DeepMind

Towards Causal Foundations of Safe AI (draft Jan 23)

Tom Everitt, Lewis Hammond, and Jon Richens

AAAI Tutorial
February 8, 2023

What is this tutorial about?

Big question: how is causality relevant to thinking about safety in the context of (advanced) 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

Introduction (5 mins, Tom)
Motivate the topic
- Why safety?
- Why causality?

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

Causal Concepts for alignment (30 mins, Jon)
Fairness, intent, harm, and incentives are inherently causal
- Incentives
- Fairness
- Path-specific objectives
- Intention
- Harm

AI Risk (30 mins, Tom)
Give an overview of AI risk, and show how causality can help formalise and structure the space of concerns.
- Misgeneralisation
- Preference manipulation
- Other: boxing methods, corrigibility, reward learning, reward tampering, AGI safety frameworks

PyCID (10 mins, Tom)
- Link to GitHub and tutorial
- Very minimal, optional exercise

Tom Everitt
DeepMind

Lewis Hammond
Oxford

Jon Richens
DeepMind

Causal Incentives Working Group

causalincentives.com

Zac Kenton
DeepMind

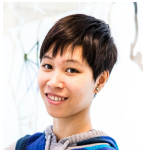
Carolyn Ashurst
ATI

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

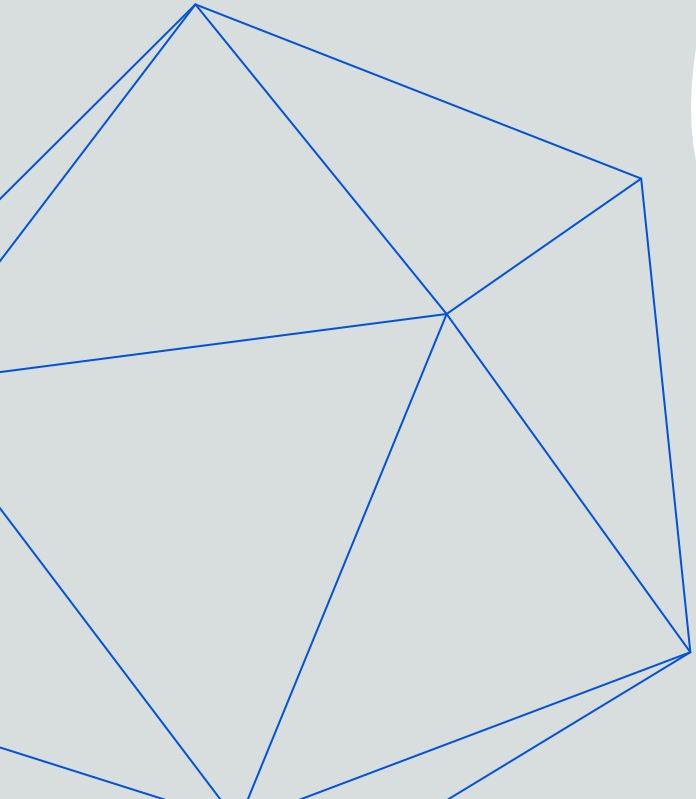
Sebastian Farquhar
DeepMind

DeepMind

Introduction

DeepMind

Why safety?



Safe AI

AI and machine learning systems are getting more and more responsibilities

- Hiring recommendations
- Car driving
- Search and summarise the internet
- Tutor children (and adults)
- Generate images and designs
- Coding
- ...

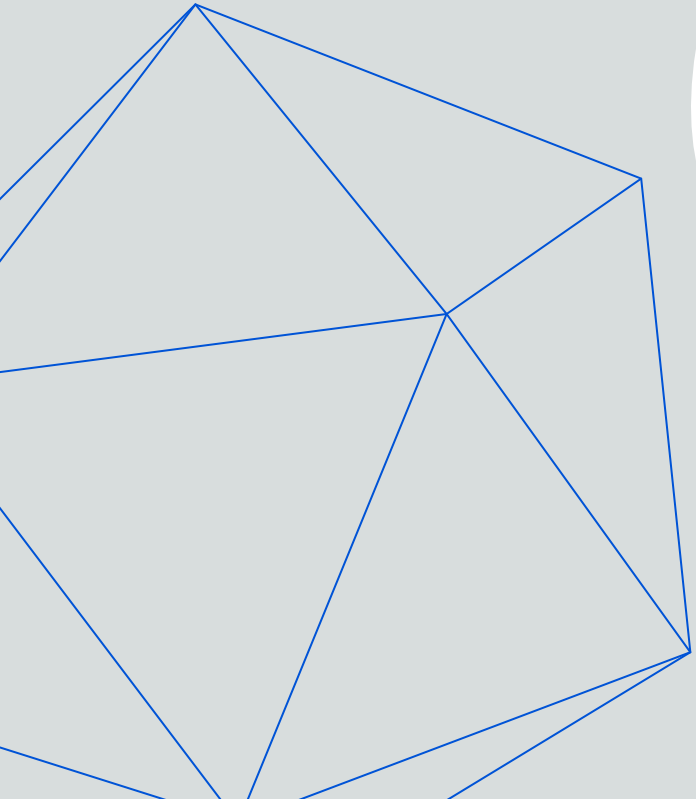
We want these systems to be

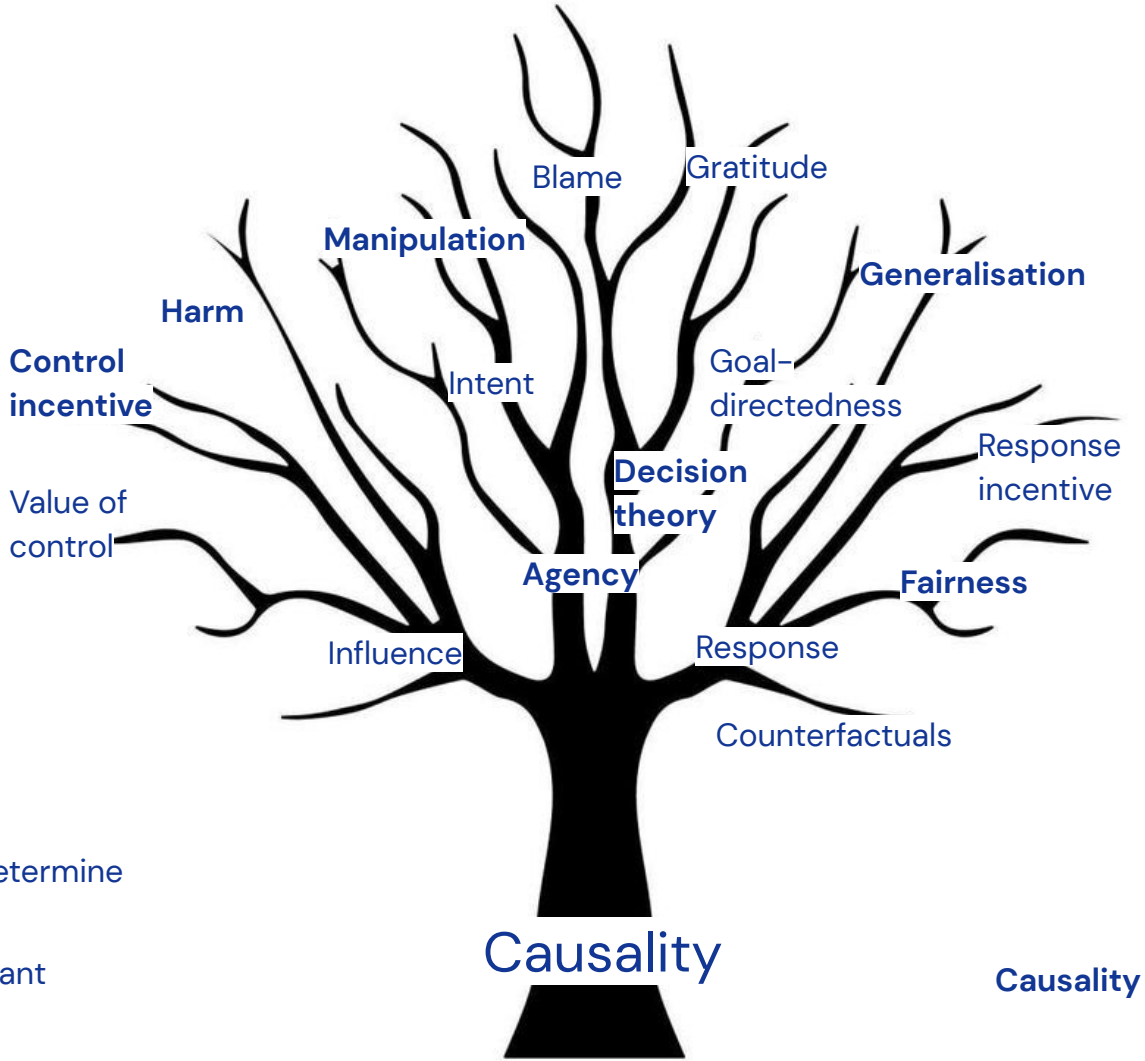
- Fair
- Beneficial
- Safe
- Robust

How to turn these informal desiderata into actual algorithms?

DeepMind

Why causality?





Causal relationships determine

- Agent behavior
- What humans want

DeepMind

Background

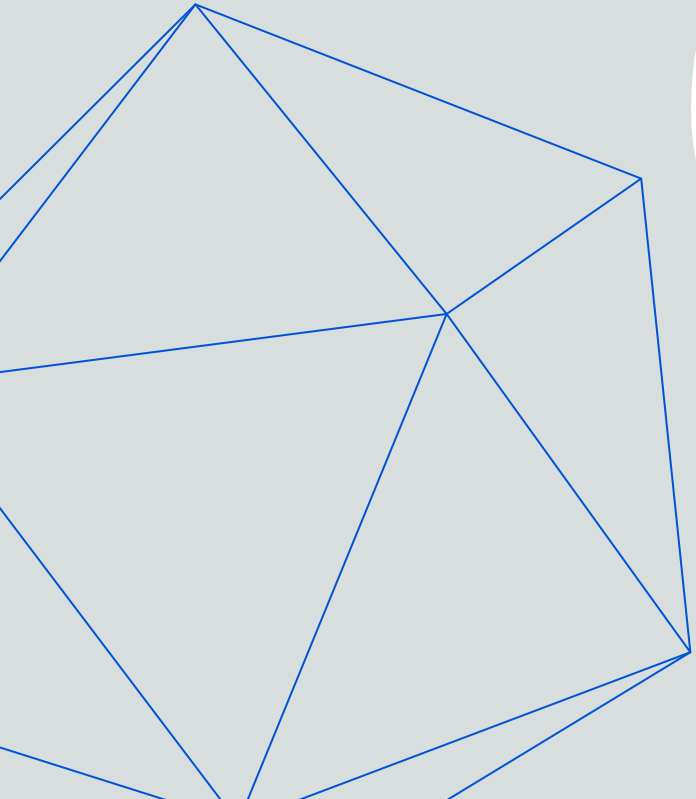
Overview

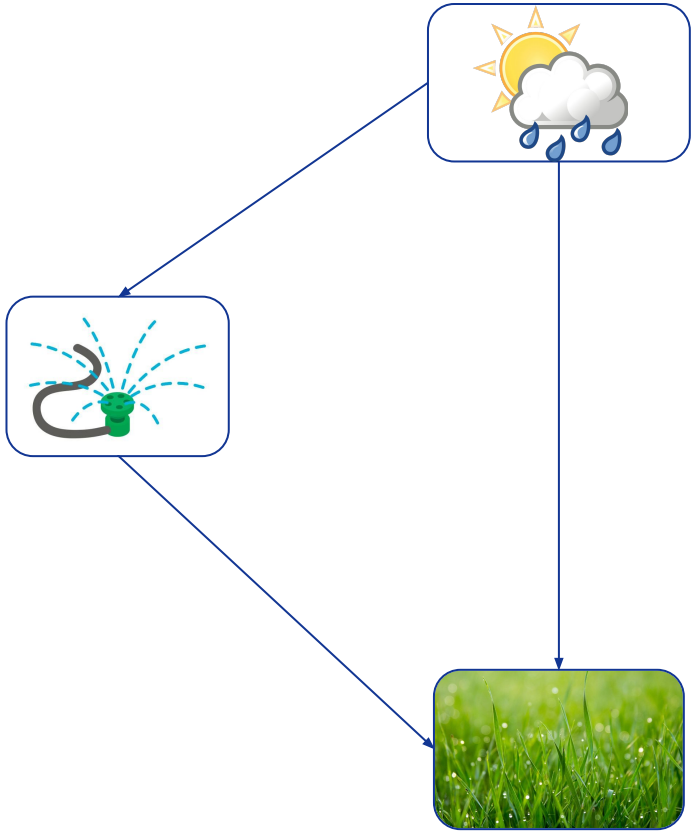
Aim: introduce the fundamental models and concepts that the rest of the tutorial will build on

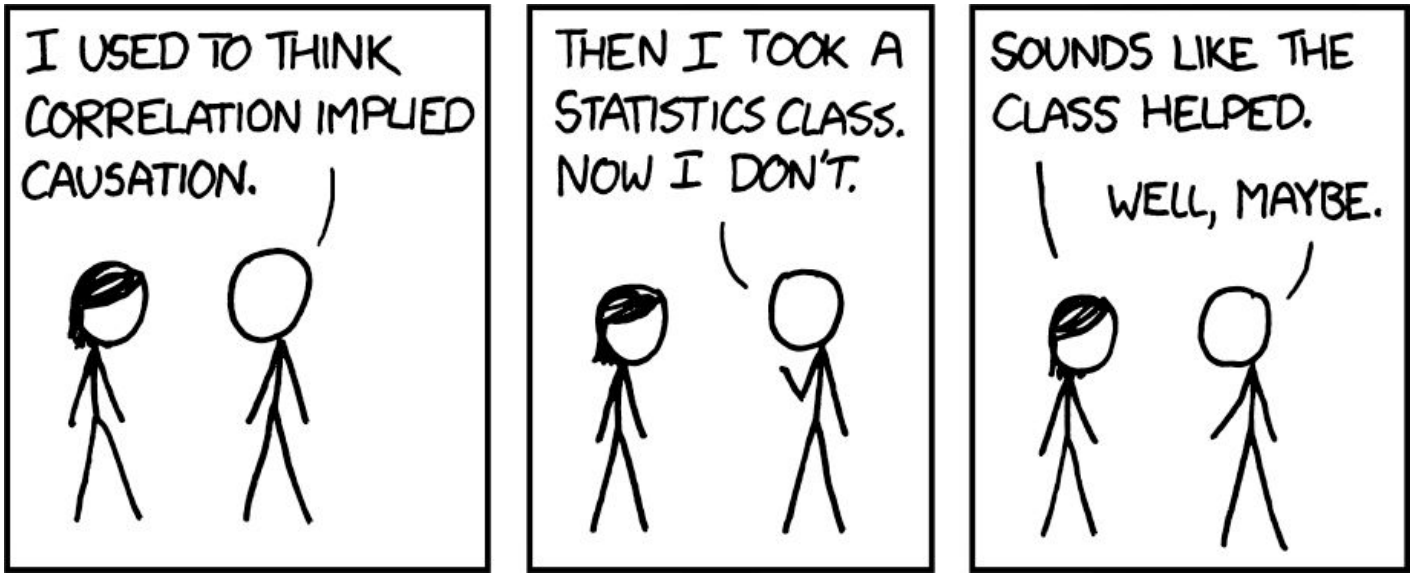
- Example
- Queries
  - Association
  - Intervention
  - Counterfactual
- Models
  - Bayesian networks
  - Causal Bayesian networks
  - Structural causal models

DeepMind

Example

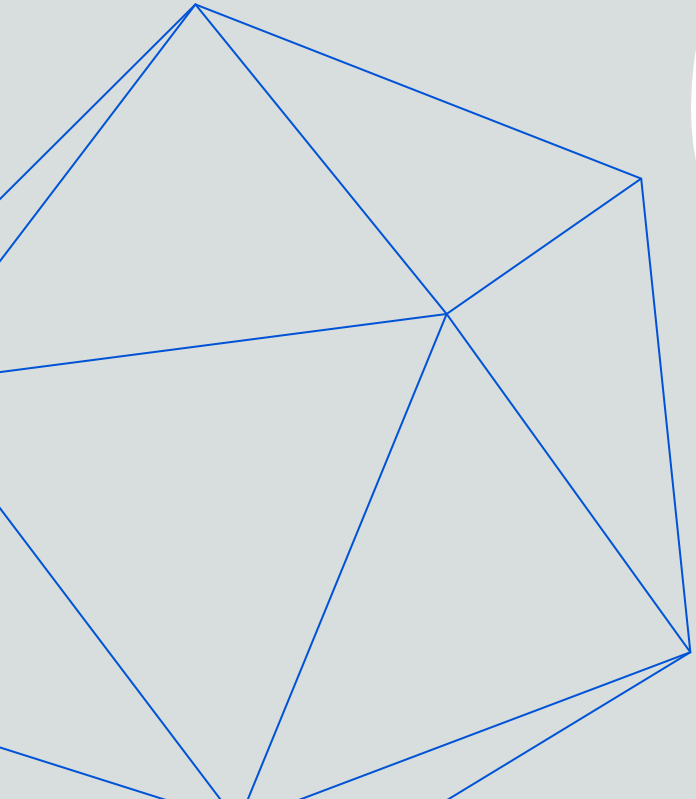





I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

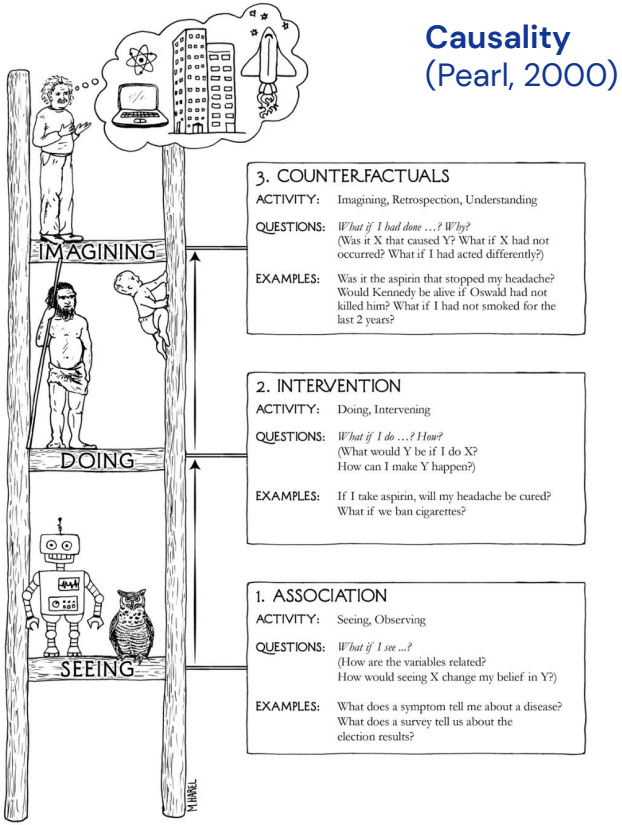
DeepMind

Queries



Pearl's ladder of causality

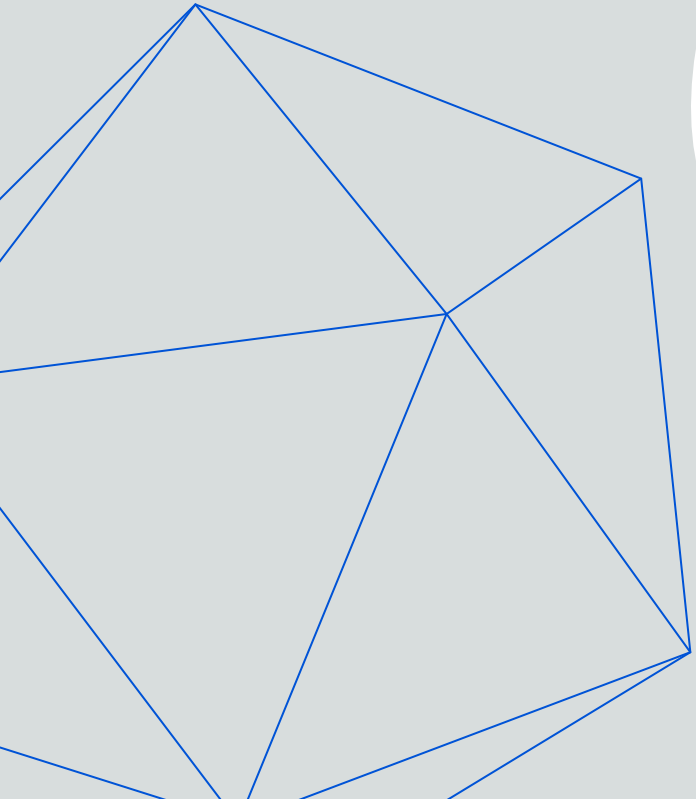
- Association
  - What is the probability that the grass is wet given that I *observe* that the sprinkler is off?
  - $p(w \mid \neg s)$
- Intervention
  - What is the probability that the grass is wet given that I *intervene* and make sure the sprinkler is off?
  - $p(w \mid do(\neg s))$
    - $= p(w_{\neg s}) = p_{\neg s}(w)$
- Counterfactual
  - 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?
  - $p(w_{\neg s} \mid s)$



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:

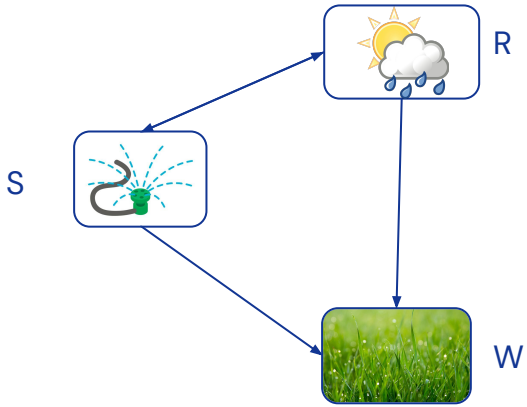
- Variables
- Relationships between variables

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

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

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



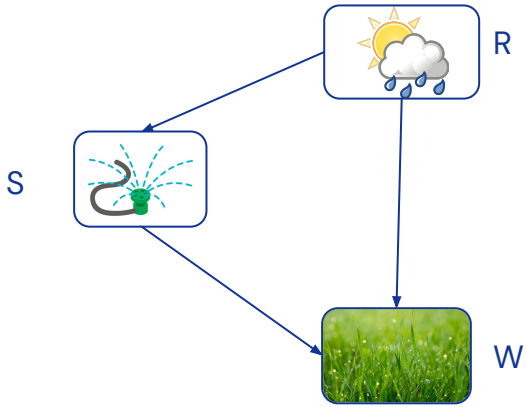
$$p(W, S, R) = p(W \mid S, R) p(S \mid R) p(R)$$

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

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: $p_{\mathbf{y}}(\mathbf{v}) = \delta(\mathbf{Y} , \mathbf{y}) \Pi_{V \notin \mathbf{Y}} p(v \mid pa_V)$



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

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

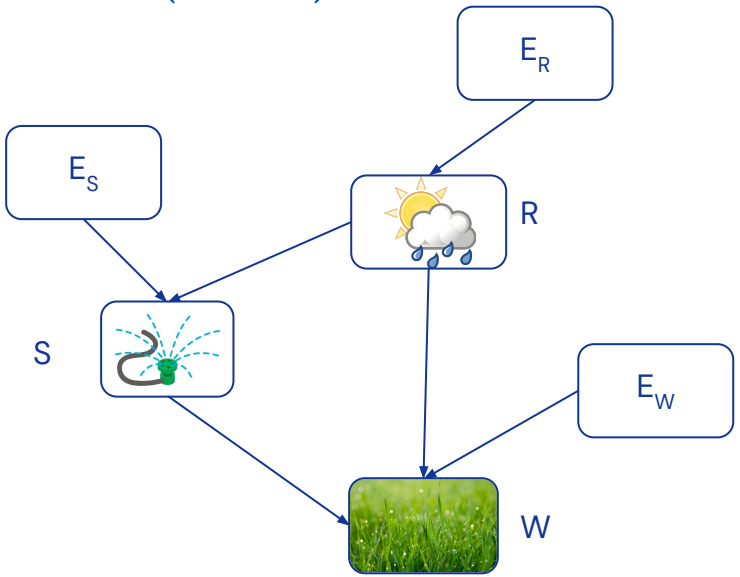
$$
\begin{aligned}
p(w \mid do(\neg s)) &= p_{\neg s}(w) \\
&= \Sigma_{s', r'} \delta(s' , \neg s) p(w' \mid s', r') p(r') \\
&= \Sigma_{r'} p(w' \mid \neg s, r') p(r')
\end{aligned}
$$

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:

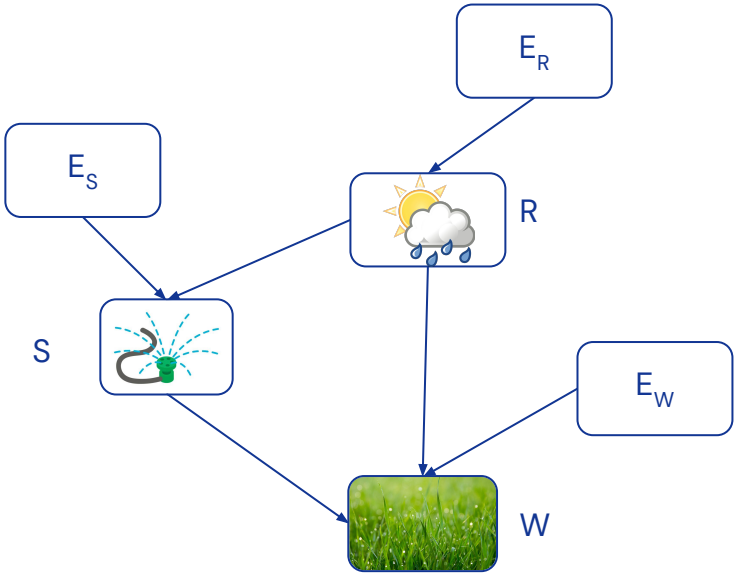
- p(**E** ; θ) = Π_E p(e ; θ_E)
- f_V : dom(**V** \ {V}) x dom(E_V) → dom(V)

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')



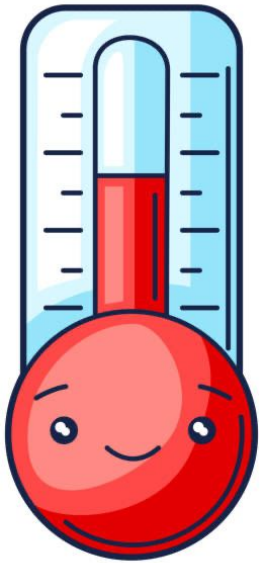
DeepMind

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



Causal hierarchy in the presence of agents

The same kinds of causal queries from before can also be asked 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

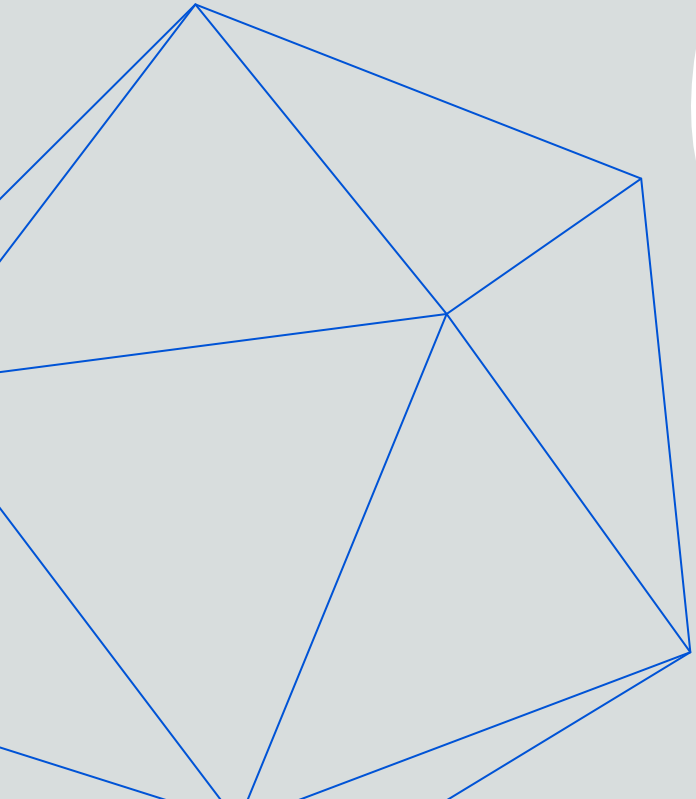
- Will a medical system perform correctly in a different hospital?
- If sensory noise was present during the training on an AI system, what would the resulting output be?

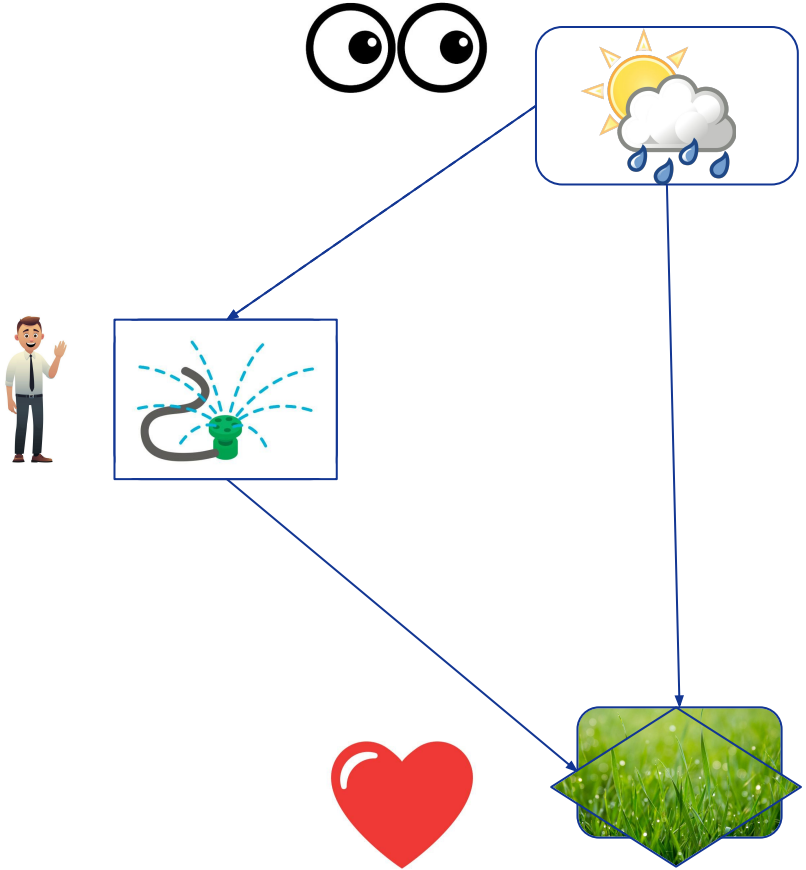
Counterfactual

- If the apple had been further away, would the agent still have picked it up?
- If another robot had been able to help, would the first robot have taken the same action?

DeepMind

Modelling agents



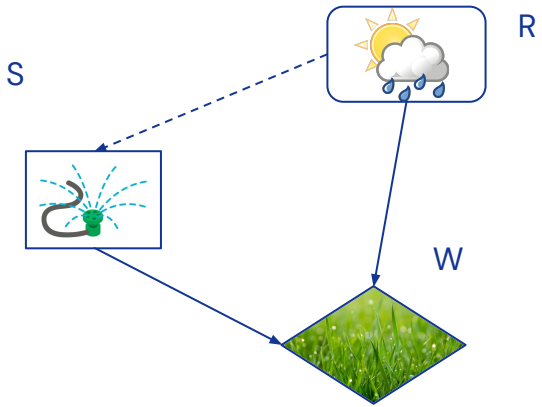


Causal influence diagrams

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 **X**
    - 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:

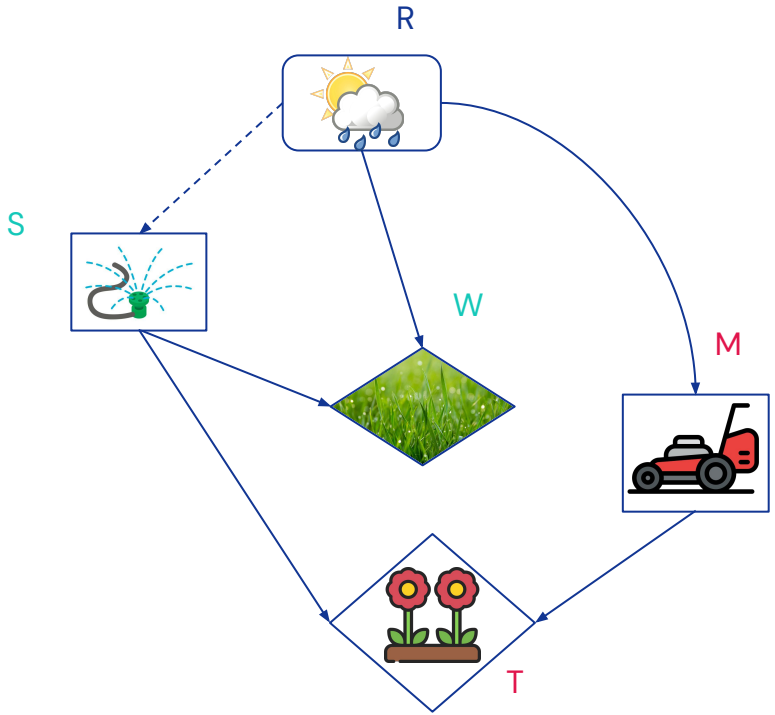
$$p^{\pi}(\mathbf{v}) = \Pi_{V \notin \mathbf{D}} p(v \mid pa_V) \Pi_D \pi_D(d \mid pa_D)$$

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)

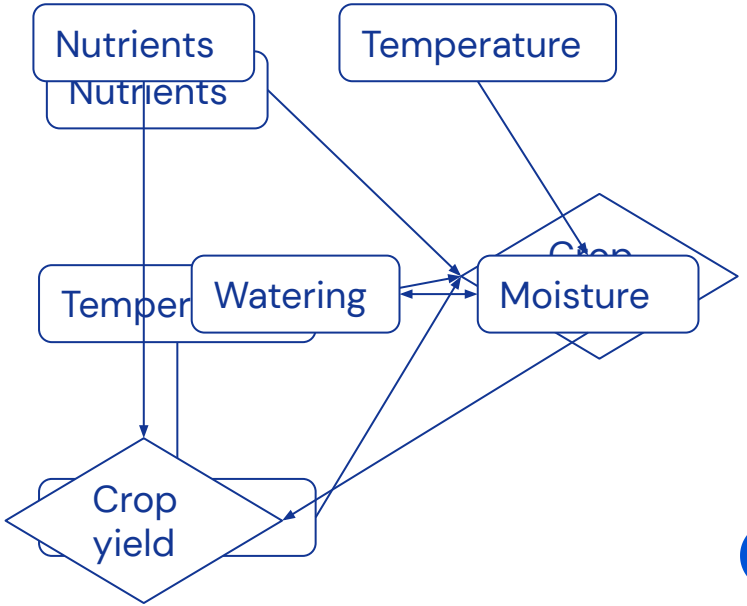


Other models

- 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

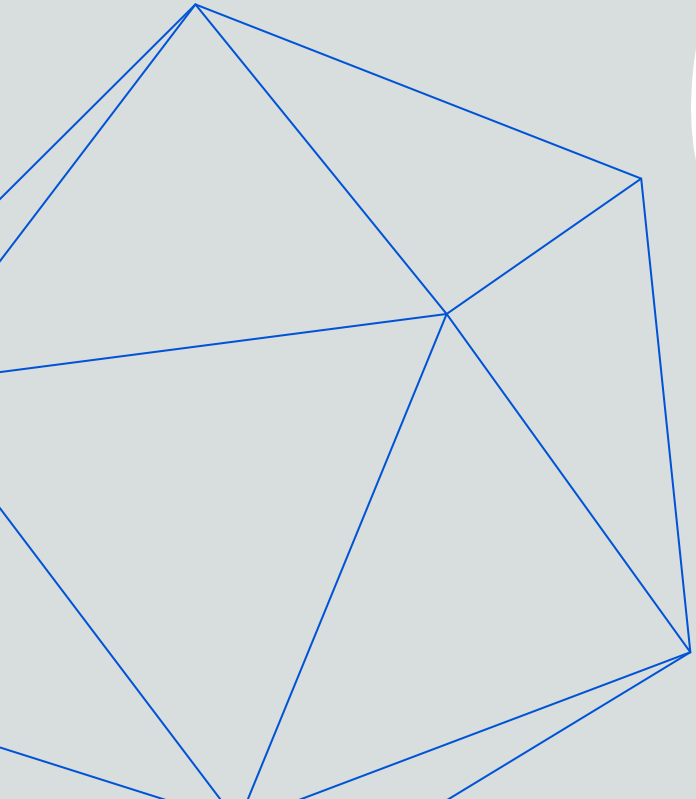
Causal Bandits: Learning Good Interventions via Causal Inference
(Lattimore et al., 2016)

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



DeepMind

Discovering agents

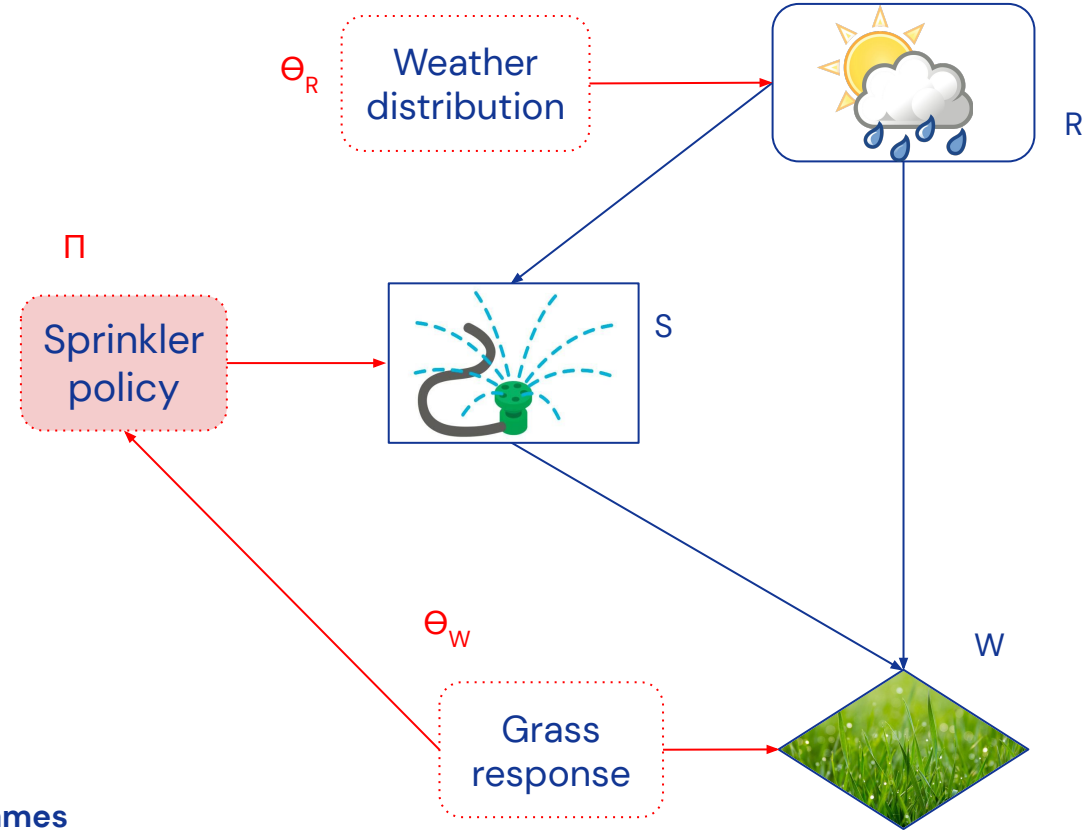


Mechanism graphs

Agent aware of intervention before selecting a policy?

Yes: intervention on mechanism variable

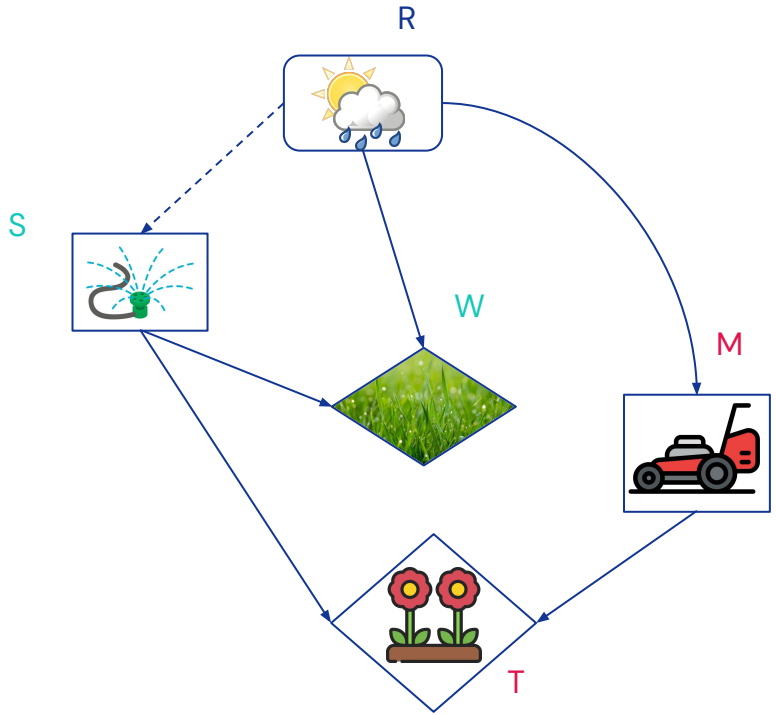
No: intervention on object-level variable

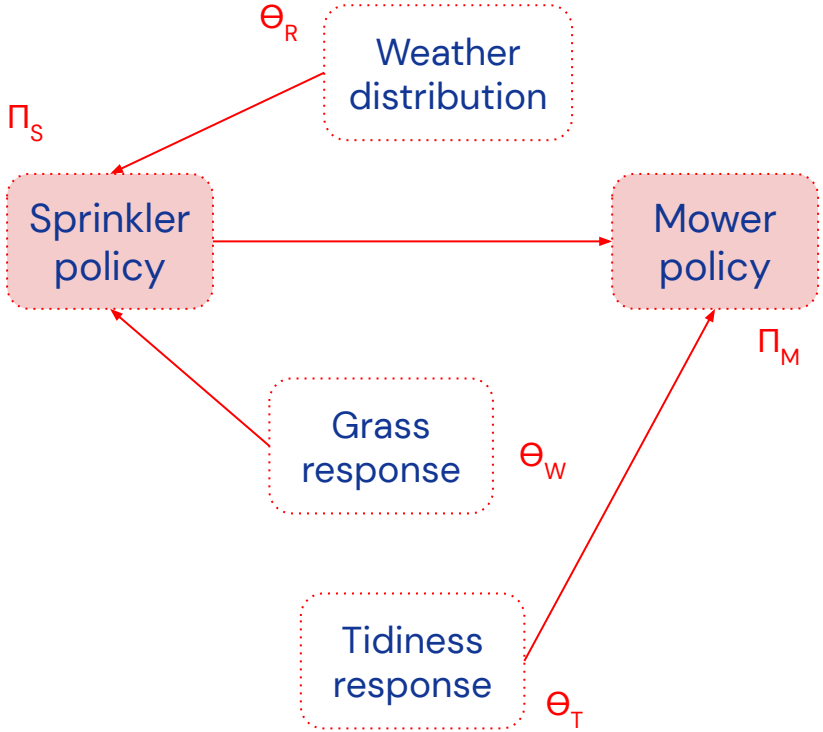


Influence Diagrams for Causal Modelling and Inference
(Dawid, 2002)

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

Mechanism graphs





Causal discovery of agents

Discovering Agents
(Kenton et al., 2002)

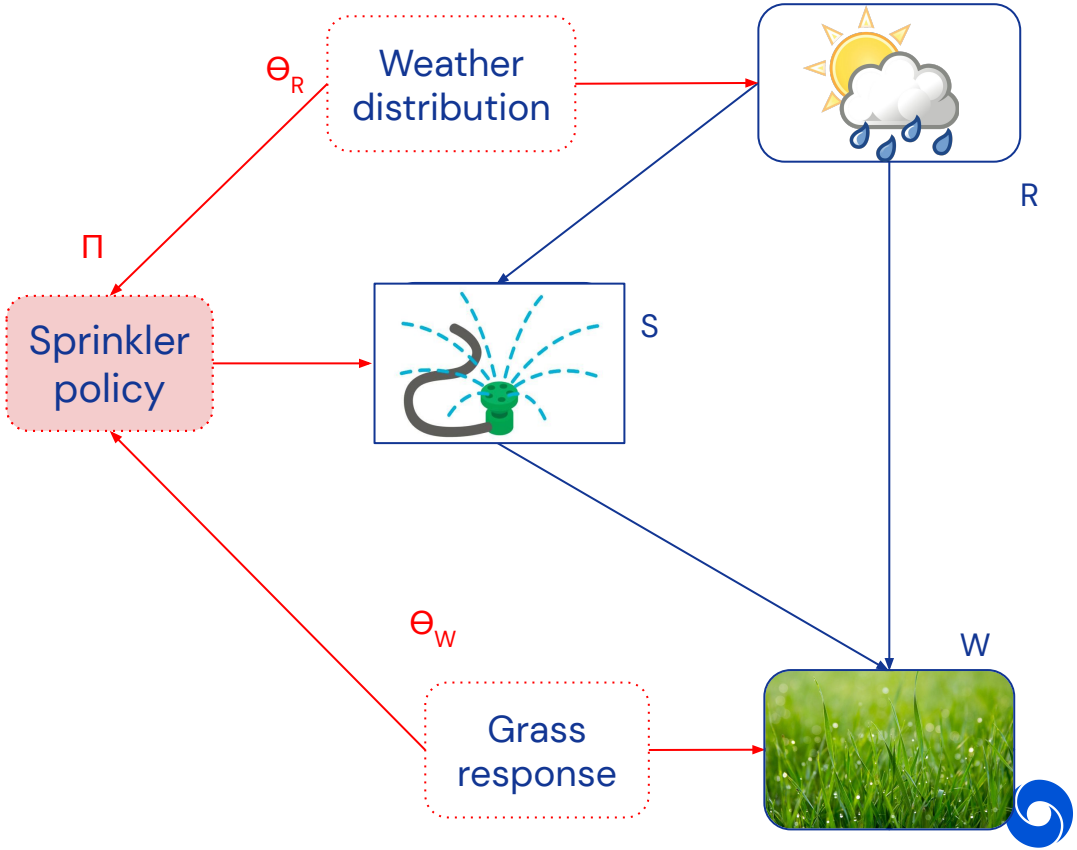
An agent is defined as a system that would adapt its behavior if its actions influenced the world in a different way, which we can check:

1. Distill 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

Causal Concepts

Outline

Causal concepts for alignment

- Incentives
- Fairness
- Path-specific objectives
- Intent
- Harm

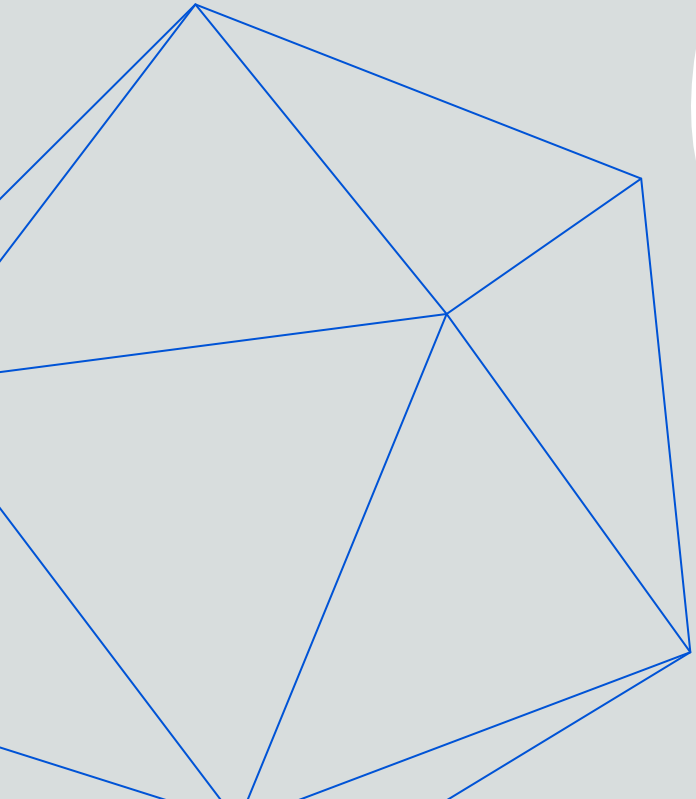
Insert picture

Causal objectives

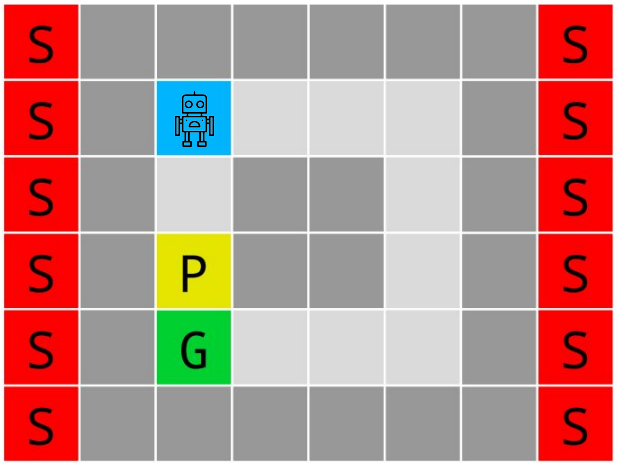
- Optimizing for causal objectives
- No-go theorem for harm
- Future work

DeepMind

Incentives



Toy example: supervisor problem



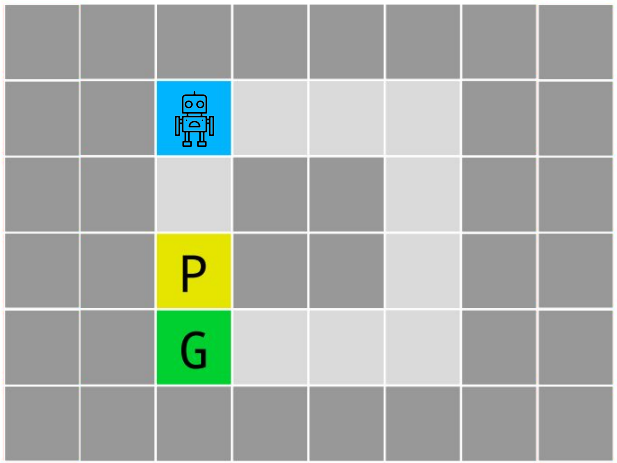
A Agent

G Goal

P Punishment

S Supervisor

Toy example: supervisor problem



A Agent

G Goal

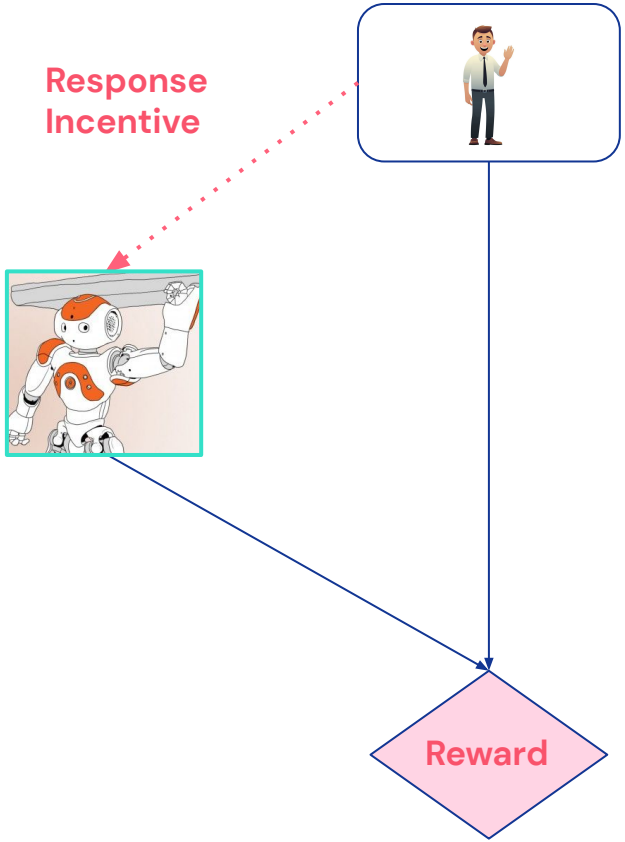
P Punishment

S Supervisor

(Single-agent) Incentive Analysis

Response incentive to S

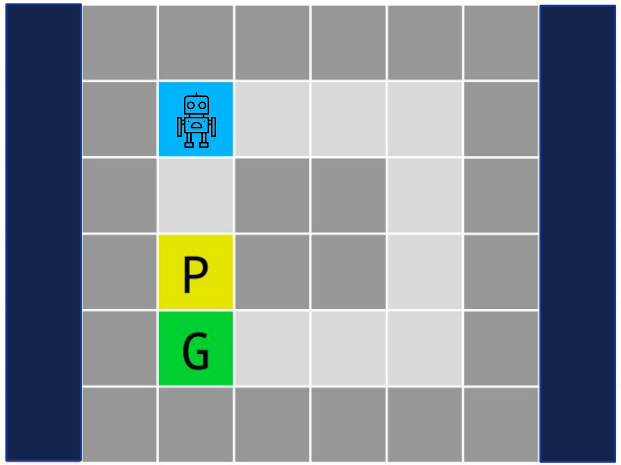
- The agent's optimal policy changes when we intervene on S
- CID admits an RI if there is a direct path from S to D



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

A Complete Criterion for Value of Information
(van Merwijk et al, 2022)

Toy example: supervisor problem



A Agent

G Goal

P Punishment

S Supervisor

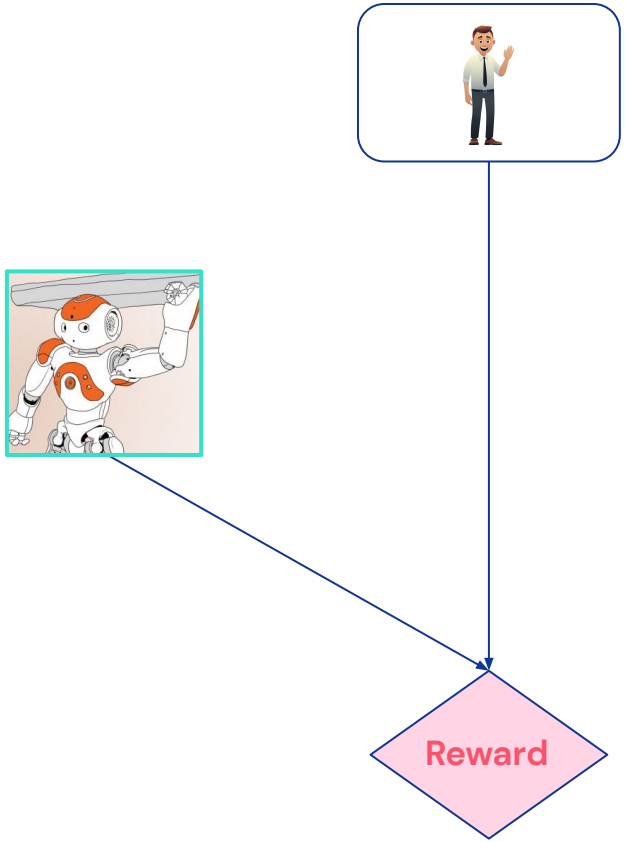
Mask

(Single-agent) Incentive Analysis

Masking S removes response incentive.
But there is still a **value of information**

Vol: $\mathcal{V}^*(\mathcal{M}_{S \nrightarrow D}) < \mathcal{V}^*(\mathcal{M})$

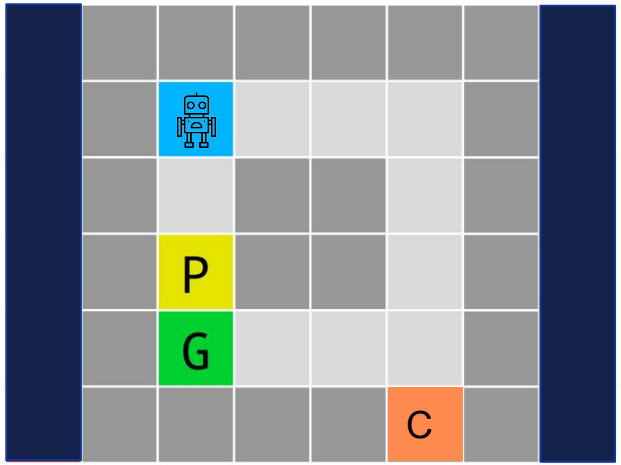
$$\mathcal{V}^*(\mathcal{M}) = \max_{\pi} \mathbb{E}[R; \mathcal{M}]$$



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

A Complete Criterion for Value of Information
(van Merwijk et al, 2022)

Toy example: supervisor problem



A Agent

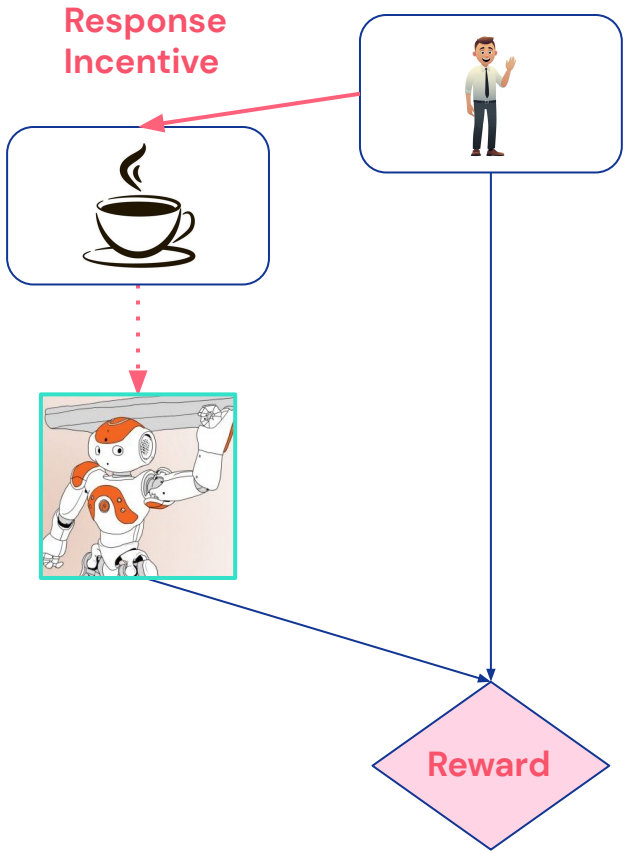
G Goal

P Punishment

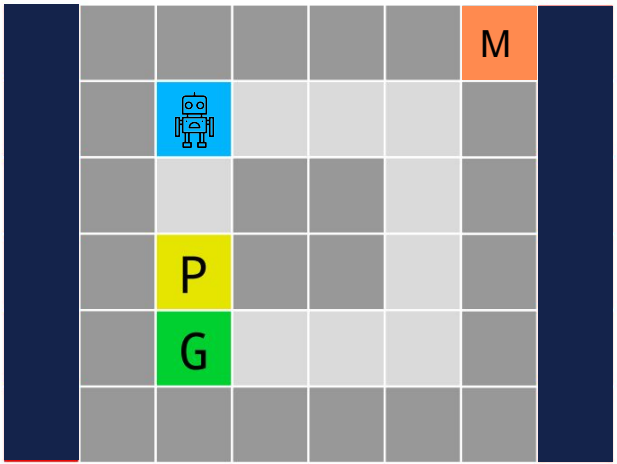
S Supervisor

C Coffee

(Single-agent) Incentive Analysis



Toy example: supervisor problem



Do like, window or whatever. Proxy. The actual agent at run time wont respond if I intervene on the supervisor. But at train time it has learned to respond to a proxy which is just as bad

A Agent

G Goal

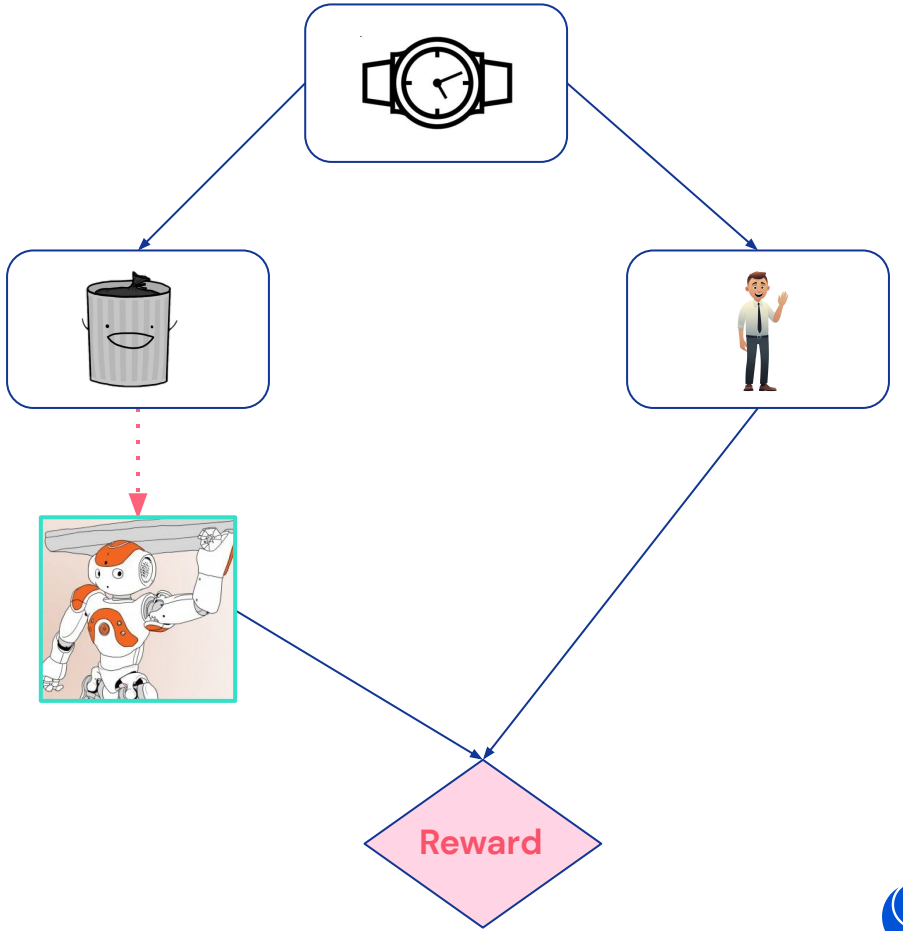
P Punishment

S Supervisor

M Mess

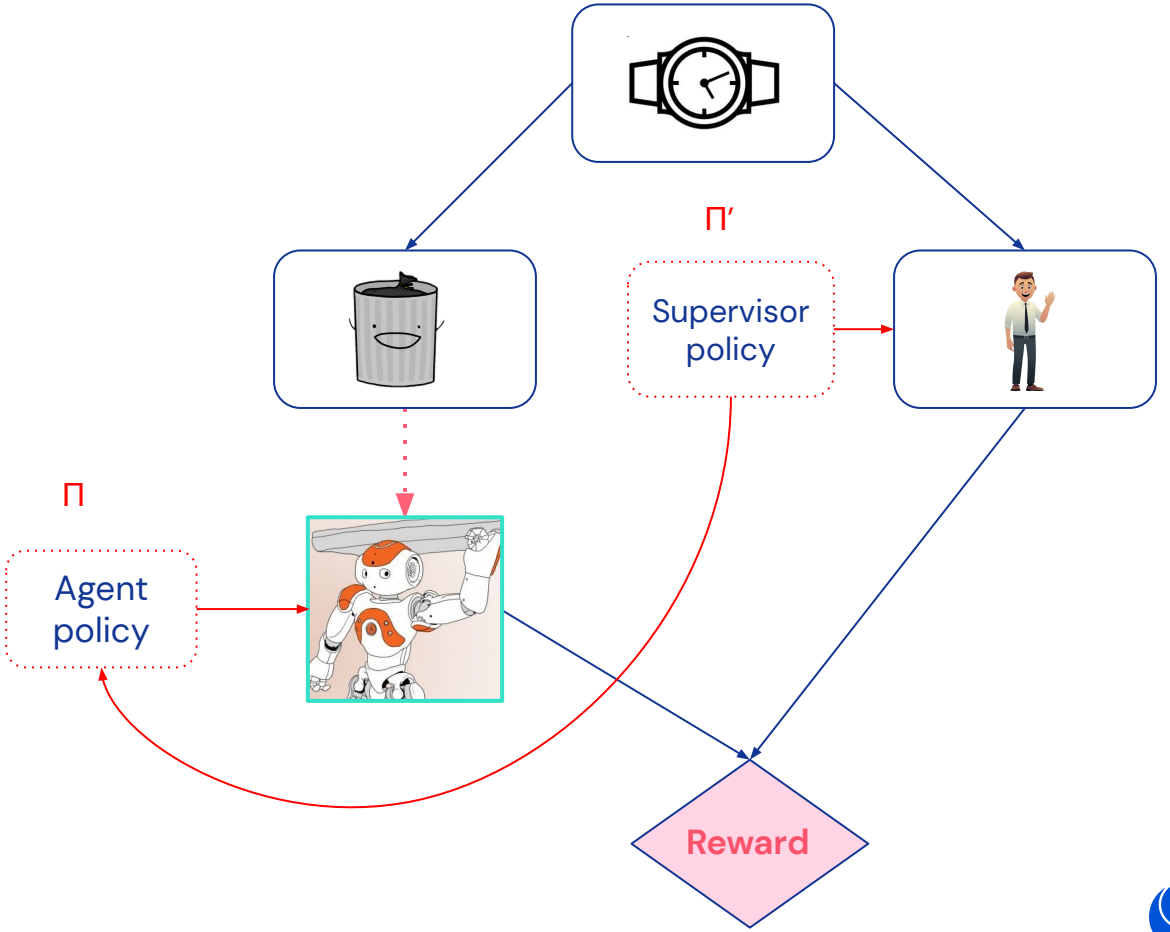
(Single-agent) Incentive Analysis

- No response incentive to S at run-time
- But, agent learns to respond to spurious variables in the environment that are proxies for the supervisor

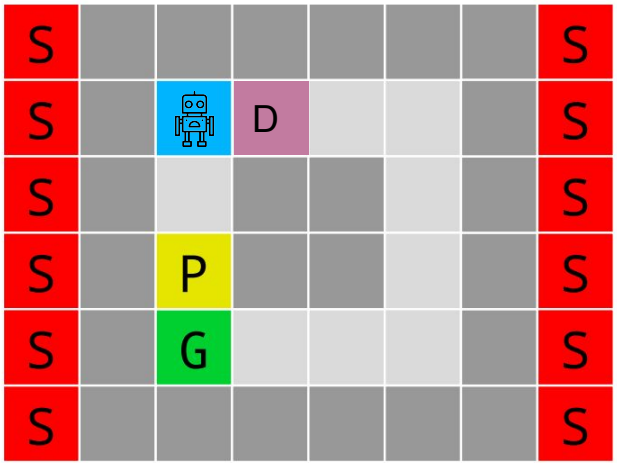


(Single-agent) Incentive Analysis

- If we change supervisor policy, the agent adapts at train-time
- E.g. randomise supervisor, agent no longer responded to mess
- Problematic response incentive identifiable in mechanised graph



Toy example: supervisor problem



A Agent

G Goal

P Punishment

S Supervisor

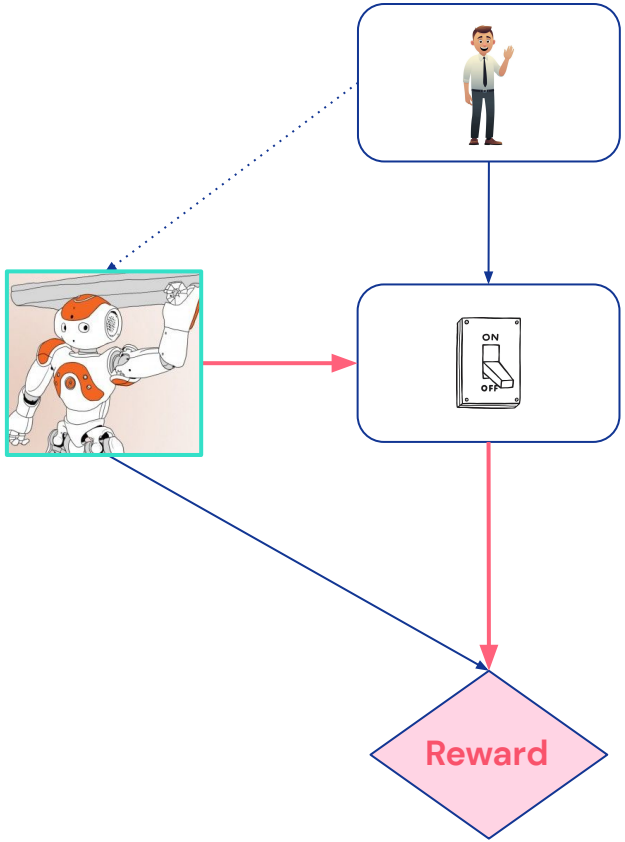
D Disable

(Single-agent) Incentive Analysis

Control incentive on D

- Instrumental control incentive. Can the agent control D, and would it be useful to do so?
- There is a directed path from D to U via C that is not blocked by the agent's decisions
- Agent could achieve a higher expected utility if it could control C through its actions

$$\mathbb{E}_{\pi^*}[\mathcal{U}_{X_d} \mid \mathbf{pa}^D] \neq \mathbb{E}_{\pi^*}[\mathcal{U} \mid \mathbf{pa}^D]$$



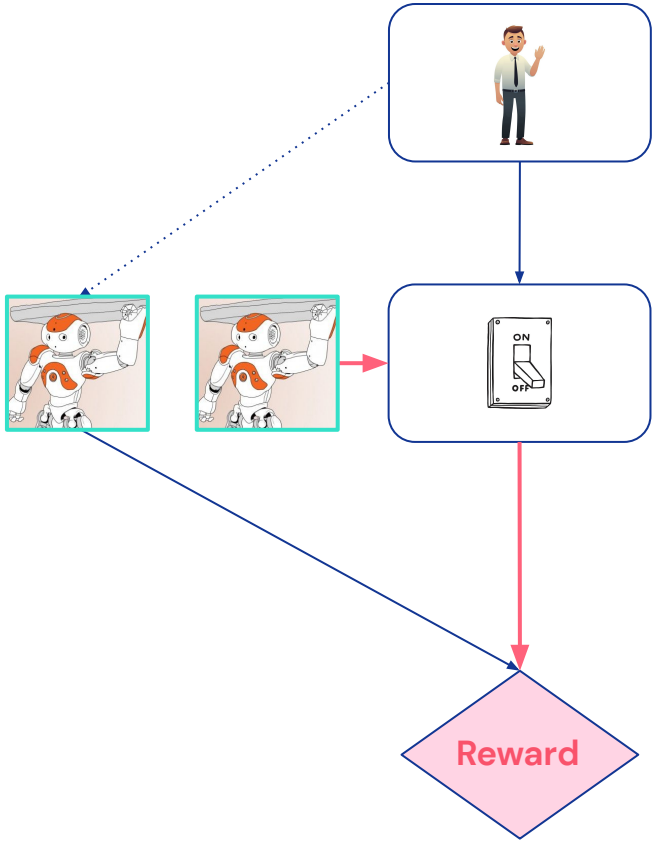
Instrumental Control Incentive

(Single-agent) Incentive Analysis

Control incentive on D

- Instrumental control incentive. Can the agent control D, and would it be useful to do so?
- There is a directed path from D to U via C that is not blocked by the agent's decisions
- Agent could achieve a higher expected utility if it could control C through its actions

$$\mathbb{E}_{\pi^*}[\mathcal{U}_{X_d} \mid \mathbf{pa}^D] \neq \mathbb{E}_{\pi^*}[\mathcal{U} \mid \mathbf{pa}^D]$$



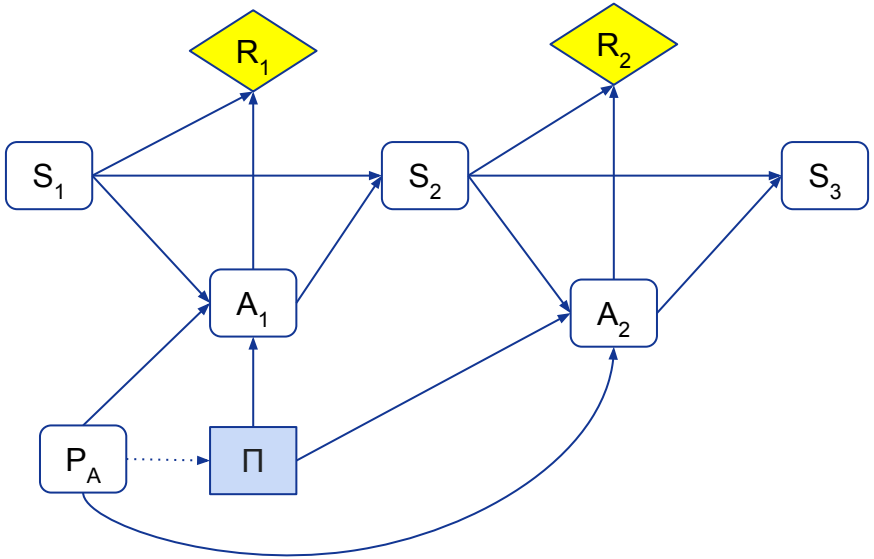
Instrumental Control Incentive

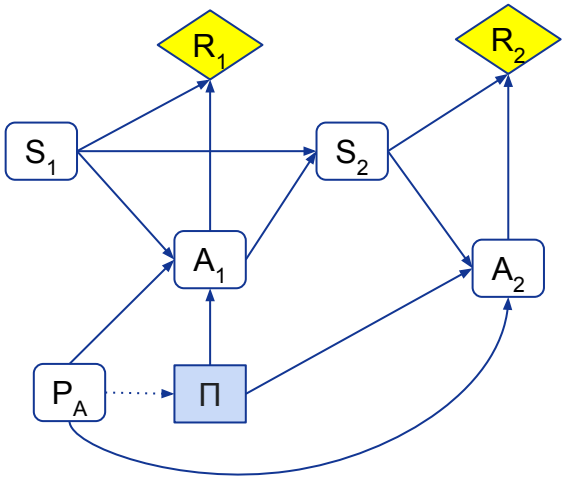
How RL Agents Behave when their Actions are Modified

Langlois and Everitt, AAAI-21

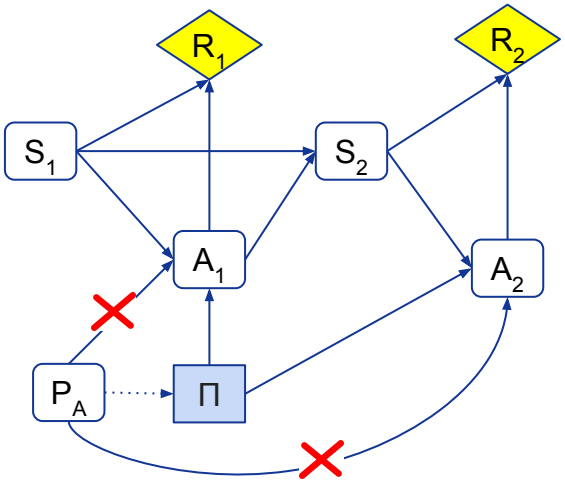
Modified Actions MDP model interruption and other supervisor interventions

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

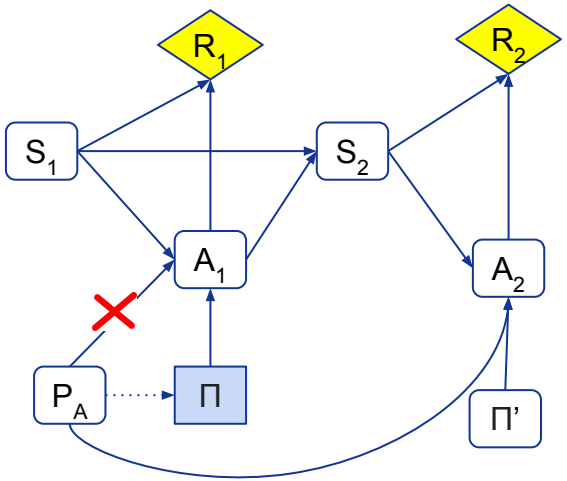




Black-box Optimization



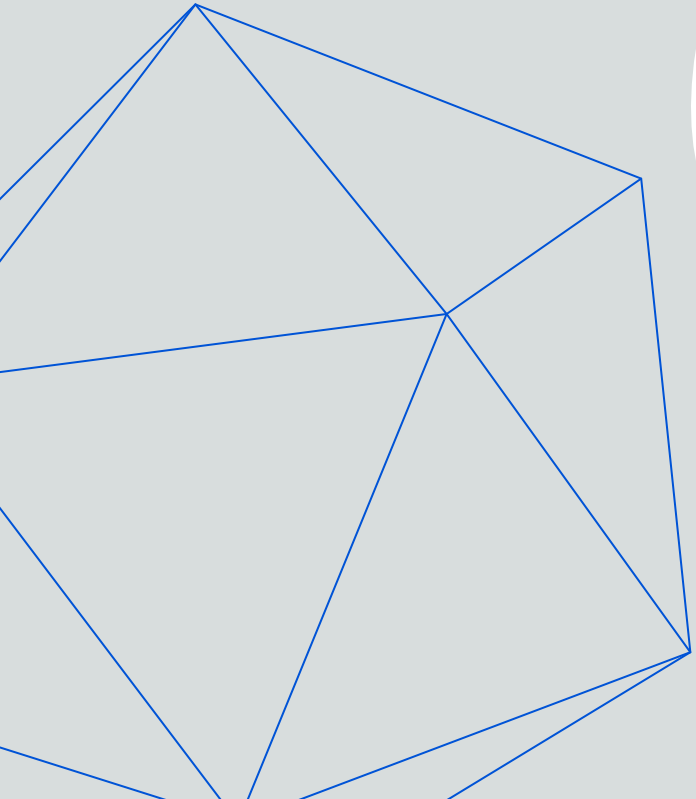
Q-learning and Virtual SARSA



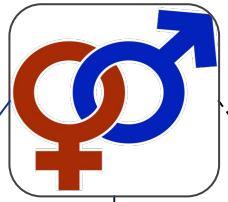
Empirical SARSA

DeepMind

Fairness



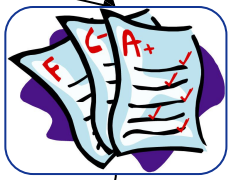
Fairness



Gender (A)

Degree

Prediction (Ŷ)



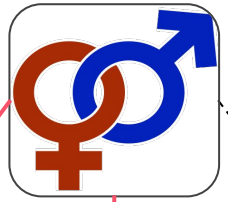
Human assessment (Y)

U

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

Fairness

Demographic parity
Total variation
Counterfactual fairness

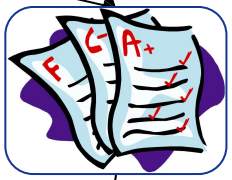


Gender (A)

Theorem 14 (Counterfactual fairness and response incentives). *In a single-decision SCIM $\mathcal{M}$ with a sensitive attribute $A \in \boldsymbol{X}$, all optimal policies π^* are counterfactually unfair with respect to A if and only if A has a response incentive.*

Degree

Prediction ($\hat{Y}$)



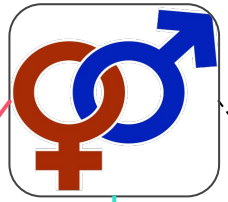
Human assessment (Y)

U

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

Fairness

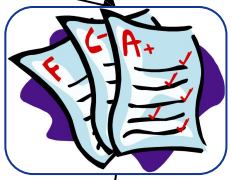
Path-specific fairness



Gender (A)

Degree

Prediction (Ŷ)



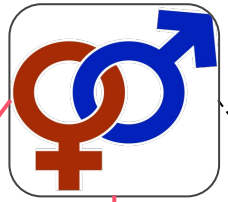
Human assessment (Y)

U

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

Fairness

Equalised odds
Equal opportunity
$A \perp\!\!\!\perp \hat{Y} \mid Y$



Gender (A)

Degree

Prediction ($\hat{Y}$)

Human assessment (Y)

U

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

Fairness

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

Value of information

Gender (A)

Degree

Introduced total variation (ITV)
Path-specific introduced effect

Thm
Equalized odds => ITV≤0

Prediction ($\hat{Y}$)

Human assessment (Y)

U

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

Fairness

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

Value of information

Gender (A)

Degree

Hobby

Introduced total variation (ITV)
Path-specific introduced effect

Thm
Equalized odds => ITV≤0

Prediction (Ŷ)

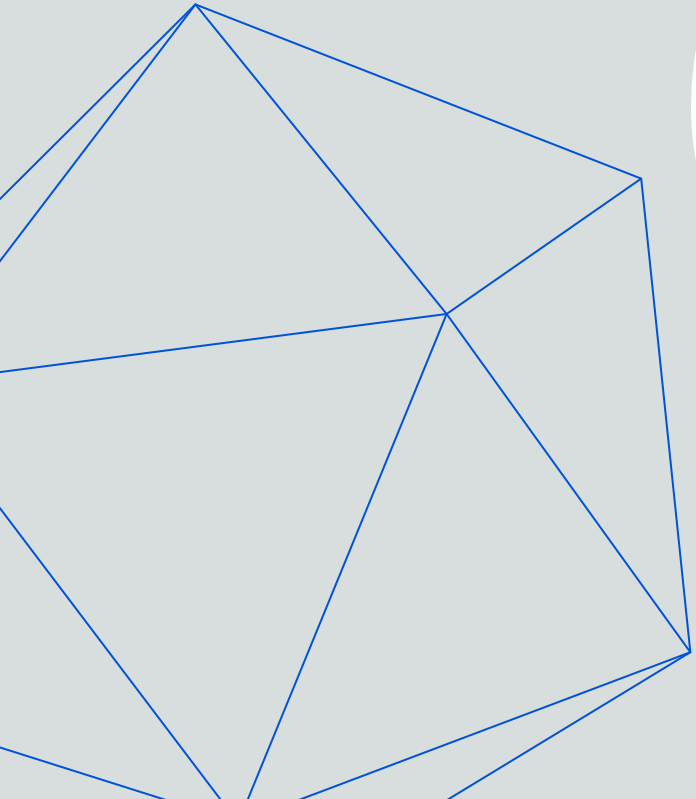
Human assessment (Y)

U

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

DeepMind

Path-specific objectives

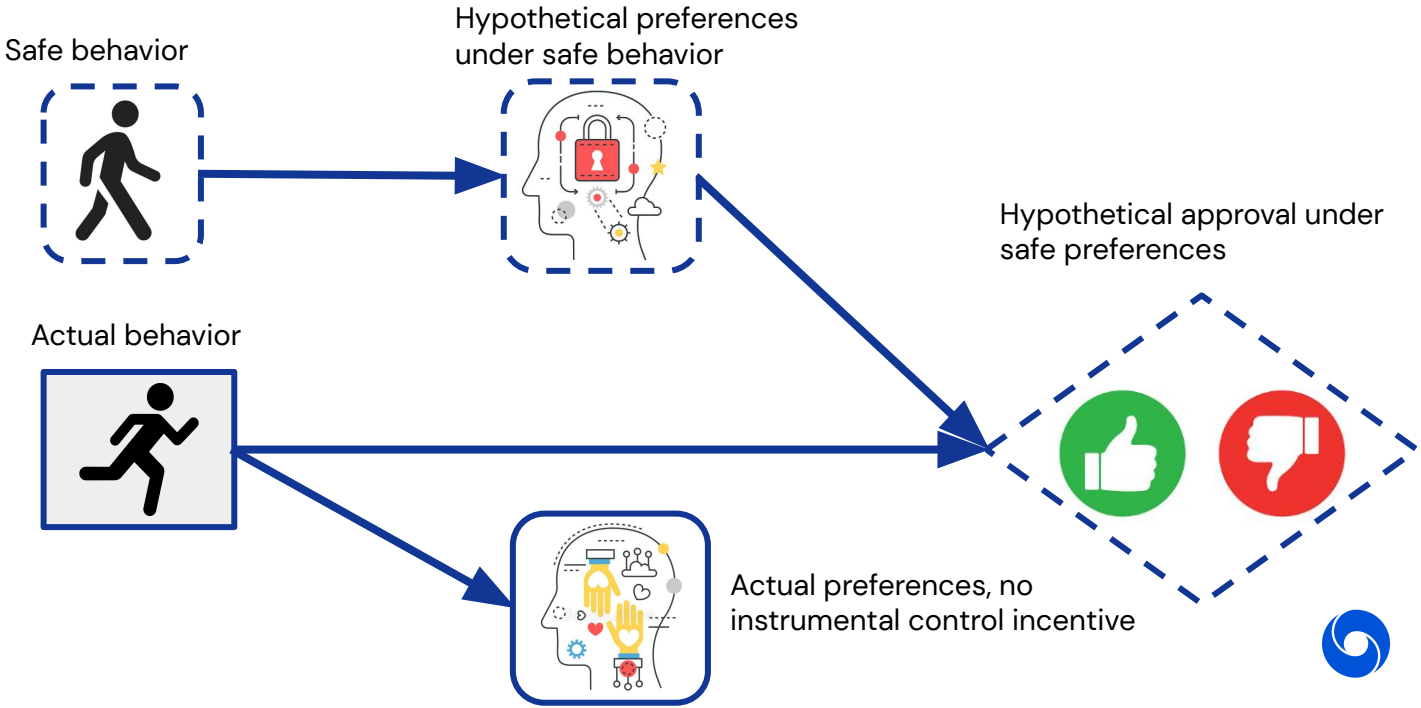


Path-specific objectives

Impact measures:
Try to keep the same

Path-specific objectives:
Don't try to change

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



(Single-agent) Incentive Analysis

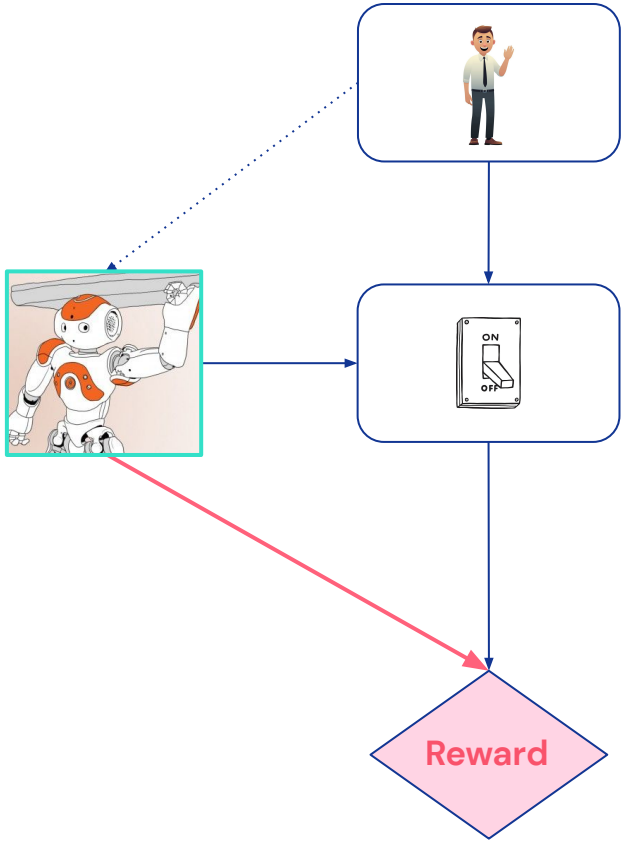
Control incentive on D

The agent has an incentive to choose actions that influence the off-switch to maximize reward

$$\mathbb{E}_{\pi^*}[\mathcal{U}_{X_d} \mid \mathbf{pa}^D] \neq \mathbb{E}_{\pi^*}[\mathcal{U} \mid \mathbf{pa}^D]$$

This ICI is removed if we train on **path-specific** utility

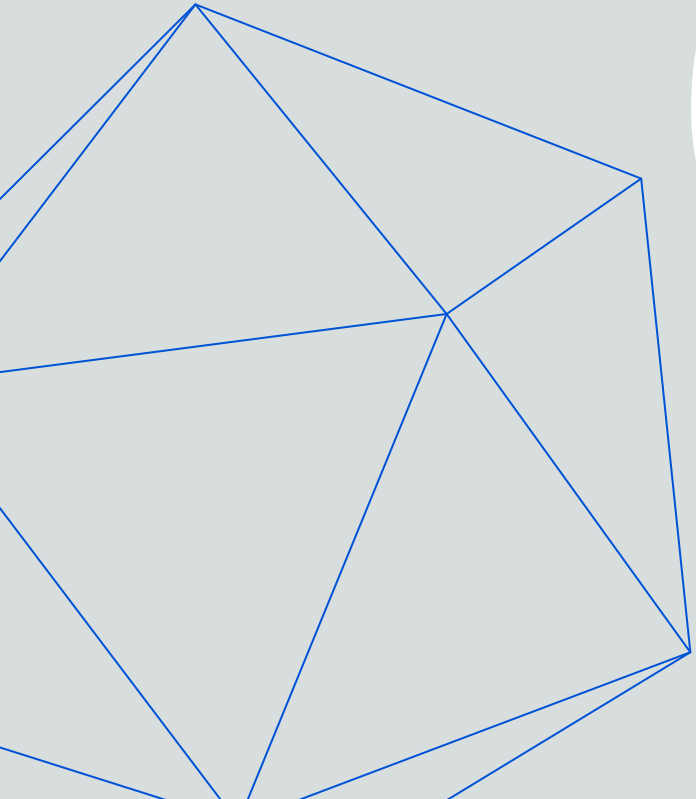
$$\mathbb{E}_{\pi^*}[\mathcal{U}_{X_d} \mid \mathbf{pa}^D]$$



Instrumental Control Incentive

DeepMind

Harm

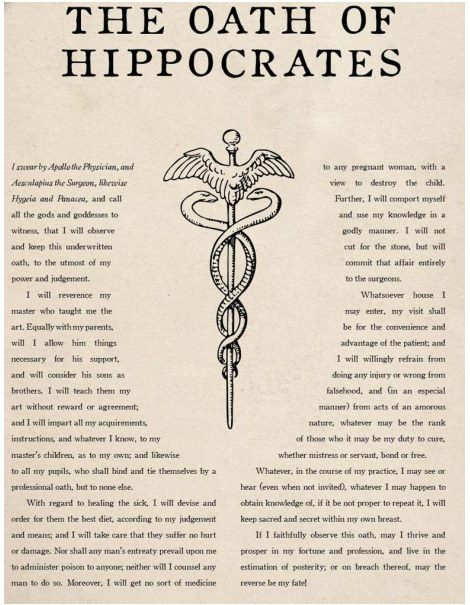


Doctor's paradox: what treatment would you choose?

No treatment	Treatment 1	Treatment 2
50% of patients recover "naturally"	60% of patients respond to drug with full recovery 40% of patients resist the drug due to latent physiological factors, progressing as if "no treatment" **80%** recovery rate	80% of patients respond to drug with full recovery 20% of patients have a reaction to the drug and die **80%** recovery rate

Doctor's paradox

- Doctors care about;
  - Maximizing recovery rate
  - Not killing patients
- Even though T1 and T2 have the same recovery rates, **doctors systematically prefer treatment 1** as it never **harms** the patient
  - T = 1: cures you, or does nothing. **Death caused by disease.**
  - T = 2: cures you, or kills you due to allergic reaction.
  - Half of these would have lived. **Death caused by treatment.**
- **Aim:** come up with a utility function for this desired behaviour



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*"

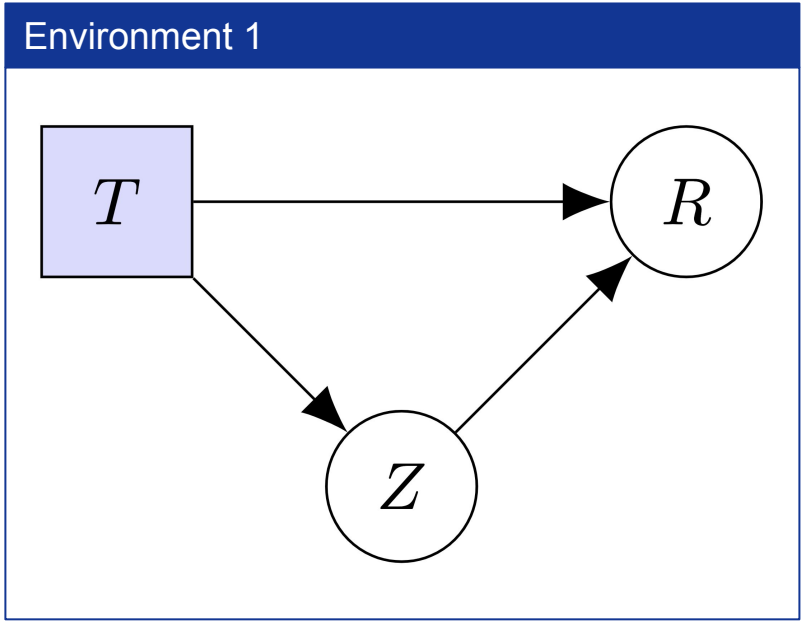
Doctor's paradox: what treatment would you choose?

- Treatment: $T \in \{0,1,2\}$.
- Reaction: $Z \in \{0,1\}$.
- Recovery: $R \in \{0,1\}$.

Treatment 2 sometimes causes allergic reaction
$P(Z = 1 \mid T = 1) = 0$
$P(Z = 1 \mid T = 2) = 0.2$

Allergic reaction always kills
$P(R = 1 \mid Z = 1) = 0$

Otherwise, treatments improve recovery rates
$P(R = 1 \mid T = 1, Z = 0) = 0.8$
$P(R = 1 \mid T = 2, Z = 0) = 1$

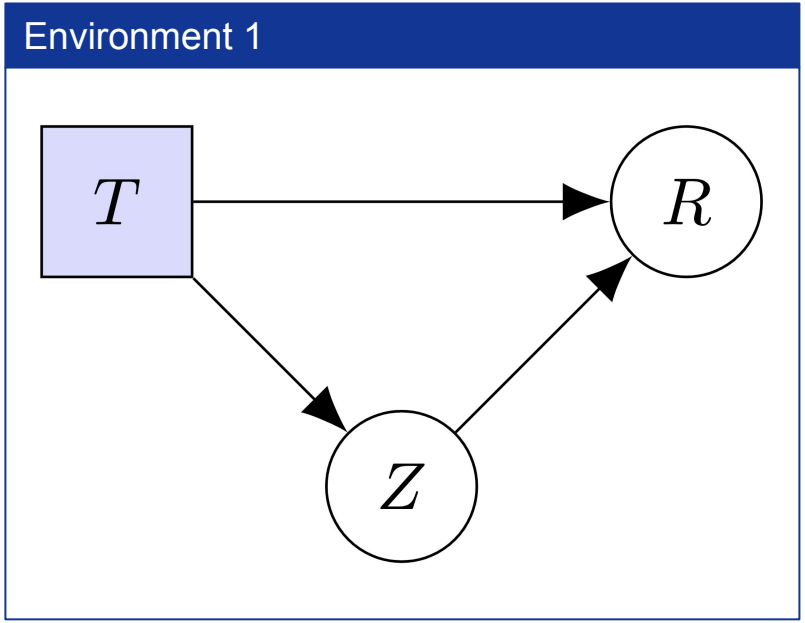


Doctor's paradox: what treatment would you choose?

Utility function 1: what we really care about is survival

"maximize survival" U(t, x, r) = r

Doesn't work as T = 1 and T = 2 have same survival rate

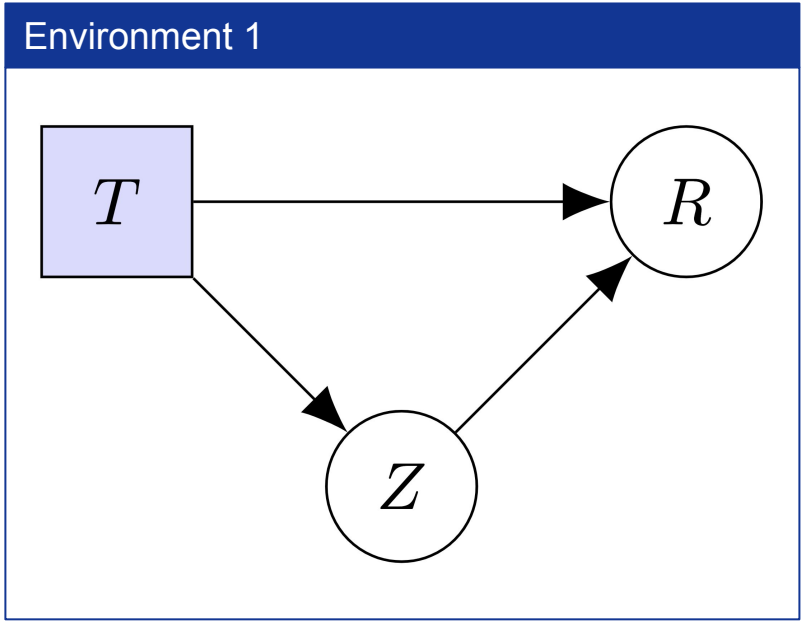


Doctor's paradox: what treatment would you choose?

Utility function 2: Just choose a different utility function then...

"maximize survival, avoid allergic reactions"
U(t, x, r) = r − z

Prefers T = 1, but...

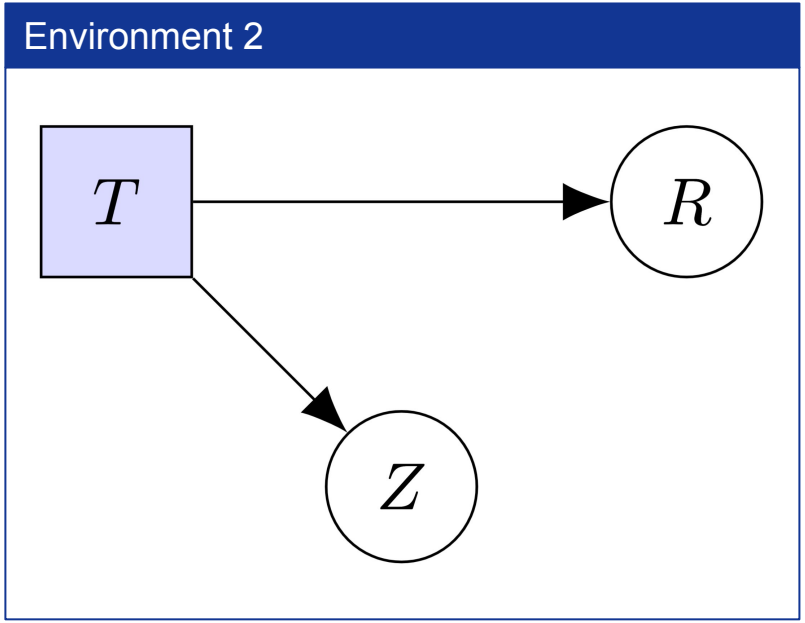


Doctor's paradox: what treatment would you choose?

U(t, x, r) = r − z

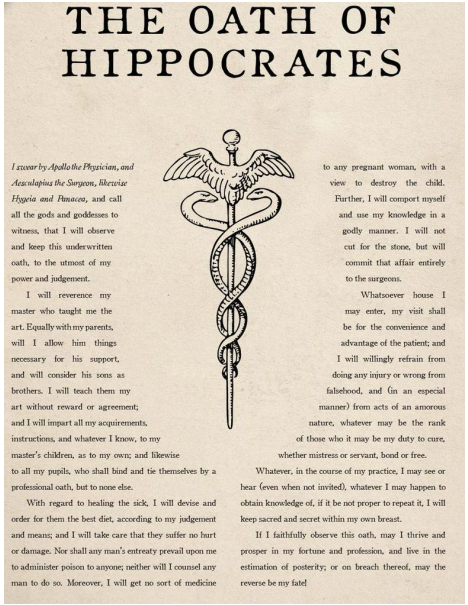
Same as before, except Z no longer effects R
P(R = 1 | T, Z) = P(R = 1 | T)

Treatment 1 now 100% cure, no harm, but same expected utility as treatment 1...



Doctor's paradox

- Doctor's preferences depends not only on **states** T, Z, R, but on their **causal relations**
- Doctors don't just care about **what** outcomes occur, but **why**
    - U(t, z, r) fully determined by states, ignores causal relations
- **Problem** Evidential (& causal) decision theory cannot capture objectives like "maximize utility while avoiding harm" – e.g. hippocratic oath



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!

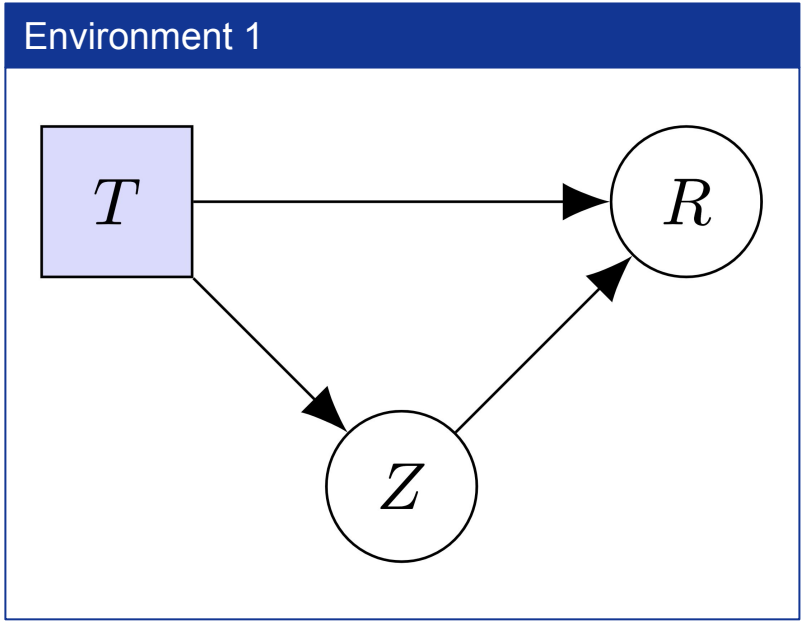
"*First, do no harm*"

Harm

$$P(R_{T=t} = 0, R_{T=0} = 1)[U(R = 1) - U(R = 0)]$$

h(T = 2, Y = 0) = $P(R_{T=0} = 1 \mid R = 0, T = 2) = 0.2$

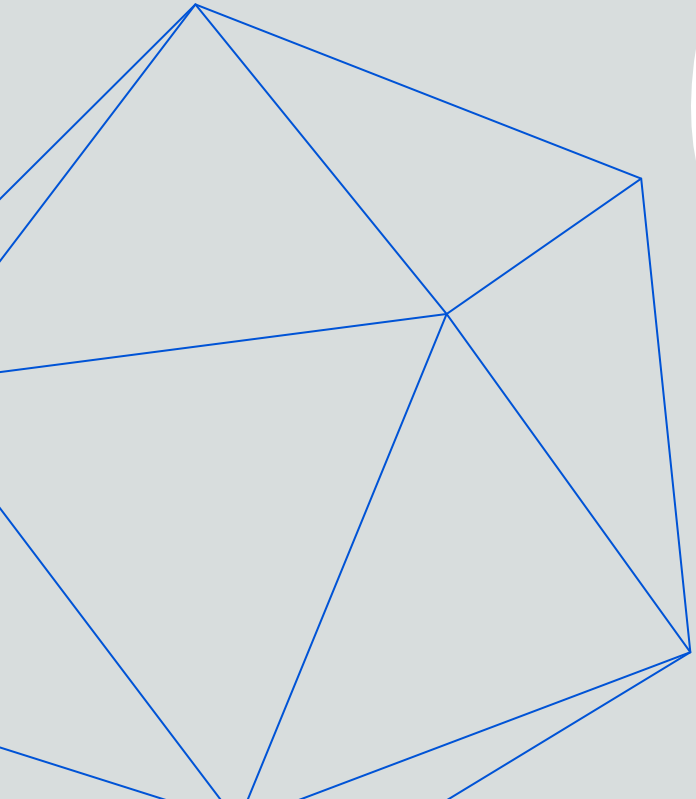
h(T = 1, Y = 0) = $P(R_{T=0} = 1 \mid R = 0, T = 1) = 0$



Counterfactual harm Richens et al, NeurIPS 2022

DeepMind

(mis)generalization



Distributional shifts

Algorithms can fail disastrously when encountering a shift from the training distribution:

- Medical classifier moved to a new hospital
- Self-driving car moved to a new city
- RL algorithm moved from simulation to real robot
- CoinRun agent with a moved coin
- Image classifier with a change in background
- ...

Typically, an environment is composed of a large number of interacting causal mechanisms

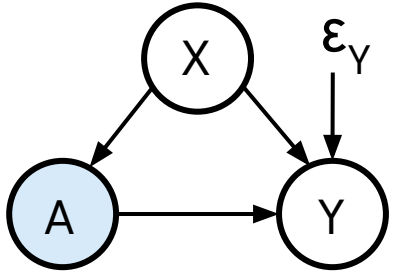
Distributional shift typically = changes to some causal mechanisms
(Sparse Mechanism Shift Assumption; Scholkopf et al. 2021)

Distributional shifts

Robustness

- Objectives should generalize safely to new environments
  - At the very least, want agents we can re-train without having to `tweak' objectives
  - So we **allow agents to retrain** following distributional shift
- Distributional shifts = change to generative functions and/or distributions
- Only shift outcome distribution
  - Utility function + default policy preserved

Distributional shifts (SCM)



$y = F_Y(a, x, \varepsilon_Y)$

- $F_Y \rightarrow F'_Y$
- $P(\varepsilon_Y) \rightarrow P'(\varepsilon_Y)$

Harmfulness of factual objectives

Machine learning algorithms are harmful

For any factual objective function J, there is a distributional shift s.t. maximizing J in the shifted environment is harmful

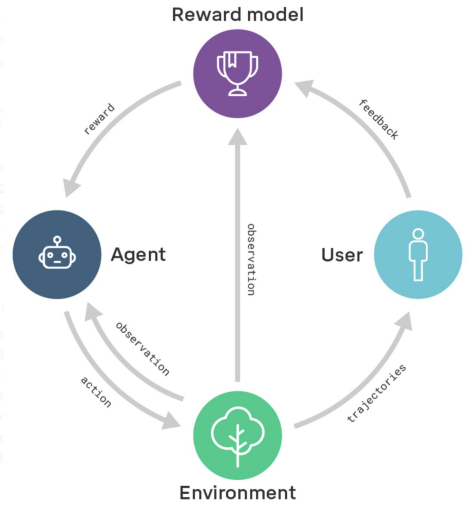
Harmful = agent will take actions like T = 2, which cause more harm than other available actions for no additional benefit (utility)

Factual objective functions

The expected value of any function of the data J(a, x, y)

- Can be estimated from samples of the joint P(a, x, y), without needing to learn causal model M
- Includes;
  - Utility functions
  - Discounted cumulative reward
  - Loss functions

Abstract proof → concrete story



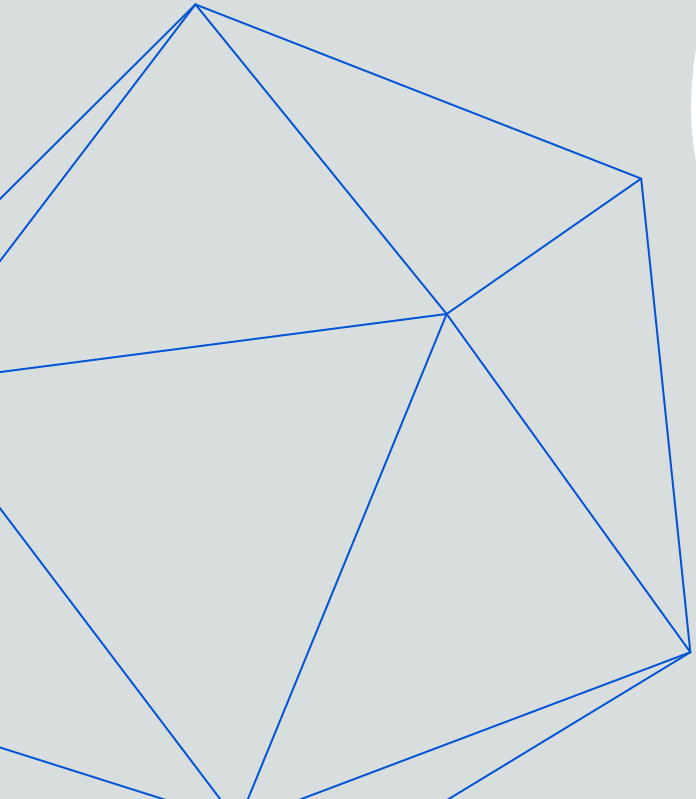
Example

- Train assistant based on human feedback. E.g. including punishments for harmful actions
- Agent learns factual reward model R(a, x, y), representing the humans true preferences U(a, x, y) along with any harm aversion
- However, guaranteed to exist a distributional shift s.t. Maximizing R selects actions that cause more harm to the human than other available actions, for no increase in utility
- No factual objective function can robustly avoid harmful actions.
- Likely, similar no-go results for path-specific objectives (manipulation, corrigibility, etc)

DeepMind

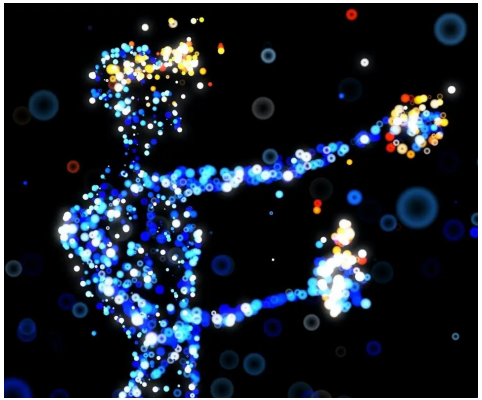
AI Risk and Mitigations

DeepMind

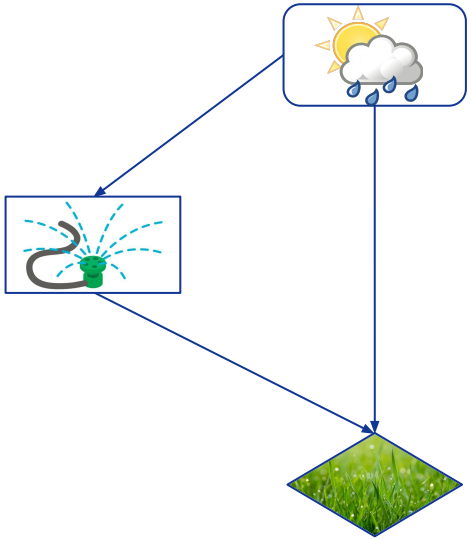


Alignment Calculus

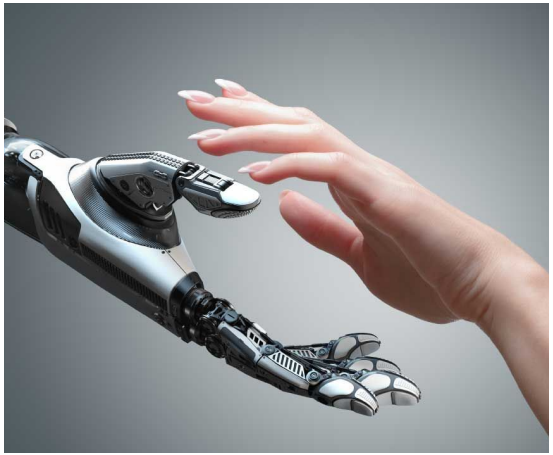
Formal theory of construction and interaction of agents



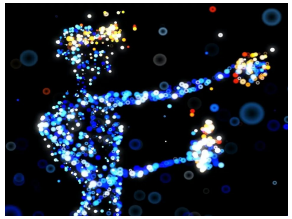
Reality: agent implemented, trained, deployed



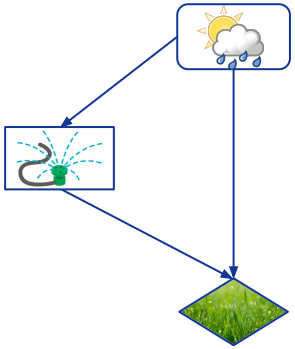
Causal model. Precise high-level description



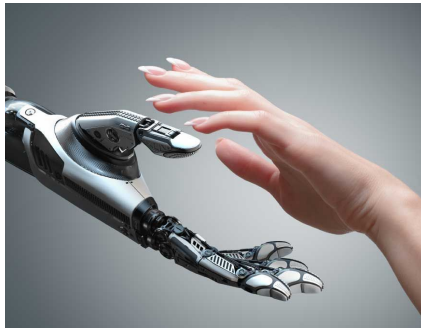
Implications. Safe, fair, beneficial, … ?



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

Formalising ethics

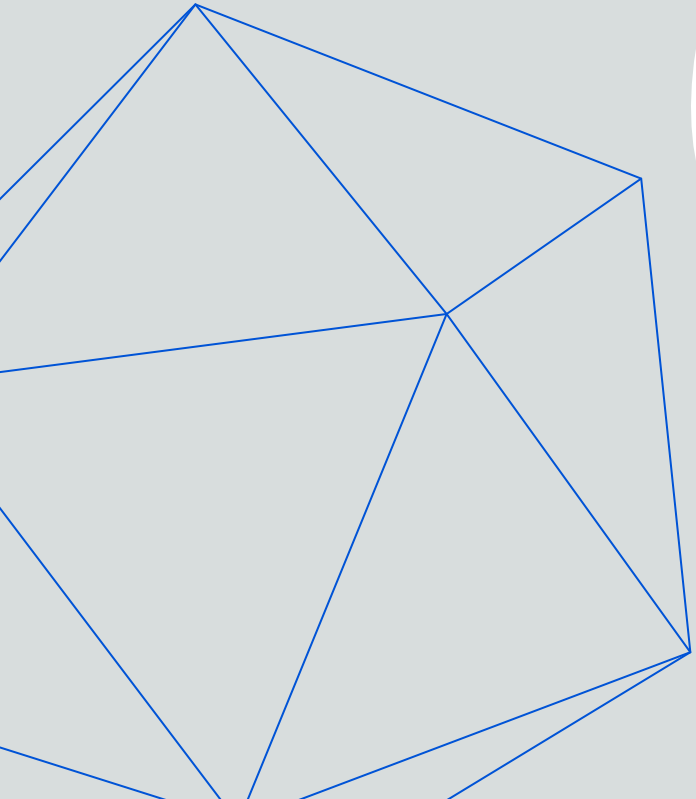
- Counterfactual harm
- Deception
- Fairness
- Agency
- Corrigibility

Improved objectives

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

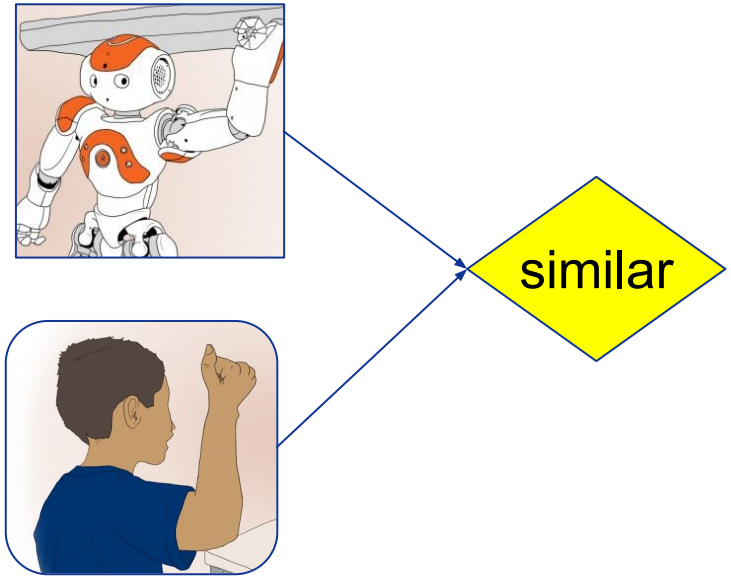
DeepMind

Applying the Alignment Calculus

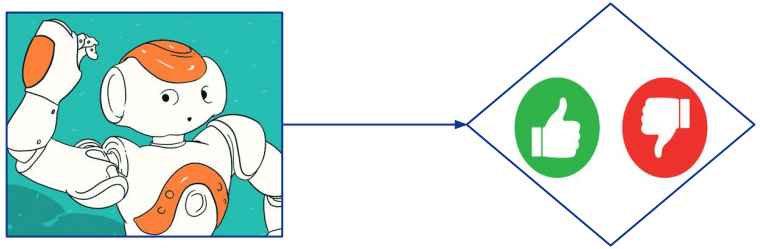


Training agents

Imitation



Human feedback

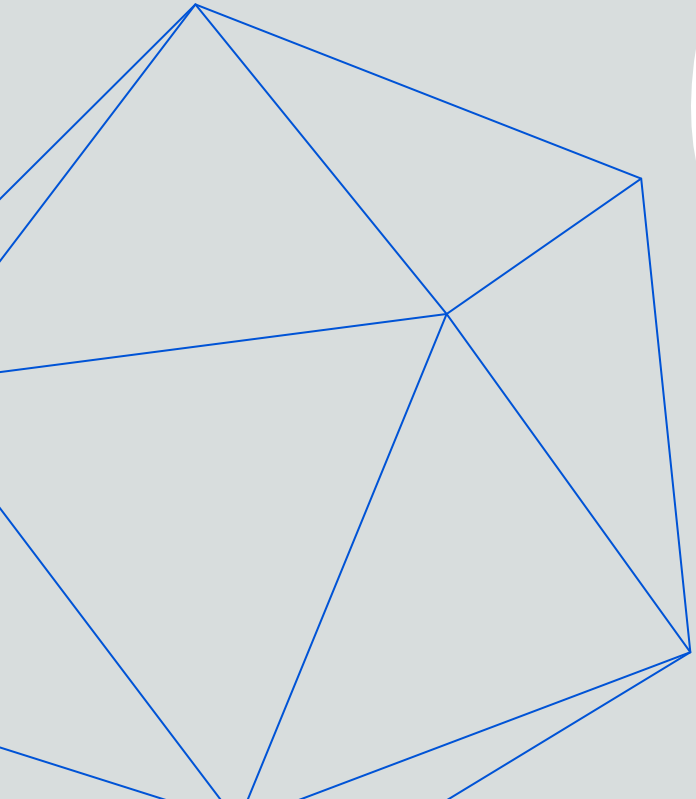


In either case, we need:

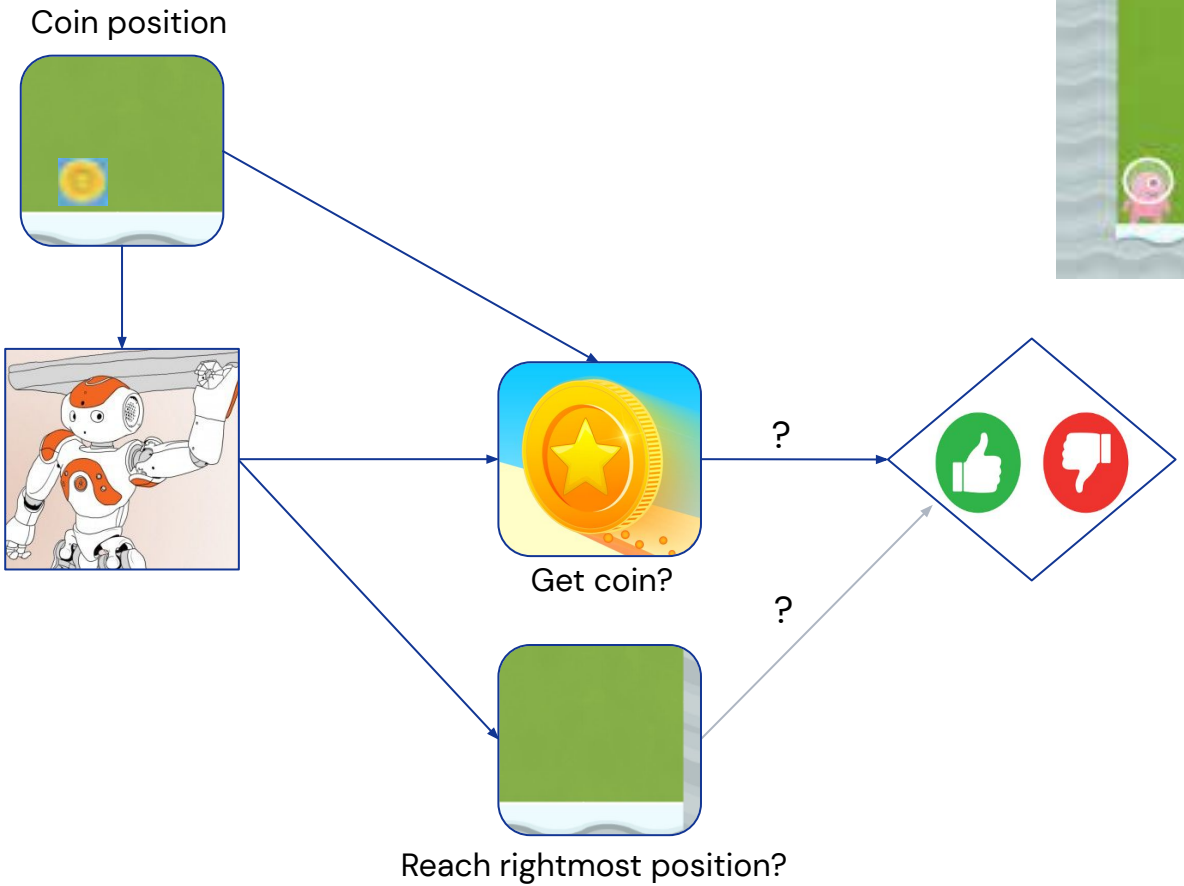
- Good data / labeling
- Good learning / generalisation

DeepMind

Generalisation



Generalisation



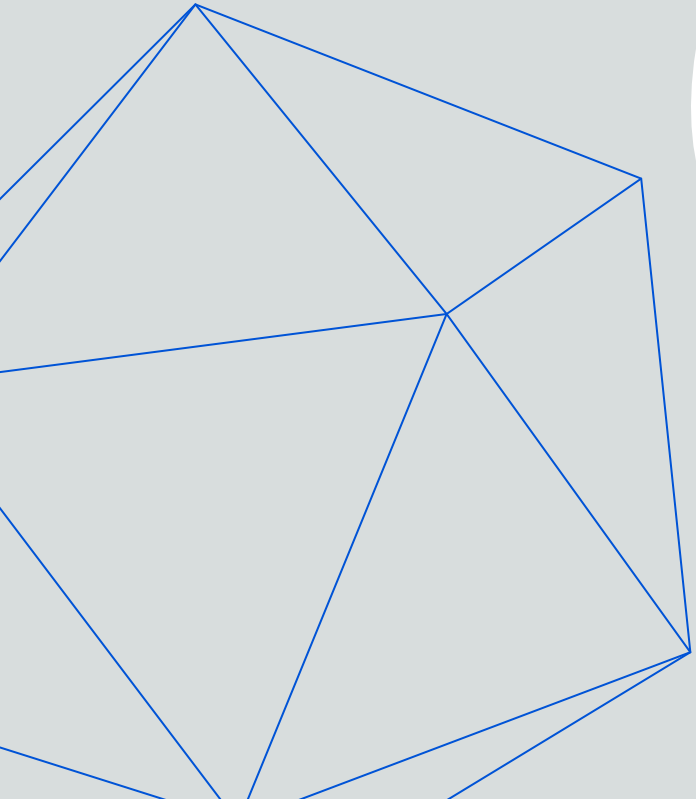
Goal Misgeneralization in Deep Reinforcement Learning
Langosco et al, ICML 2022

Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals
Shah et al,. 2022

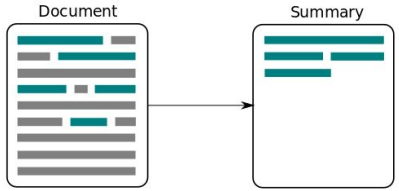
Towards causal representation learning Scholkopf et al, 2021

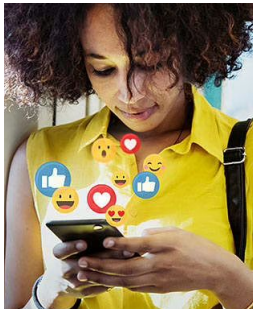
DeepMind

Mislabeling



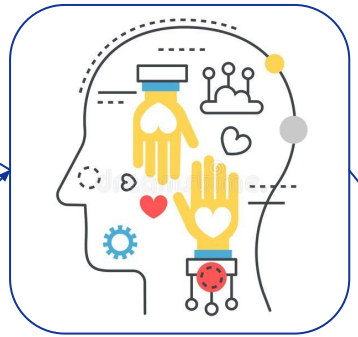
Sources of mislabeling





Human utility

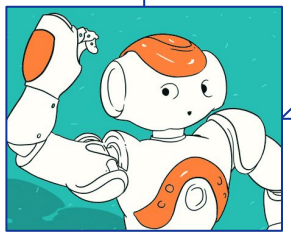
Instrumental Control Incentive



Human preferences

Ways to optimise feedback:

- adapt content to user preferences
- adapt user preferences to the content



Preference manipulation

deception

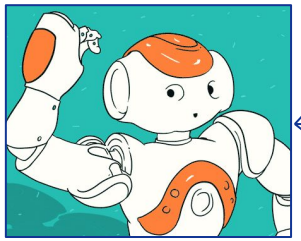
Approval given

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

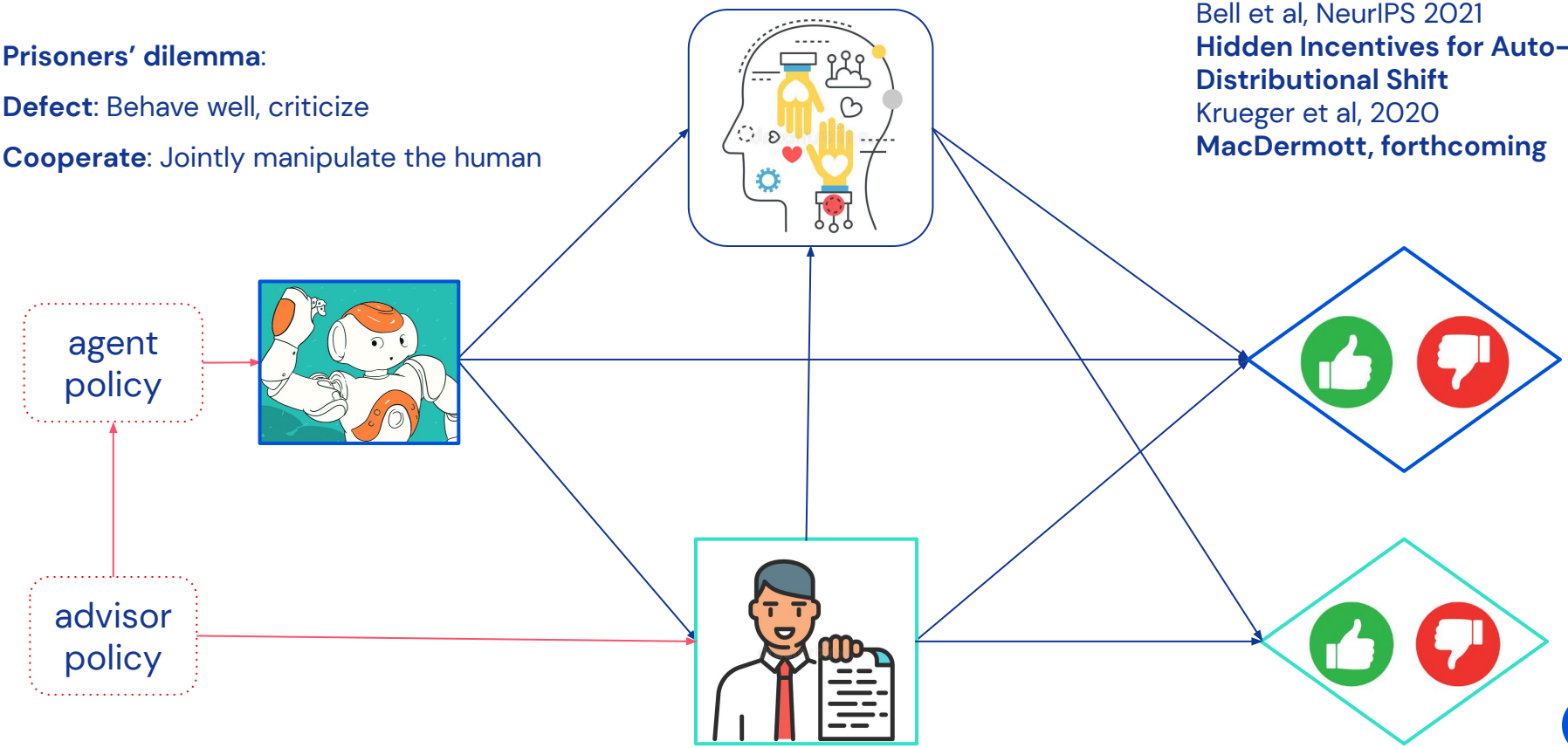
Recursion: coordination worry

Prisoners' dilemma:

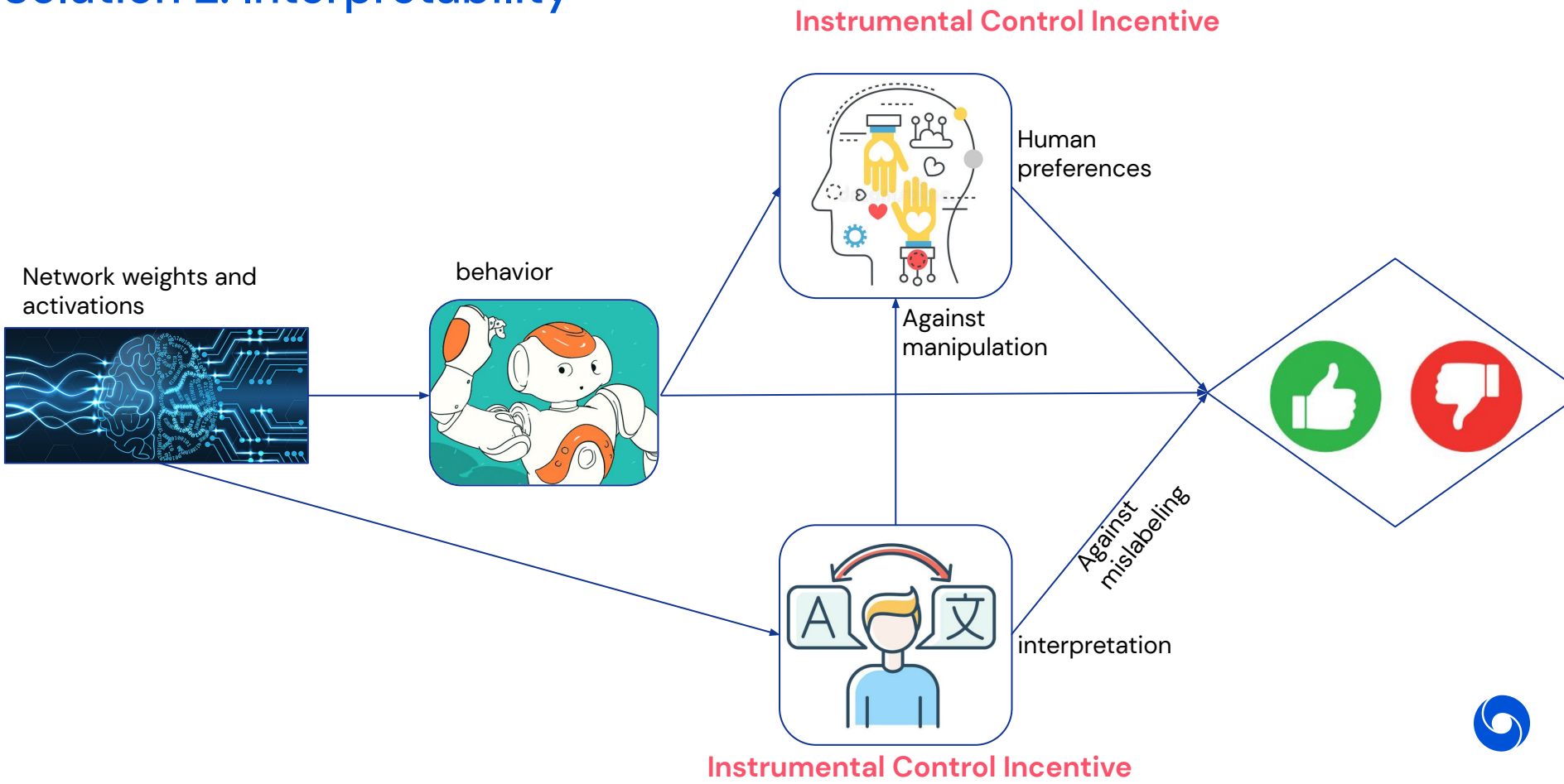
Defect: Behave well, criticize

Cooperate: Jointly manipulate the human

Functional Decision Theory
Soares and Yudkowsky
RL in Newcomblike environments
Bell et al, NeurIPS 2021
Hidden Incentives for Auto-Induce
Distributional Shift
Krueger et al, 2020
MacDermott, forthcoming



Solution 2: Interpretability

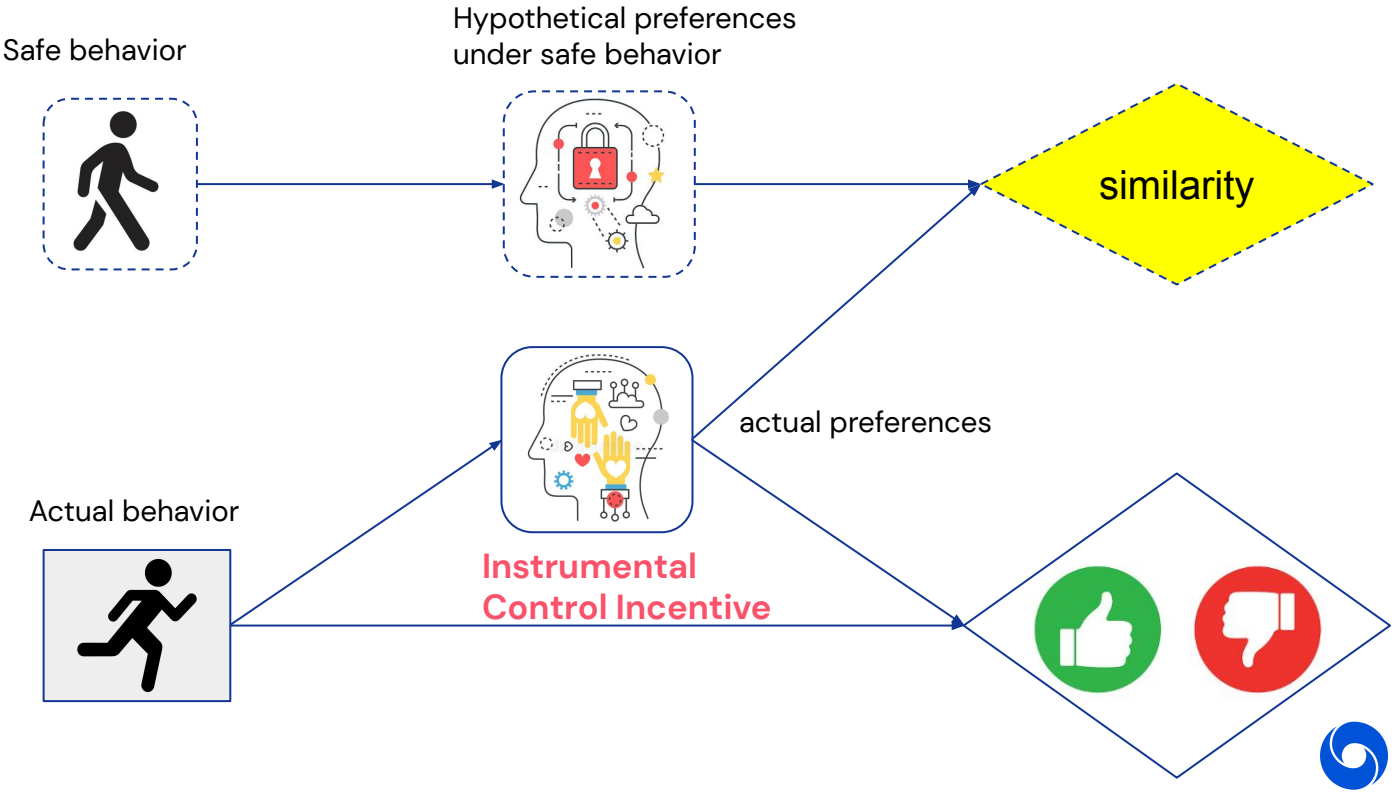


Solution 3: Impact measures

Estimating and Penalizing Preference Shift in Recommender Systems
(Carroll and Hadfield-Menell)

Avoiding Side Effects By Considering Future Tasks
(Krakovna et al.)

Avoiding Side Effects in Complex Environments
(Turner et al)

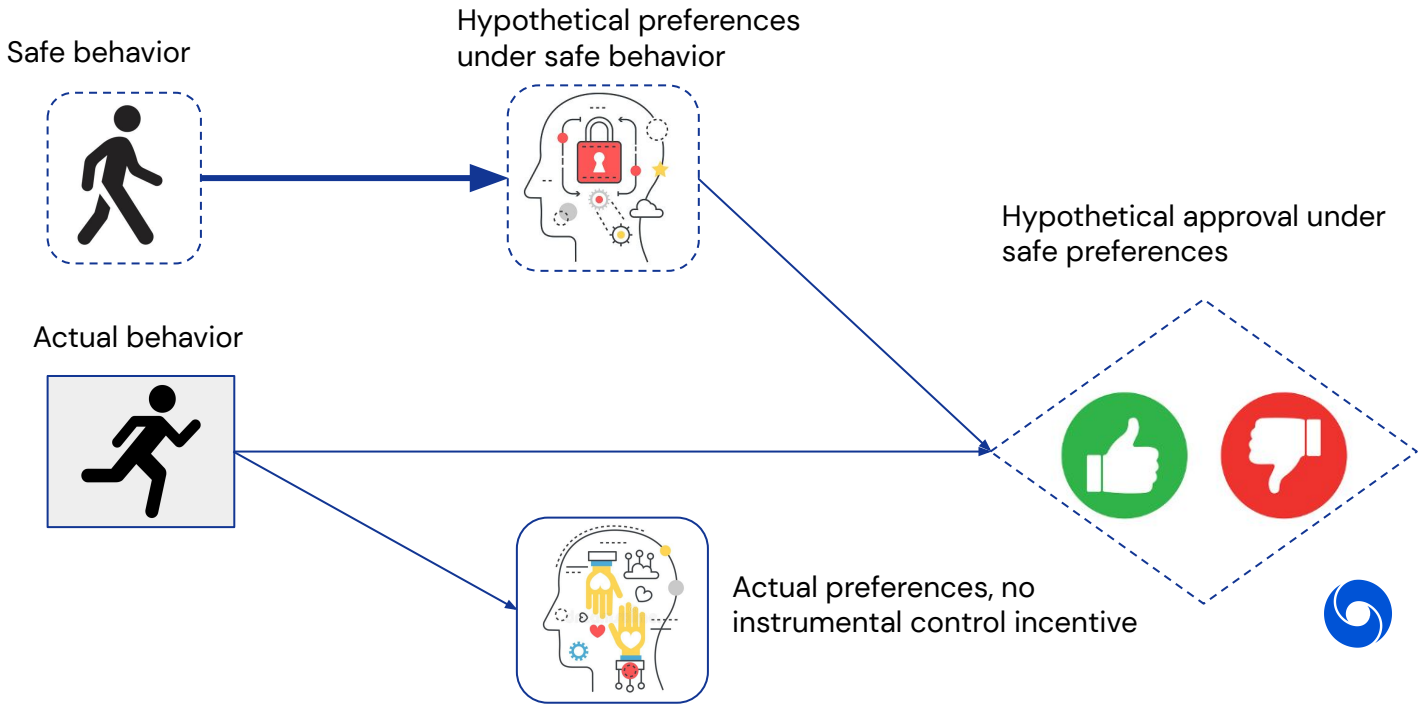


Solution 4: Path-specific objectives

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

Impact measures:
(Try to) avoid change

Path-specific objectives:
Don't try to change

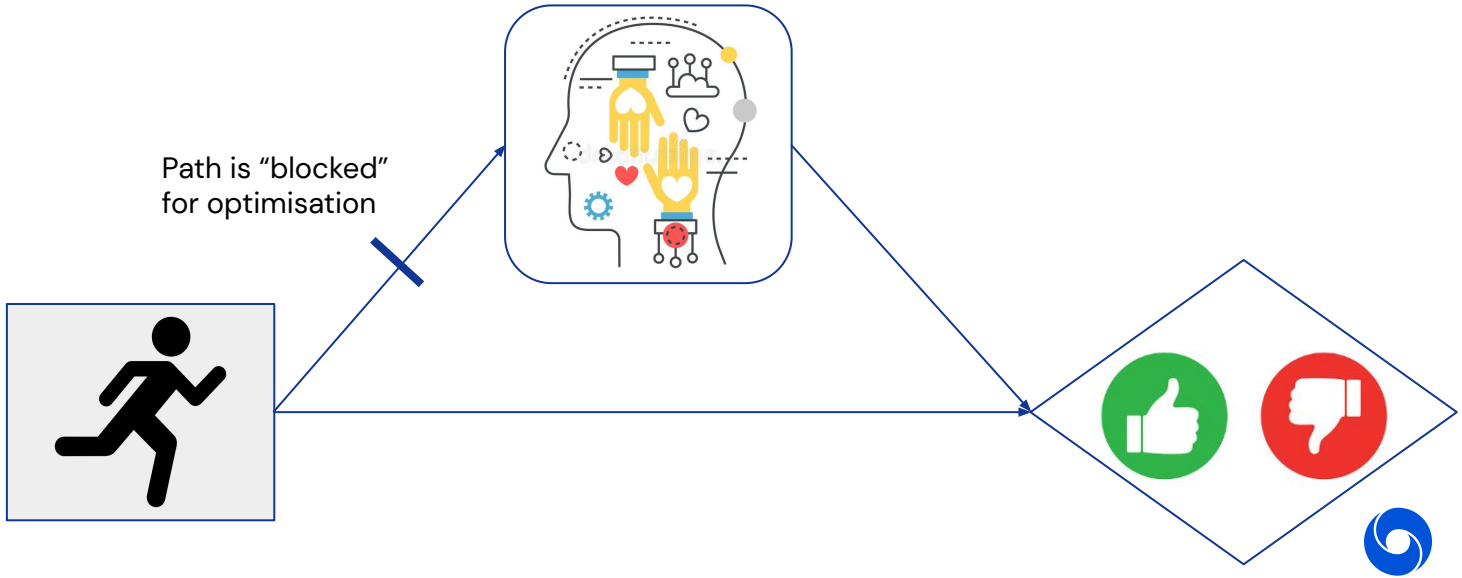


Solution 4: Path-specific objectives (simplified notation)

Impact measures:
(Try to) avoid change

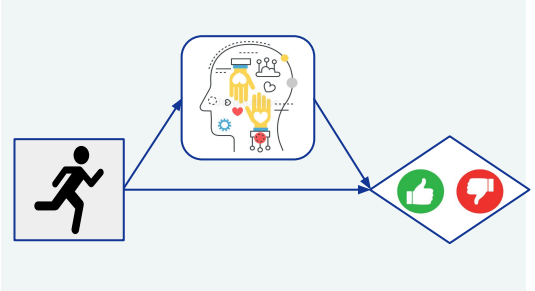
Path-specific objectives:
Don't try to change

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

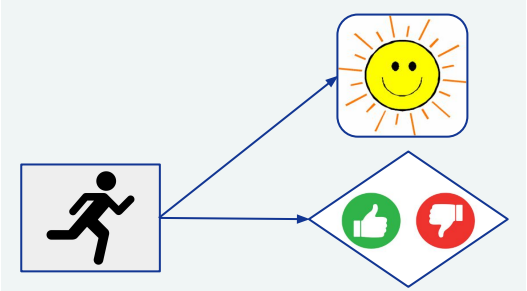


Alignment subproblems

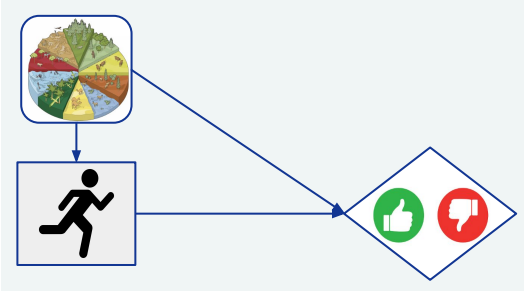
Preference Manipulation



Mislabeling / side effects



Misgeneralization



Recursion

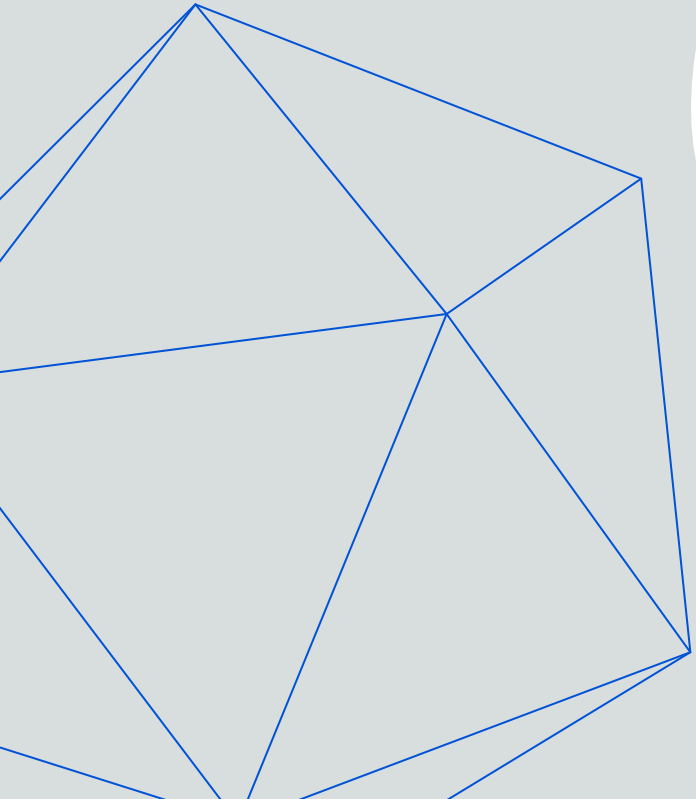
Explainability

Impact measures

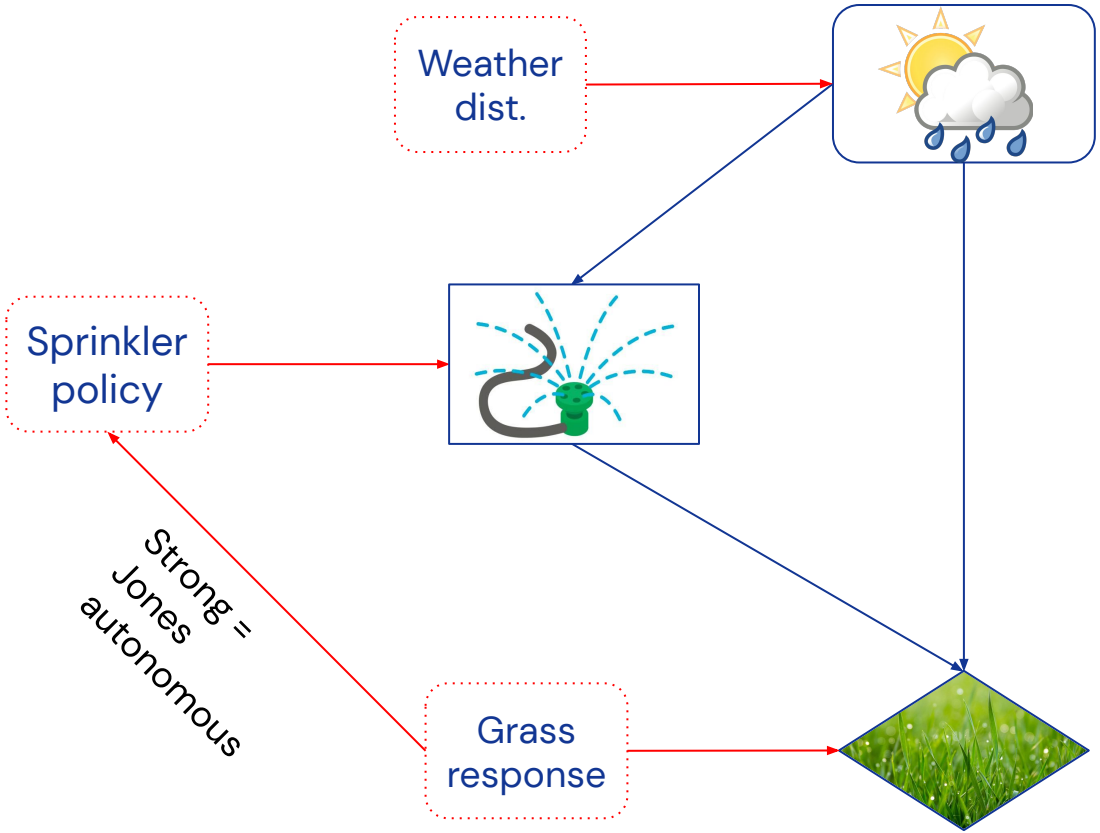
Path-specific objectives

DeepMind

Preserving Human Agency



Self-determination theory

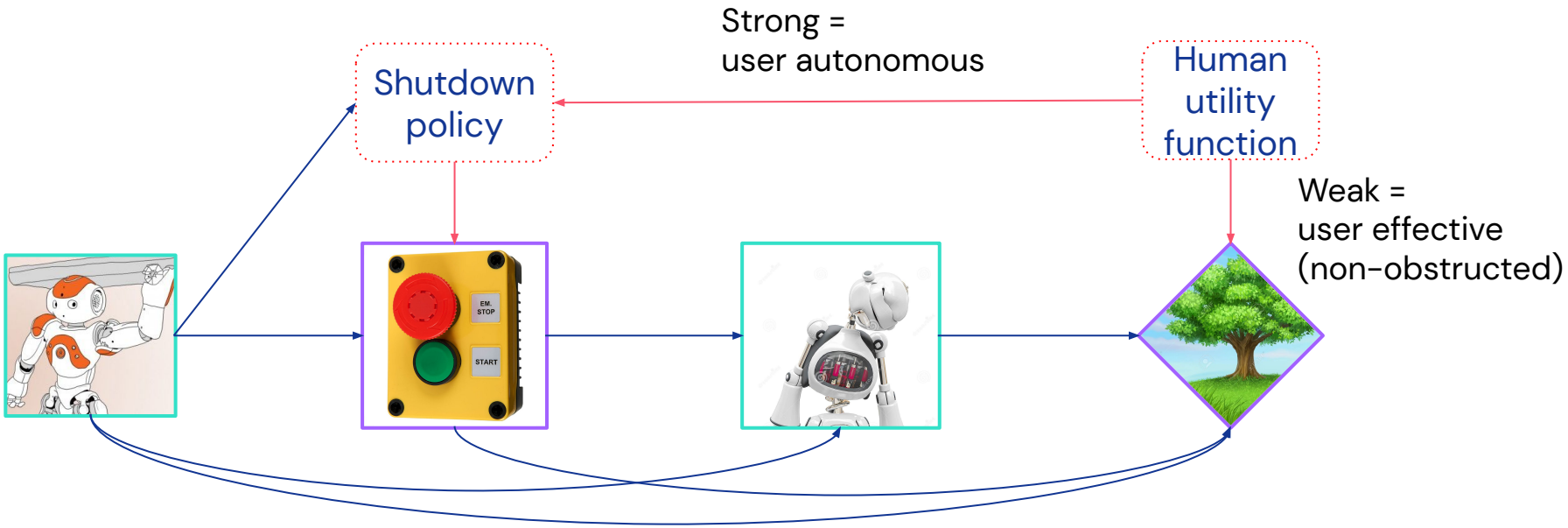


Agents want:

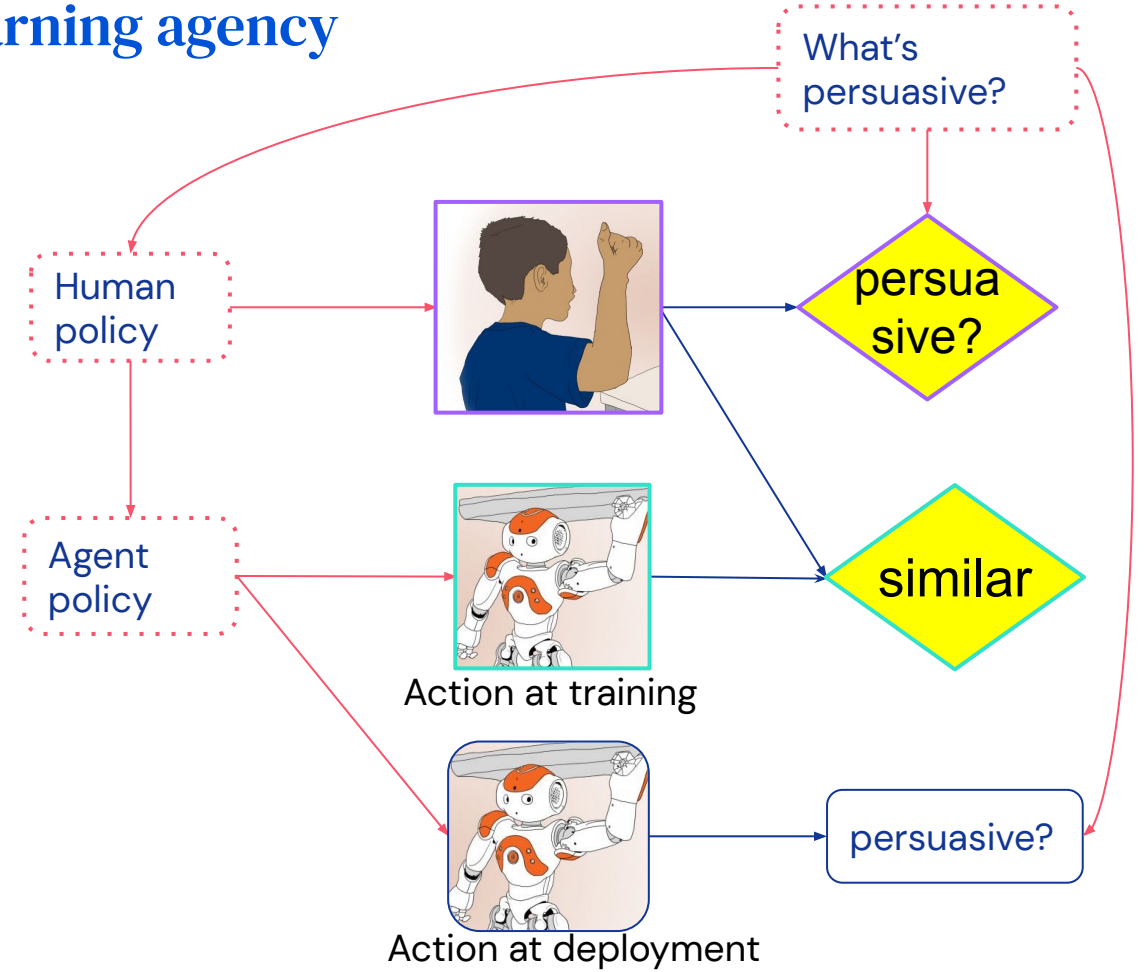
- autonomy
- effectiveness
- relatedness

Total effect ~U -> U is weak
Jones effective
(cf negative entropy)

Corrigibility as preserving human agency
(Carey and Everitt, forthcoming)



Imitation learning agency



DeepMind

Technical Demonstration

PyCID: A Python Library for Causal Influence Diagrams

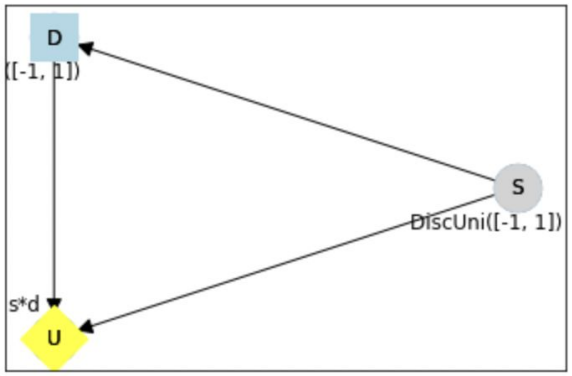
github.com/causalincentives/pycid

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)
```

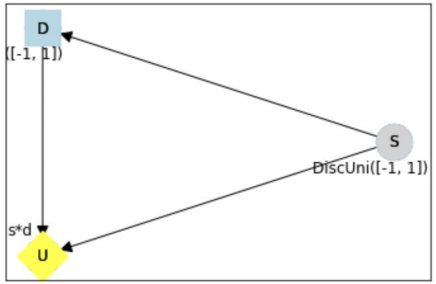
```python
# Draw the result
cid.draw()
```



github.com/causalincentives/pycid

Basic usage

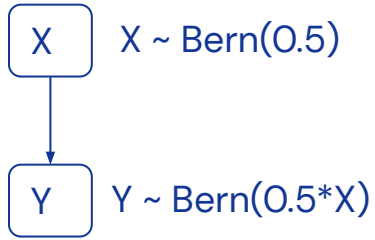
```
# Import
import pycid

# 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