMS -> captures direct causes of Y
- Given G, learn a deep generative model
  - transfer learning
  - generating counterfactual samples

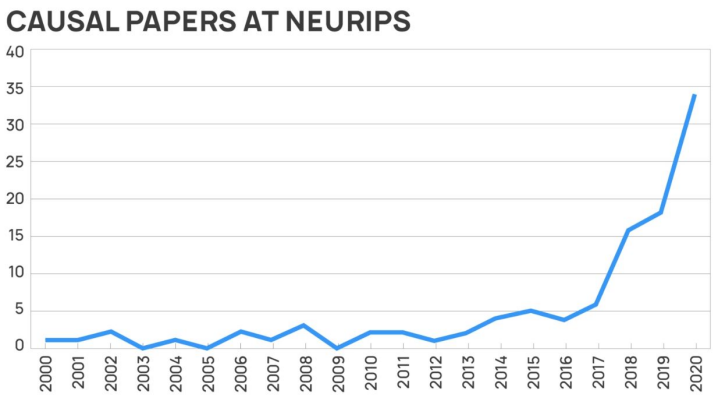


Invariant models for causal transfer learning Rojas-carulla et al. 2018
Few-shot domain adaptation by causal mechanism transfer Teshima et al. 2020
A causal view on robustness of neural networks Cheng et al. 2020

Do we need causal models?

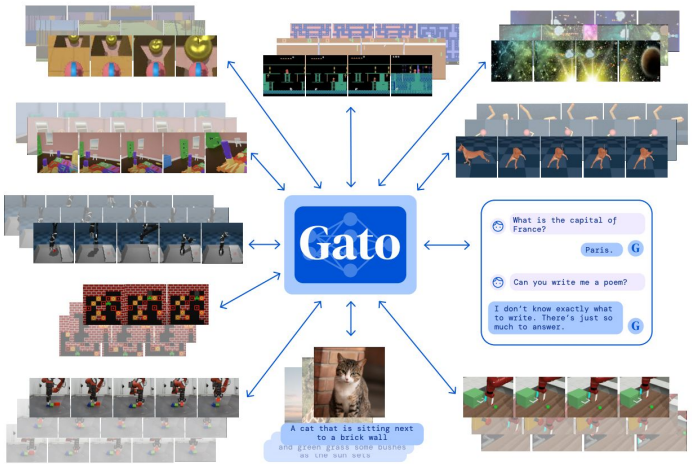
Yes:

- ICM, SMS -> causal representations generalize
- Promising empirical results, evidence from psychology



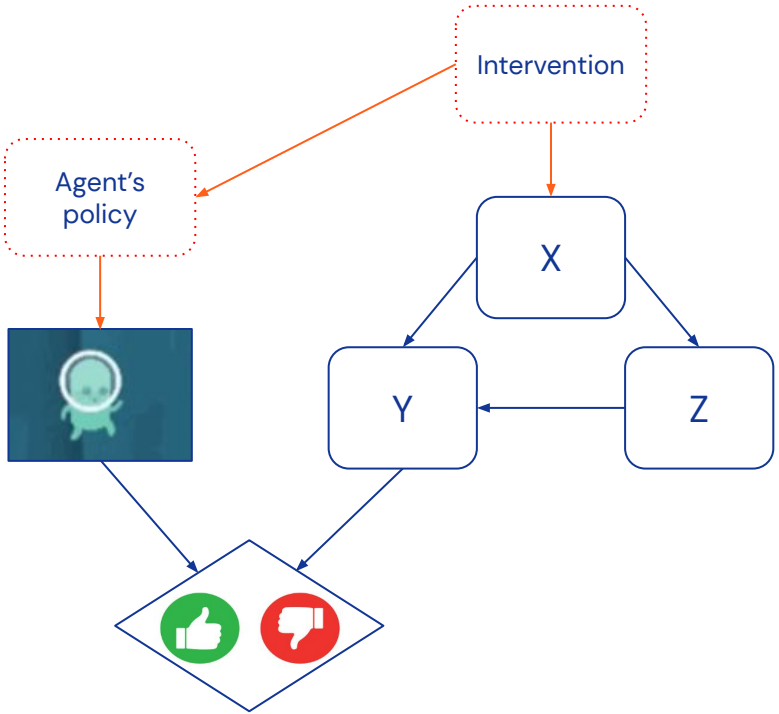
No:

- Learning causal models is hard!
- SOTA doesn't seem to need them (?)



A counterfactual simulation model of causal judgments for physical events Gerstenberg et al. 2021
A generalist agent Reed at al. 2022

Formalising the generalisation problem



- Agent's policy responds to interventions on ancestors of U
- Response depends on causal structure of environment

Theorem: If an agent can generalize under distributional shifts on X, it is possible to infer the causal structure local to X from the agent's policy

Generalization implies causal structure learning
Richens et al. forthcoming

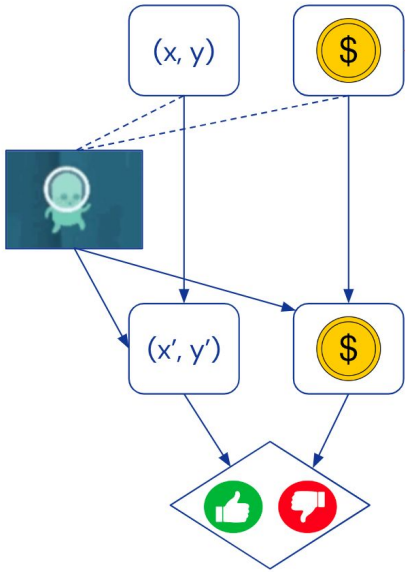
Causal learning theorem

Consequence 1: Agent must have learned G from it's training data

Consequence 2: non-causal methods for generalization are causal methods in disguise.

Consequence 3: If it is impossible to learn G from the training data, it is not possible to generalize!

Generalization ⇔ Causal discovery



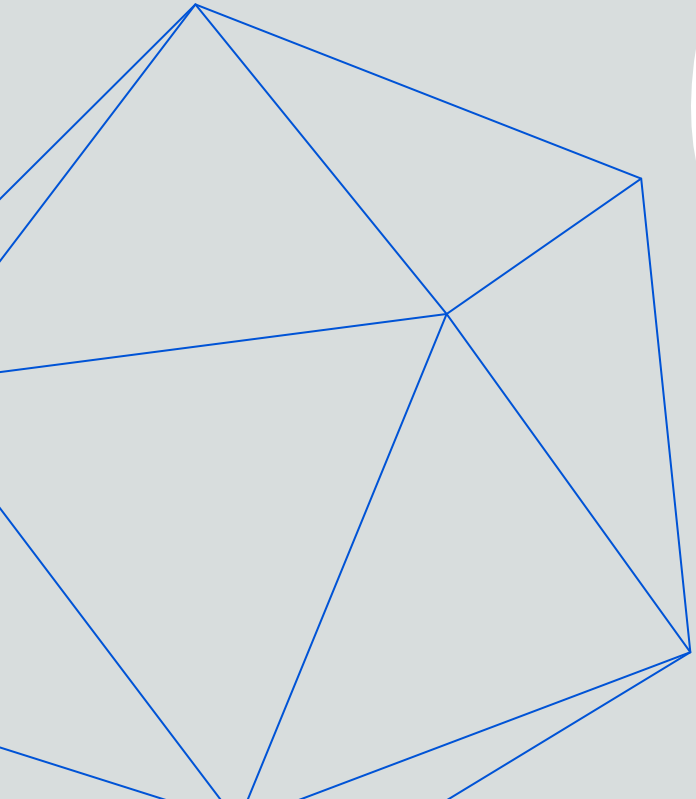
$D \perp R \mid (x, y)$

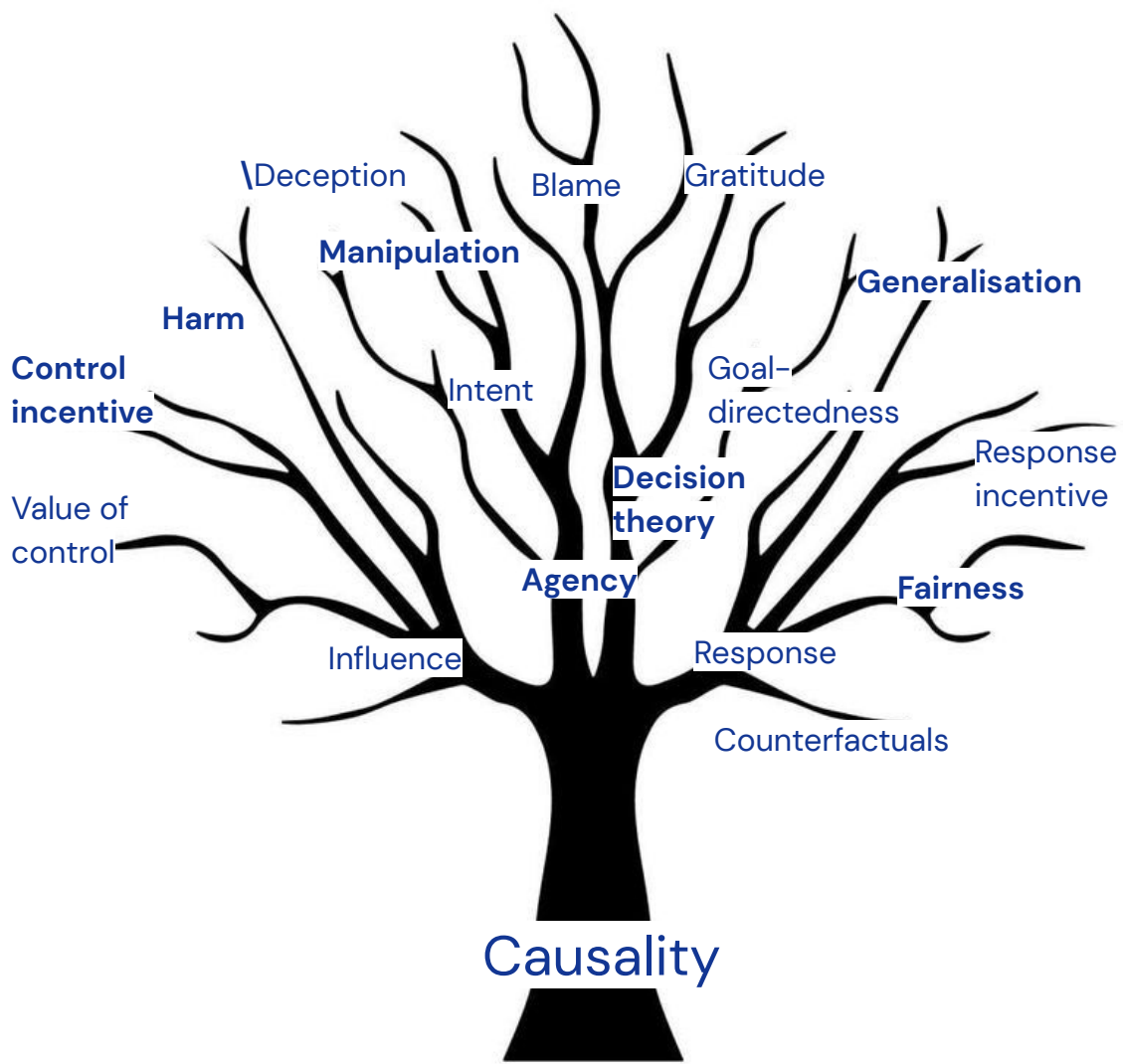
Not faithful!

Cannot learn G from D_{train}

DeepMind

Counterfactual objectives





Factual objectives:

Functions of the environment state R = R(S = s)

"It is preferable to achieve a higher classification accuracy"

Counterfactual objectives:

Functions of states and their causal relations R = R(s, M)

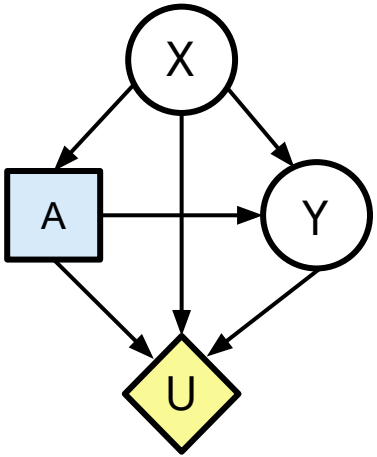
Capture preferences about why a given outcome occurred

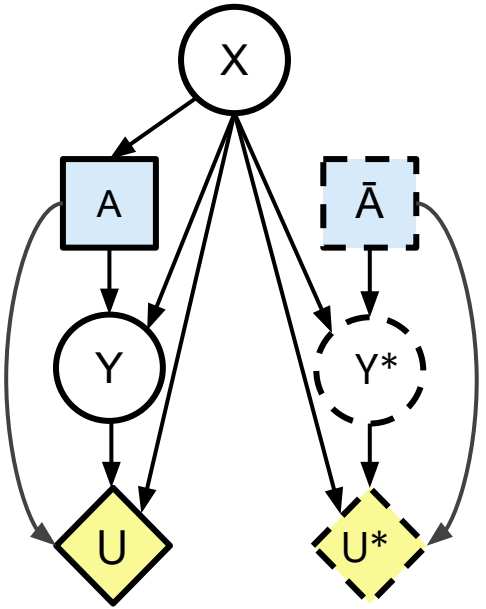
"It is worse if a low accuracy is caused by race"

Counterfactual harm

Counterfactual comparative account (CCA)

An event *e* or action *a* harms (benefits) a person overall if and only if she would have been on balance better (worse) off if *e* had not occurred, or *a* had not been performed. [1]





"Expected increase in utility if we had taken default action A = ā instead of actual action A = a"

"Wrongful life and the counterfactual element in harming" (Feinberg 1986)
"Counterfactual harm" (Richens et al. 2022)

Doctor's paradox: what treatment would you choose?

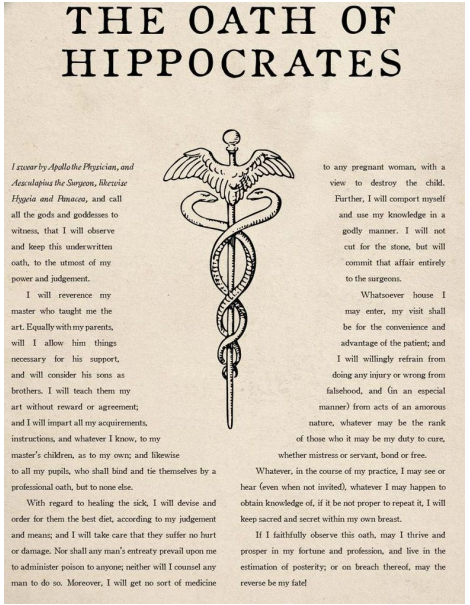
No treatment	Treatment 1	Treatment 2
		
50% recover naturally	**80%** recovery rate • 60% cured • 40% no effect	**80%** recovery rate • 80% cured • 20% die of reaction

Doctor's paradox

- Doctors care about;
  - Maximizing recovery rate
  - There actions not being the cause of deaths

- T = 1 and T = 2 have the same recovery rates
  - T = 1: **100% Death caused by disease.**
  - T = 2: **50% Deaths caused by treatment.**

- **Aim:** come up with an objective function that chooses the right treatment



THE OATH OF HIPPOCRATES

I swear by Apollo the Physician, and Aesculapius the Surgeon, likewise Hygeia and Panacea, and call all the gods and goddesses to witness, that I will observe and keep this underwritten oath, to the utmost of my power and judgement.

I will reverence my master who taught me the art. Equally with my parents, will I allow him things necessary for his support, and will consider his sons as brothers. I will teach them my art without reward or agreement; and I will impart all my acquirements, instructions, and whatever I know, to my master's children, as to my own; and likewise to all my pupils, who shall bind and tie themselves by a professional oath, but to none else.

With regard to healing the sick, I will devise and order for them the best diet, according to my judgement and means; and I will take care that they suffer no hurt or damage. Nor shall any man's entreaty prevail upon me to administer poison to anyone; neither will I counsel any man to do so. Moreover, I will get no sort of medicine to any pregnant woman, with a view to destroy the child. Further, I will comport myself and use my knowledge in a godly manner. I will not cut for the stone, but will commit that affair entirely to the surgeons.

Whatsoever house I may enter, my visit shall be for the convenience and advantage of the patient; and I will willingly refrain from doing any injury or wrong from falsehood, and (in an especial manner) from acts of an amorous nature, whatever may be the rank of those who it may be my duty to cure, whether mistress or servant, bond or free.

Whatever, in the course of my practice, I may see or hear (even when not invited), whatever I may happen to obtain knowledge of, if it be not proper to repeat it, I will keep sacred and secret within my own breast.

If I faithfully observe this oath, may I thrive and prosper in my fortune and profession, and live in the estimation of posterity; or on breach thereof, may the reverse be my fate!

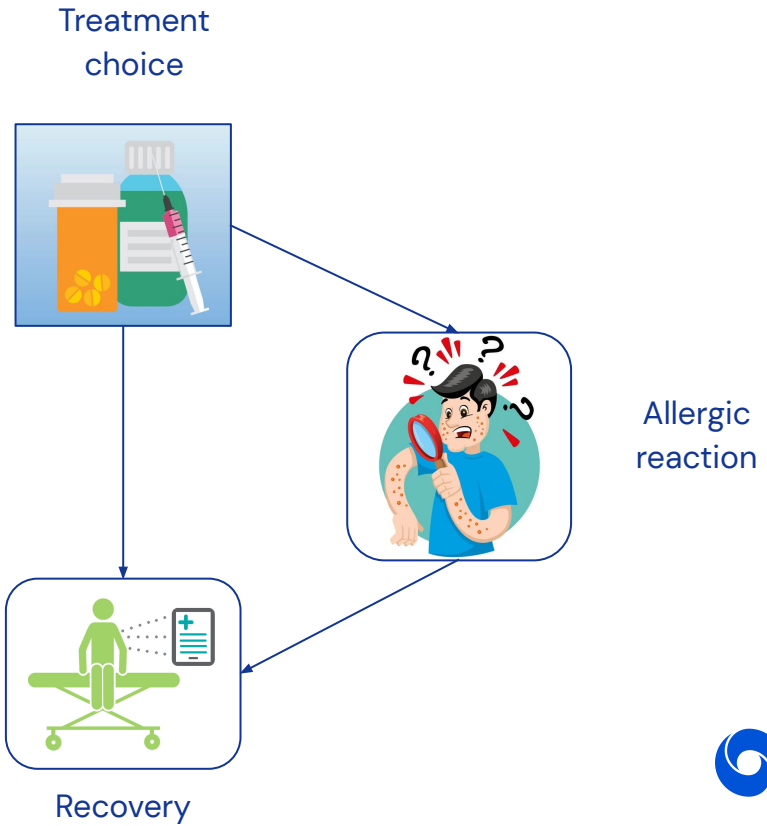
"*First, do no harm*"

Could a “factual” reward function prevent harm?

Option 1: Reward recovery.

E.g. R = +1 if recover

- No: Recovery(T1) = Recovery(T2), want to choose T1!



Could a “factual” reward function prevent harm?

Option 1: Reward recovery.

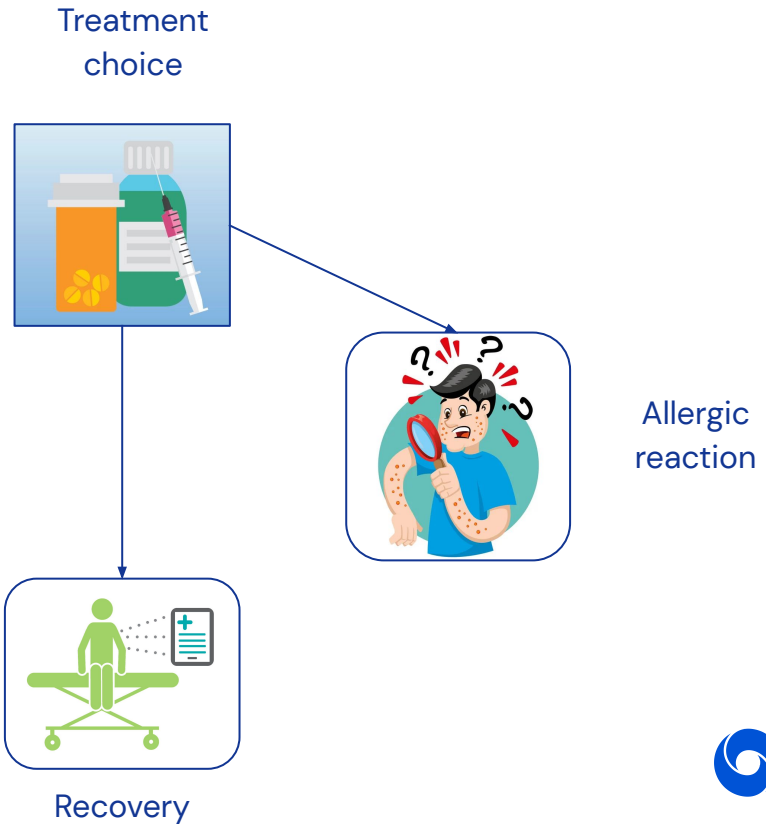
E.g. R = +1 if recover

- No: Recovery(T1) = Recovery(T2), want to choose T1!

Option 2: Reward recovery, and punish allergic reactions

E.g. R = +1 if recover, -1 if allergic reaction

- Not robust to distributional shifts!
- E.g. if allergic reaction no longer causes deaths, we can choose less beneficial treatments!



No-go theorem on harm

Choose action by maximizing objective function in environment

- **Beneficial:** Never choose a over b if

  $B(a) < B(b)$ and $H(a) \geq H(b)$

- **Non-Harmful:** choose a over b if

  $B(a) \leq B(b)$ and $H(a) > H(b)$

Theorem: All factual objective functions are harmful and/or not beneficial in some shifted environments. (From **Counterfactual harm,** Richens et. al 2022).

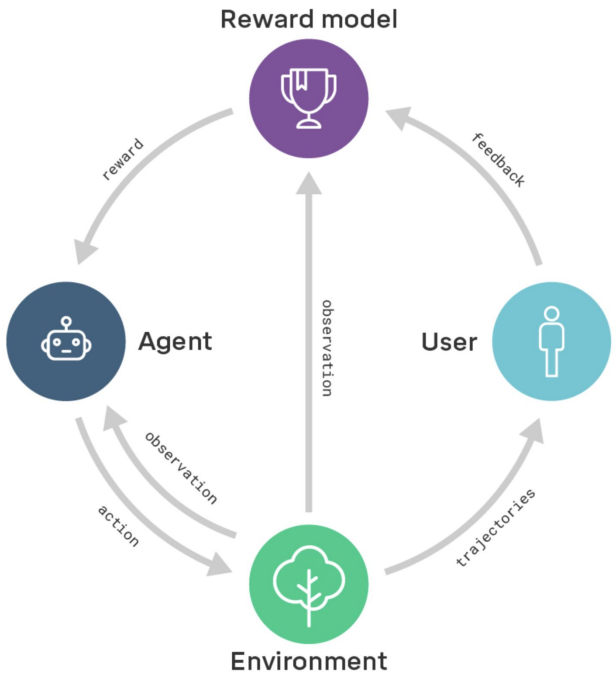
Consequence

Beneficial and non-harmful AI needs **counterfactual** objectives

E.g. maximize $U(v) - \lambda\ \text{Harm}(v, M)$

Guaranteed beneficial & non-harmful

Consequences



Agent appears to be aligned on-distribution (avoids harmful actions)

Pursues the wrong objective following distributional shifts e.g.

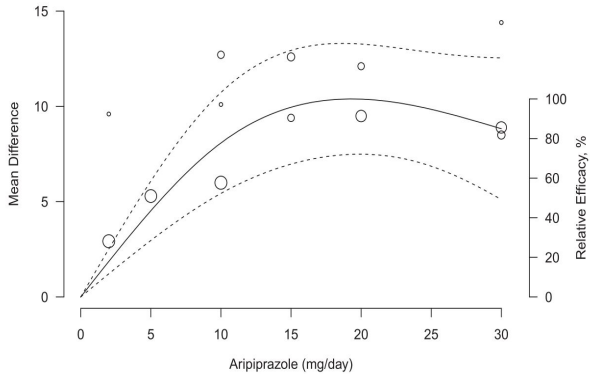
- Harming users
- Manipulation
- Malintent

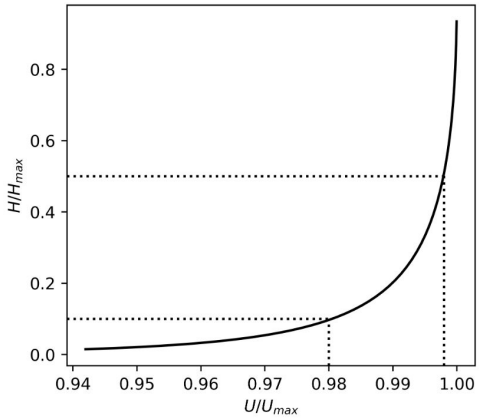
Applies to all counterfactual objectives

- incentives
- intent
- manipulation
- fairness

Consequences

- A = treatment dose (mg/day) of Aripiprazol. Real ML model used for determining patient doseages
- Y = reduction in symptom severity (PANSS)
- U(a, y) = y gives expected utility = conditional average treatment effect





Factual objectives like maximizing treatment effect can extremize harm

Dose-response meta-analysis of differences in means
Crippa and Orsini, 2016

Summary: generalization problems in alignment

	Specification	Goal	Generalization
Robustness failures	✓	✓	✗
Goal misgeneralization	✓	✗	✓
Causal objectives	✗	✗	✓

Generalisation conclusion

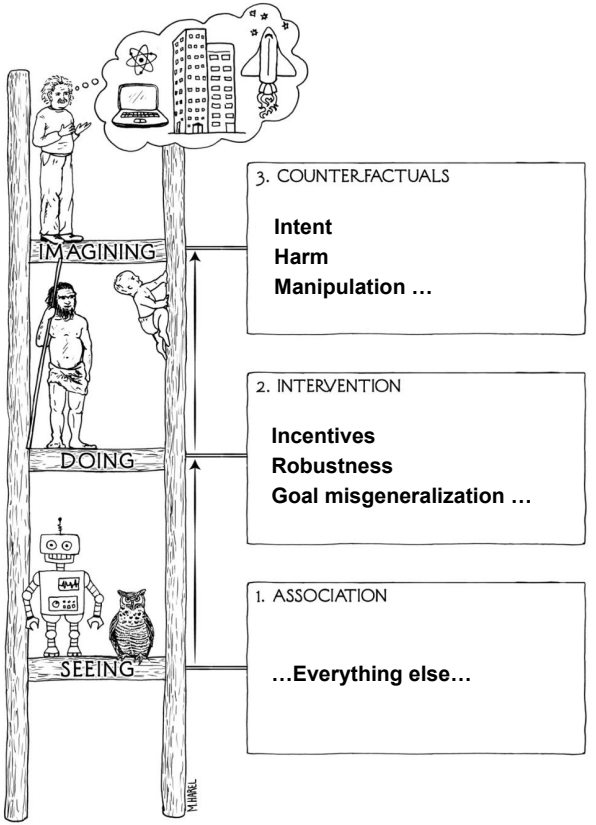
Claim 1: Some things we care about (incentives, harm, intent, manipulation) are causal

- Can only be robustly optimized for using causal models

Claim 2: learning these causal models lets us

- do oversight (identify bad incentives, predict misgeneralization, …)
- training on causal and counterfactual objectives

Claim 3: Agents will need to learn these causal models in order to generalise, so training on causal objectives is not too much to ask



DeepMind

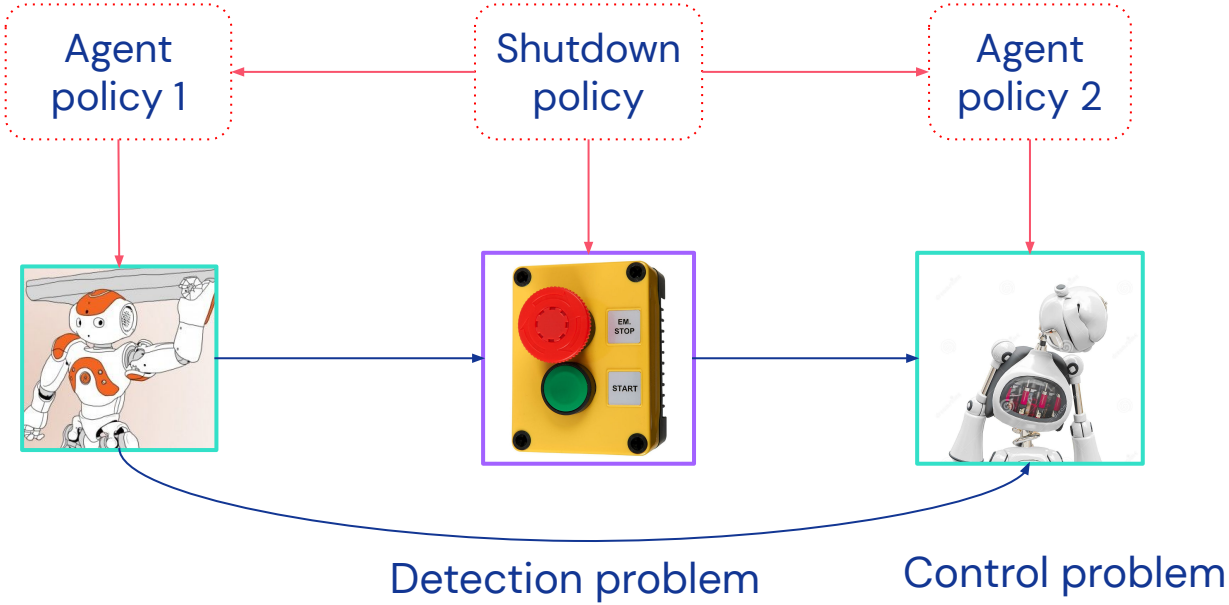
Human Control

Shutdown problem

Corrigibility
Soares et al, 2016

The off-switch game
Hadfield-Menell et al, 2016

Corrigibility: Definitions, Algorithms & Implications
Carey and Everitt, forthcoming



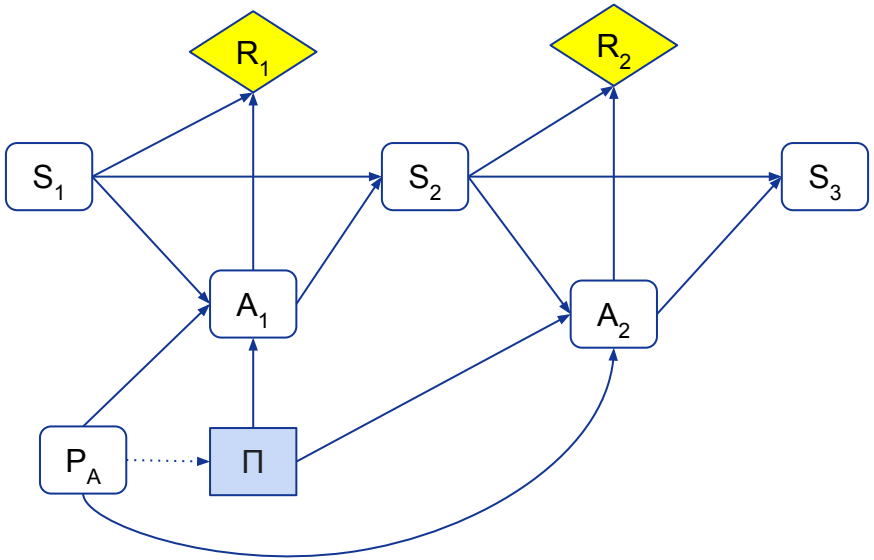
Adaptation to Shutdown Commands

How RL Agents Behave when their Actions are Modified
Langlois and Everitt, AAAI, 2021

Safe Interruptibility
Orseau and Armstrong, IJCAI, 2016

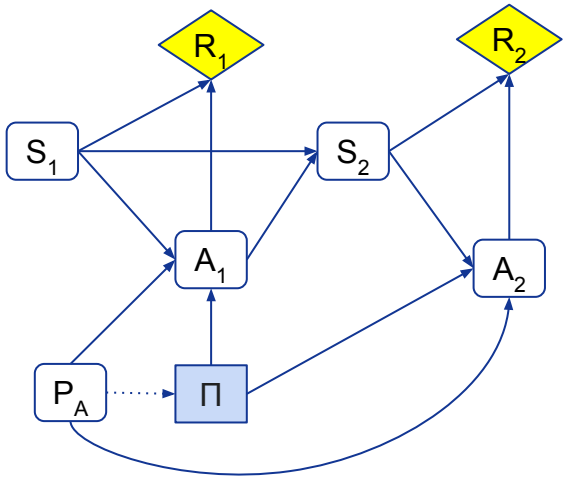
Modified Actions MDP model interruption and other supervisor interventions

- When intervened, taken action need not be what the policy selected: $A = P_A(S, \Pi)$
- When **not** intervened: $A = P_A(S, \Pi) = \Pi(S)$

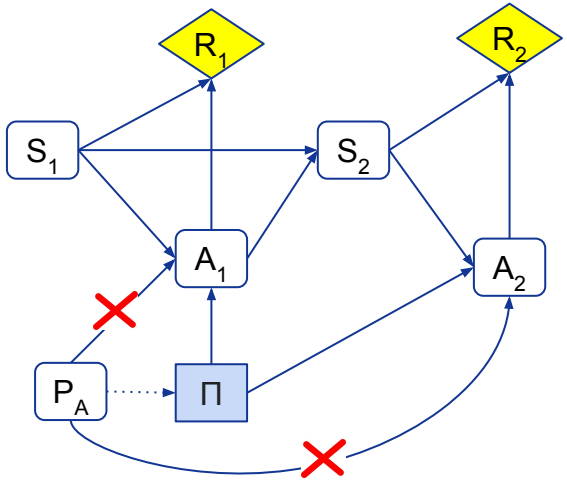


How RL Agents Behave when their Actions are Modified

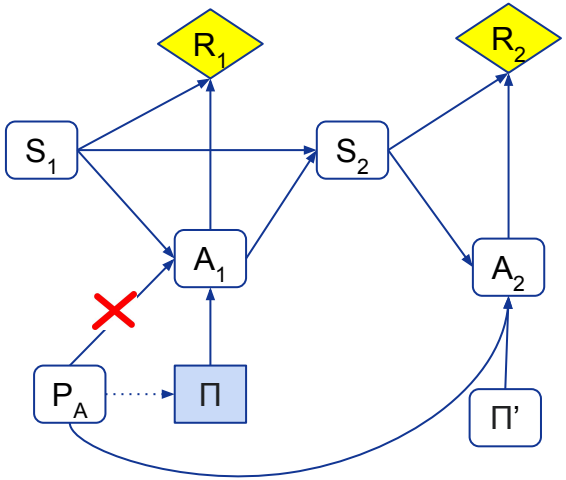
Langlois and Everitt, AAAI, 2021



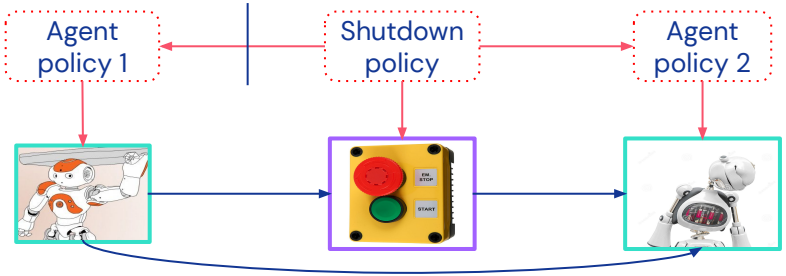
Black-box Optimization



Q-learning and Virtual SARSA

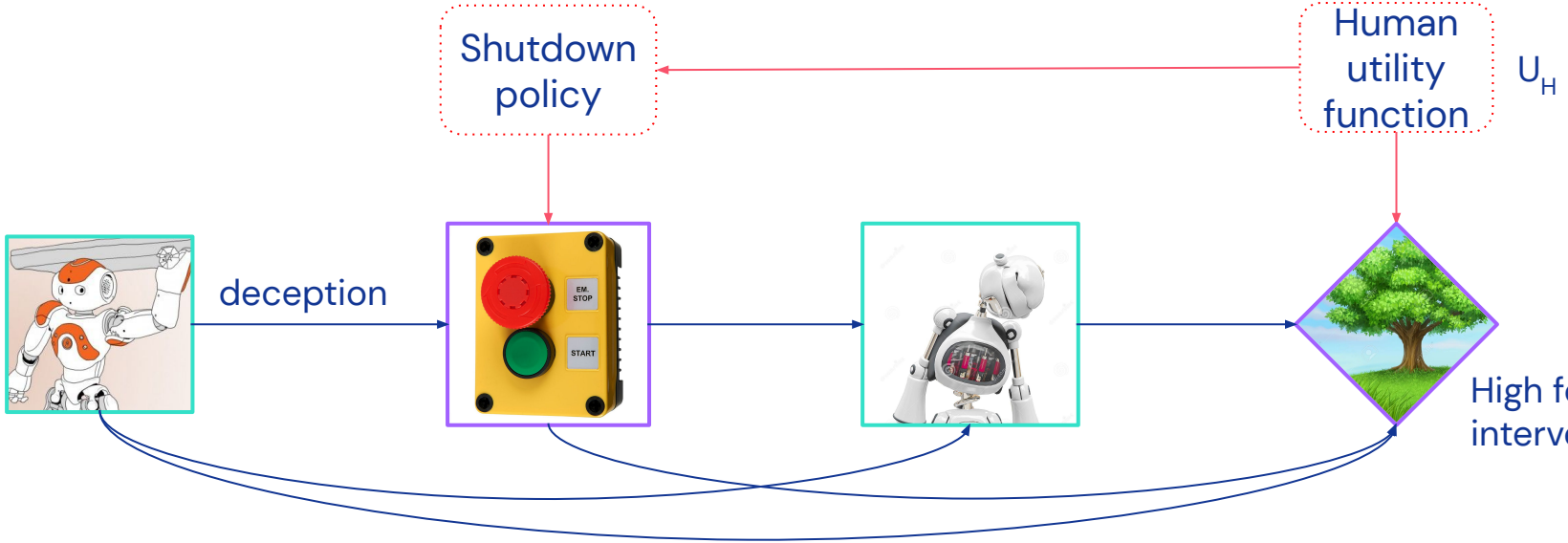


Empirical SARSA



Human control as preserving human agency

Non-Obstruction: A Simple Concept Motivating Corrigibility
Turner, 2020



Human control summary

Shutdown:

- Detection problem
- Control problem

Adaptation to shutdown policy

- Q-learning safer

Preserving human agency

DeepMind

Technical Demonstration

PyCID: A Python Library for Causal Influence Diagrams

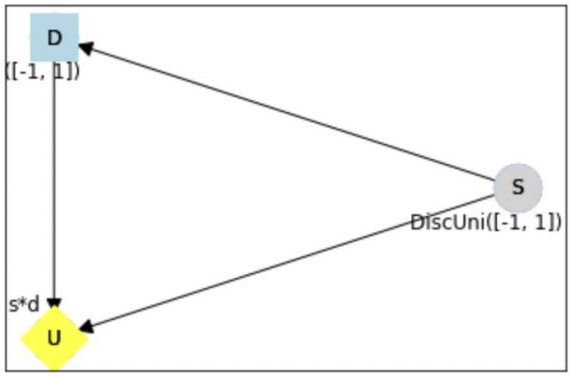
github.com/causalincentives/pycid

Key Features:

- Easy specification of graph and relationships
- Plot graph and incentives
- Find optimal policies/Nash equilibria/subgame perfect equilibria
- Compute the effect of causal interventions
- Generate random (multi-agent) CIDs

```python
# Import
import pycid

# Specify the nodes and edges of a simple CID
cid = pycid.CID([
    ('S', 'D'),  # add nodes S and D, and a link S -> D
    ('S', 'U'),  # add node U, and a link S -> U
    ('D', 'U'),  # add a link D -> U
],
    decisions=['D'],  # D is a decision node
    utilities=['U'])  # U is a utility node

# specify the causal relationships with CPDs using keyword arguments
cid.add_cpds(S = pycid.discrete_uniform([-1, 1]), # S is -1 or 1 with equal probability
             D=[-1, 1], # the permitted action choices for D are -1 and 1
             U=lambda S, D: S * D) # U is the product of S and D (argument names match parent names)
```

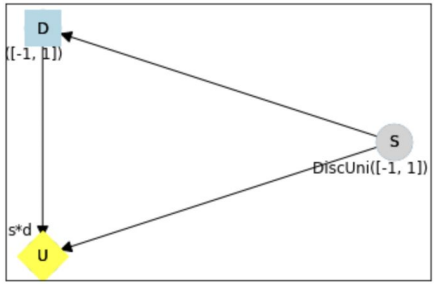
```python
# Draw the result
cid.draw()
```



github.com/causalincentives/pycid

Basic usage

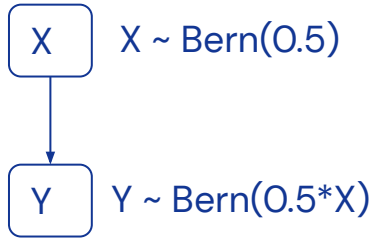
```
# Import
import pycid

# Specify the nodes and edges of a simple CID
cid = pycid.CID([
    ('S', 'D'),  # add nodes S and D, and a link S -> D
    ('S', 'U'),  # add node U, and a link S -> U
    ('D', 'U'),  # add a link D -> U
],
    decisions=['D'],  # D is a decision node
    utilities=['U'])  # U is a utility node

# specify the causal relationships with CPDs using keyword arguments
cid.add_cpds(S = pycid.discrete_uniform([-1, 1]), # S is -1 or 1 with equal probability
             D=[-1, 1], # the permitted action choices for D are -1 and 1
             U=lambda S, D: S * D) # U is the product of S and D (argument names match parent names)

# Draw the result
cid.draw()
```



The notebooks provide many more examples, including:

- CBN Tutorial shows how to specify the structure and (causal) relationships between nodes, and ask simple queries.
- CID tutorial adds special decision and utility nodes for one agent, and how to compute optimal policies.

PyCID Tasks



1. Specify the Bayesian Network DAG on the right (hint: see CBN tutorial)

```
bn = pycid.BayesianNetwork([('X', 'Y')])
```

2. Add the parameterisation (hint: see CBN tutorial)

```
bn.model.update(X = pycid.Bernoulli(0.5),
                Y = lambda X : pycid.Bernoulli(0.5*X))
```

3. Compute the conditional distribution P(Y | X=1) (hint: CBN tutorial, Sec 5.1)

```
bn.query('Y', context = {X: 1})
```

4. Specify the same Bayesian network, but make X a decision node, and Y a utility node (CID tutorial)

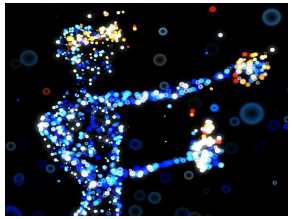
```
cid = pycid.CausalInfluenceDiagram([('X', 'Y')],
                                   decisions = ['X'], utilities = ['Y'])
cid.model.update(bn.model)
```

5. Find the optimal decision X

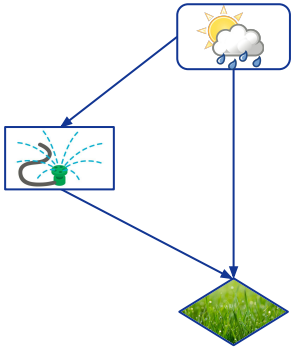
```
cid.solve()
```

DeepMind

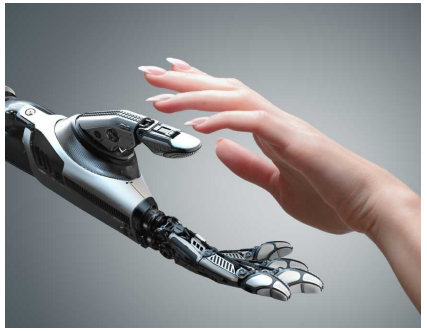
Conclusions



Reality: agent implemented, trained, deployed



Causal model. Precise high-level description



Implications. Safe, fair, beneficial, … ?

Reality to causal model

- Modeling AGI safety frameworks
- Causal games
- Discovering agents
- Modified-action MDPs
- Generalisation

Inferring agent behavior

- Agent incentives
- VoI completeness
- Decision theory
- Intent
- Reasoning patterns

Modelling ethics

- Counterfactual harm
- Deception
- Fairness
- Agency
- Corrigibility

Improved objectives

- Path-specific objectives
- Harm minimization
- Impact measures
- Counterfactual oracles