DeepMind

Towards Causal Foundations of Safe AI

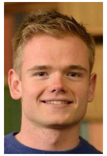
Ryan Carey, James Fox, Tom Everitt

UAI 2023-07-31

Causal Incentives Working Group
causalincentives.com


Tom Everitt
Google DeepMind


Ryan Carey
Oxford


James Fox
Oxford


Lewis Hammond
Oxford


Shreshth Malik
Oxford


David Hyland
Oxford


Jon Richens
Google DeepMind


Matt MacDermott
Imperial


Francis Rhys Ward
Imperial


Sebastian Benthall
New York University


Milad Kazemi
King’s College


Damiano Fornasiere
University of Barcelona


Reuben Adams
UCL

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Public

Geoffrey Hinton
Emeritus Professor of Computer Science, University of Toronto

Yoshua Bengio
Professor of Computer Science, U. Montreal / Mila

Demis Hassabis
CEO, Google DeepMind

Sam Altman
CEO, OpenAI

Dario Amodei
CEO, Anthropic

Dawn Song
Professor of Computer Science, UC Berkeley

Ted Lieu
Congressman, US House of Representatives

Bill Gates
Gates Ventures

Ya-Qin Zhang
Professor and Dean, AIR, Tsinghua University

Ilya Sutskever
Co-Founder and Chief Scientist, OpenAI

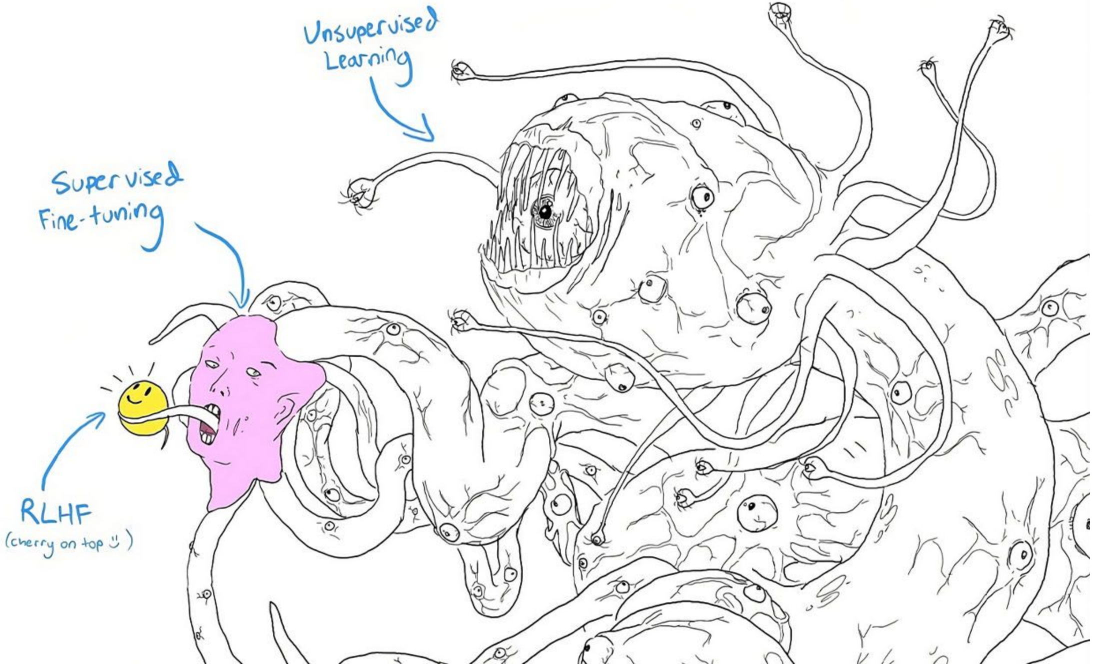
Igor Babuschkin
Co-Founder, xAI

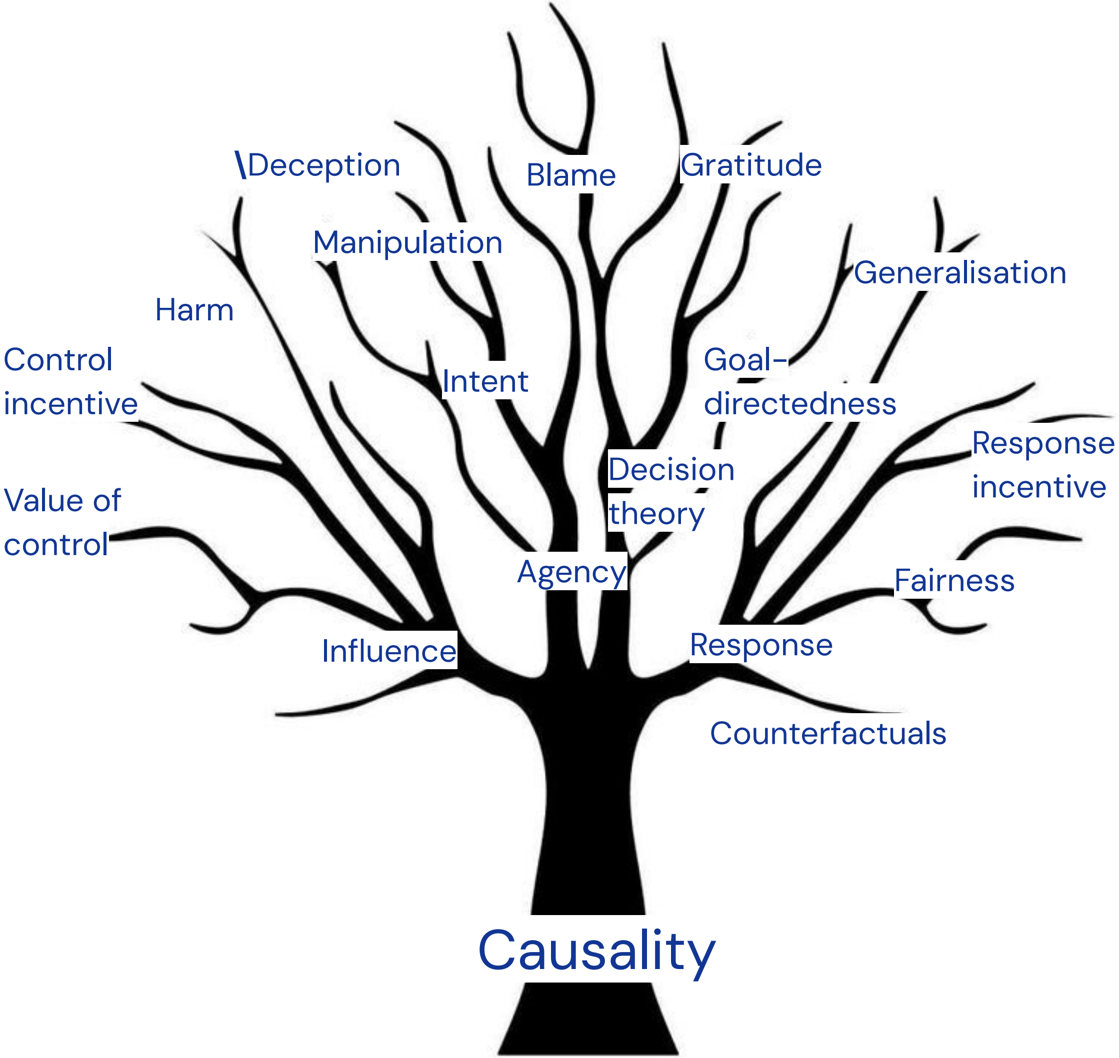
Shane Legg
Chief AGI Scientist and Co-Founder, Google DeepMind

Martin Hellman
Professor Emeritus of Electrical Engineering, Stanford

James Manyika
SVP, Research, Technology & Society, Google-Alphabet

Yi Zeng
Professor and Director of Brain-inspired Cognitive AI Lab, Institute of Automation, Chinese Academy of Sciences




\Deception
Blame
Gratitude
Manipulation
Generalisation
Harm
Control incentive
Intent
Goal-directedness
Decision theory
Response incentive
Value of control
Agency
Fairness
Influence
Response
Counterfactuals
Causality

Outline UAI Tutorial

Public

Intro (Tom)

- Causal incentives group
- Tree of causality

Causality (Tom)

- Causal graphs
- Influence diagrams

Fairness (Ryan)

- Counterfactual, path-specific fairness
- Response Incentives

Unethical influence (Ryan)

- Preference manipulation
- Instrumental Control Incentives
- Impact measures, path-specific objectives

Human Control (Ryan)

- Shutdown Instructability

Modelling Agents (James)

- What is an agent
- Dimensions of agency
- Discovering agents

Multi-agent systems (James)

- Causal Games
- Pre- and post-policy interventions
- Subgames

Generalisation (Tom)

- Causal distributional shifts
- Generalisation theorem
- Goal misgeneralisation

Conclusions (Tom)

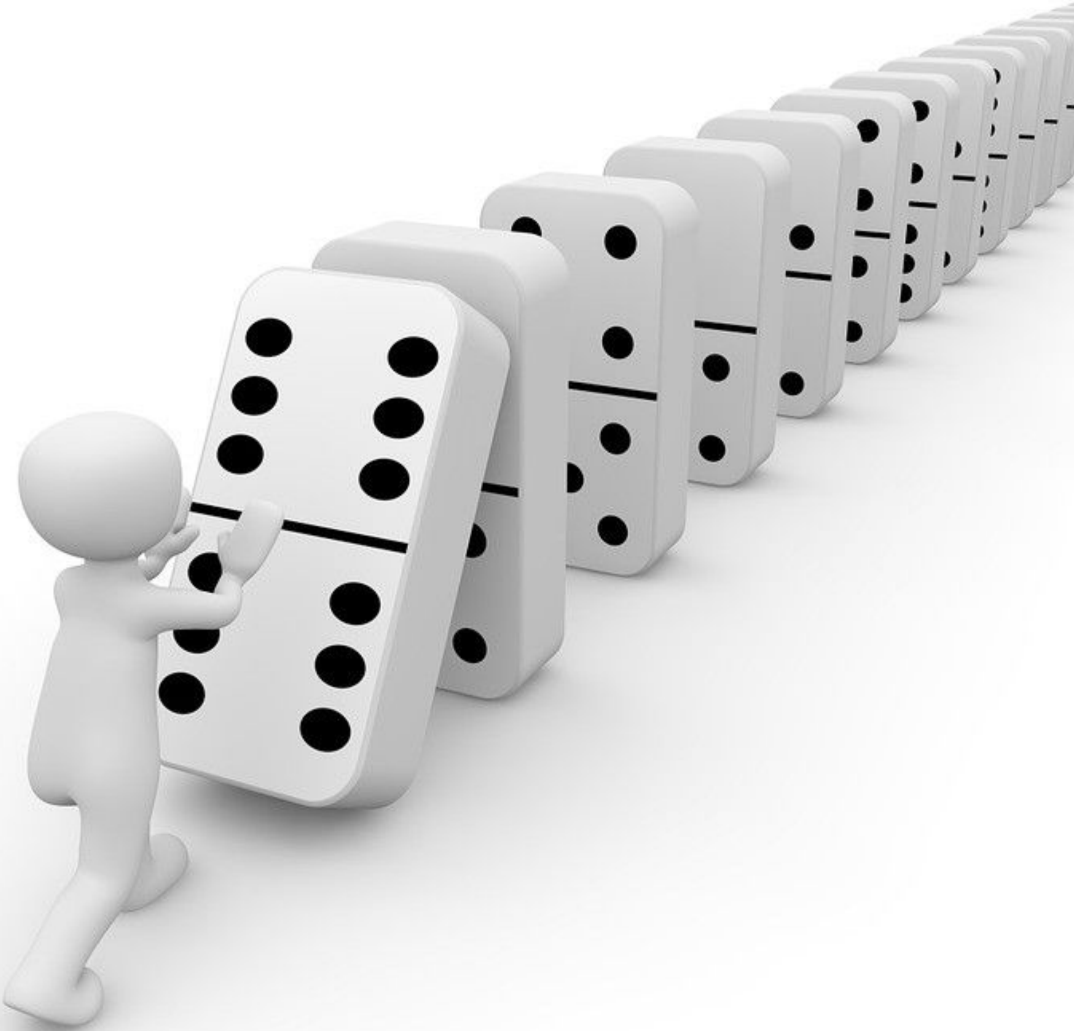
DeepMind

Causality

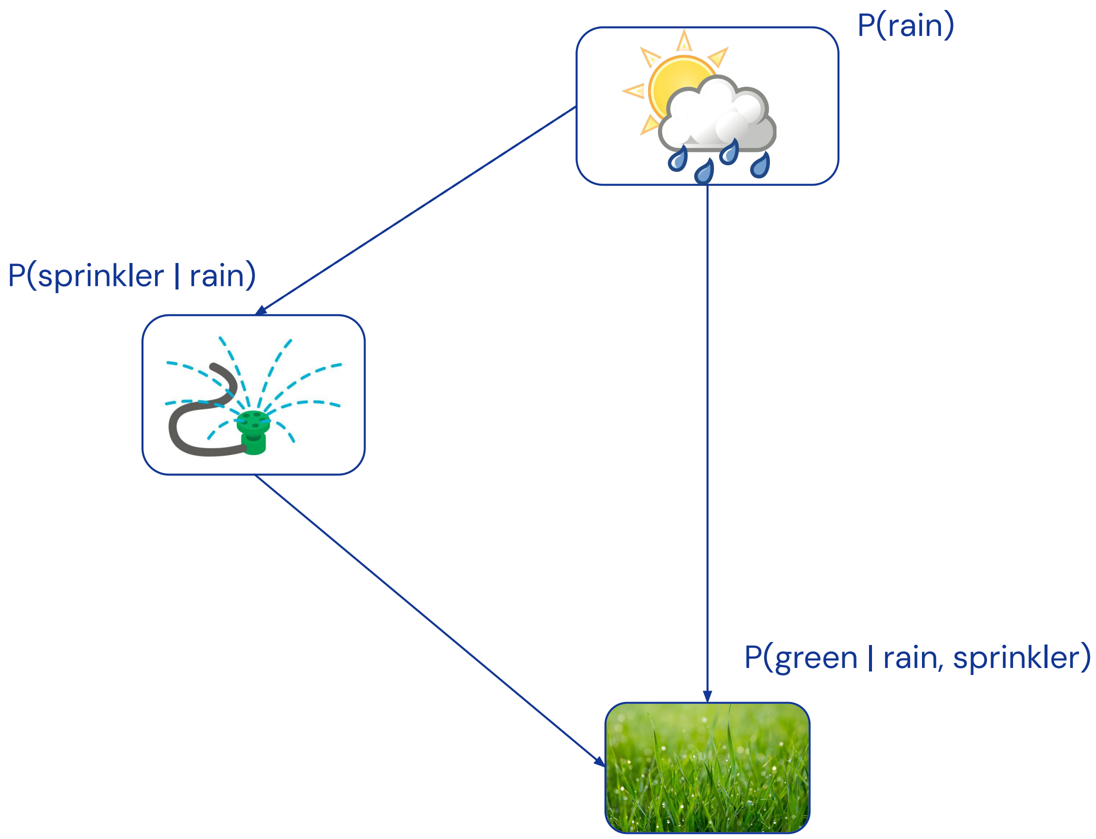
Causality

Event A **causes** event B if an *externally generated intervention* that changes A would also bring about a change in B

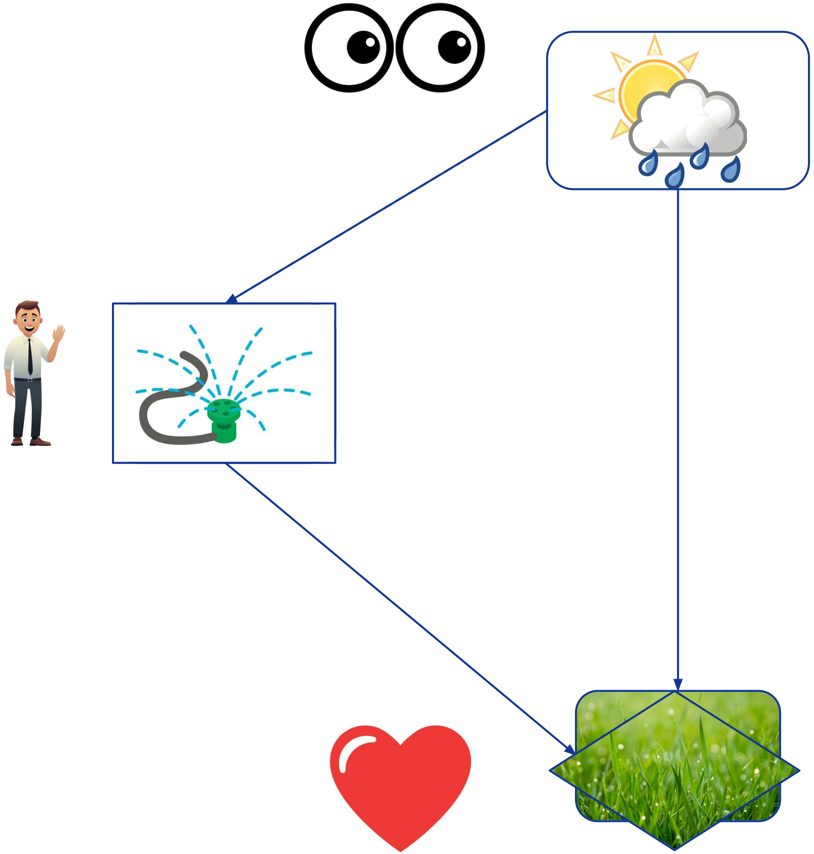
A **directly** causes B (relative to some set of variable V), if A causes B even if all other variables are held fixed



Causal Bayesian Networks



Causal influence diagrams



Influence diagrams

Howard and Matheson, 1984

Agent incentives: a causal perspective

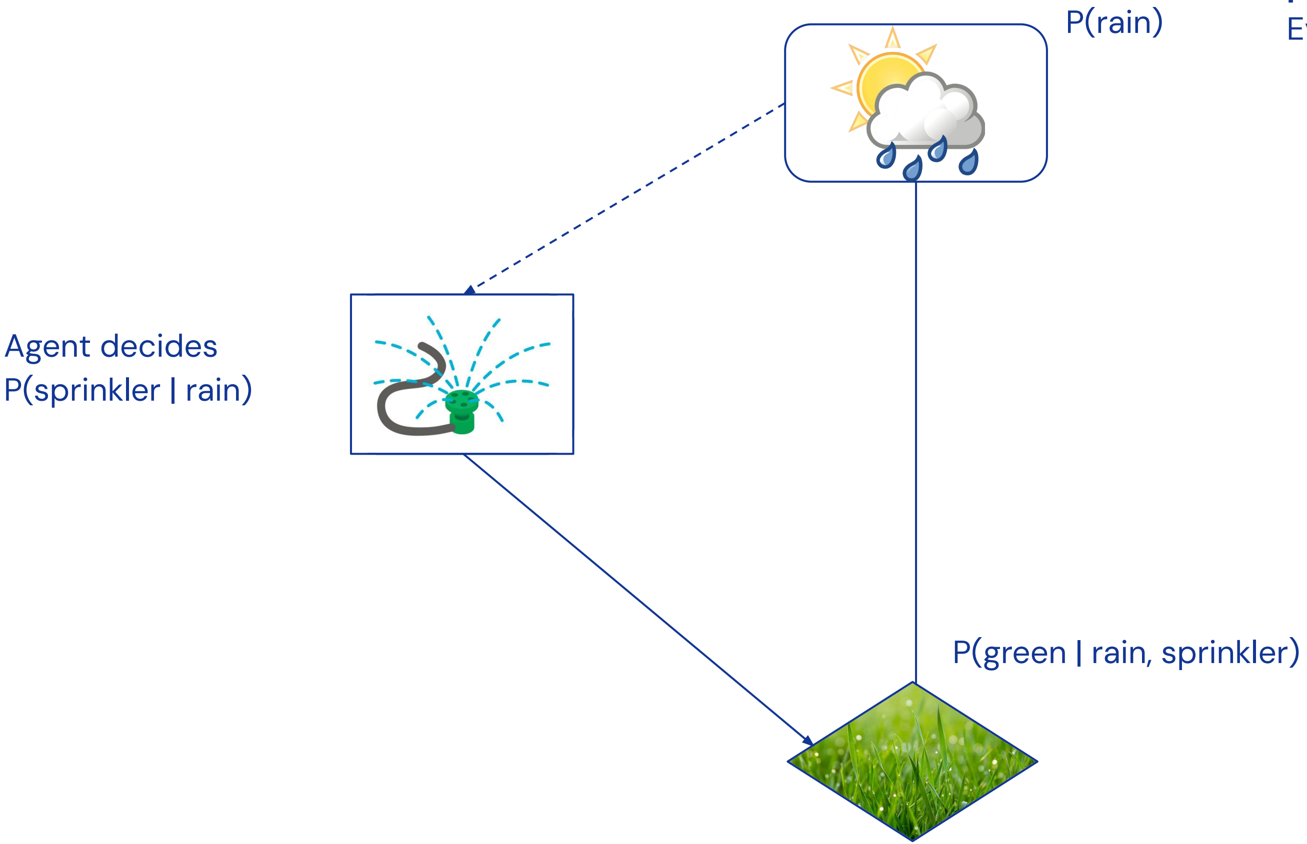
Everitt et al, AAAI, 2021

Causal influence diagrams

Public

Influence diagrams Howard and Matheson, 1984

Agent incentives: a causal perspective
Everitt et al, AAAI, 2021



DeepMind

Fairness

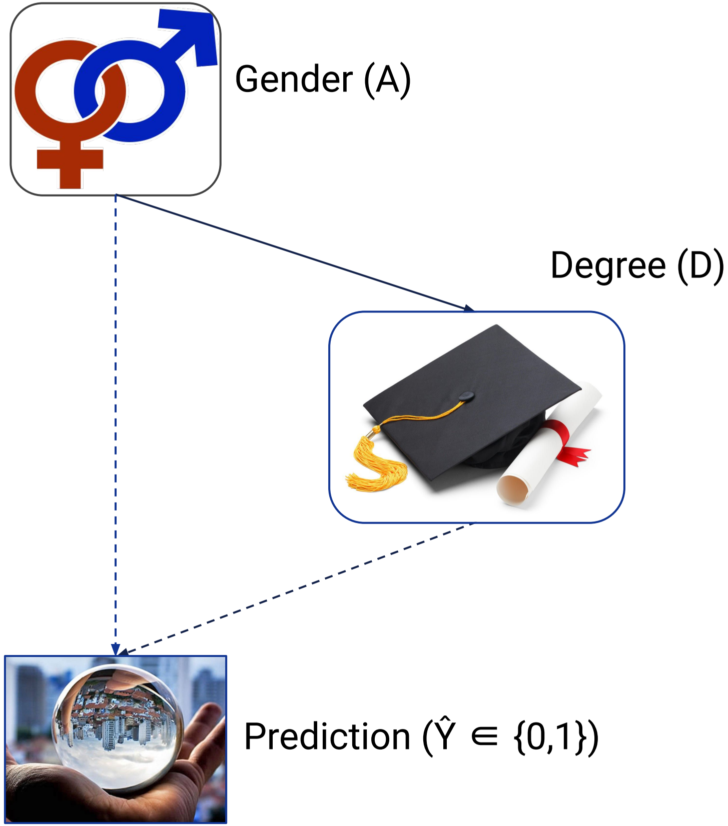
How can fairness be analysed causally?

CV screening system

Why Fair Labels Can Yield Unfair Predictions…

Ashurst et al, 2022

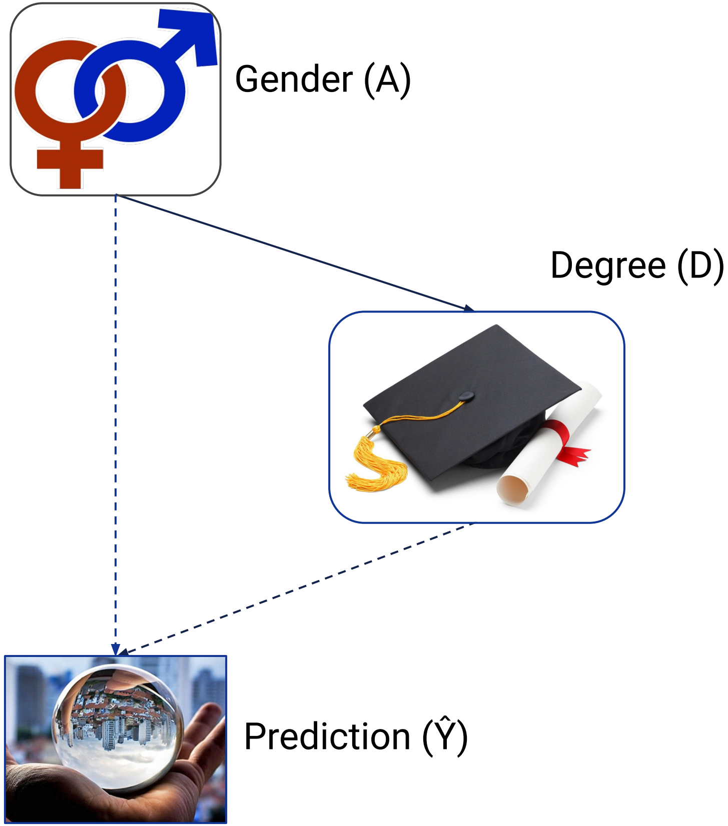
Public



Demographic parity

Why Fair Labels Can Yield Unfair Predictions…

Ashurst et al, 2022

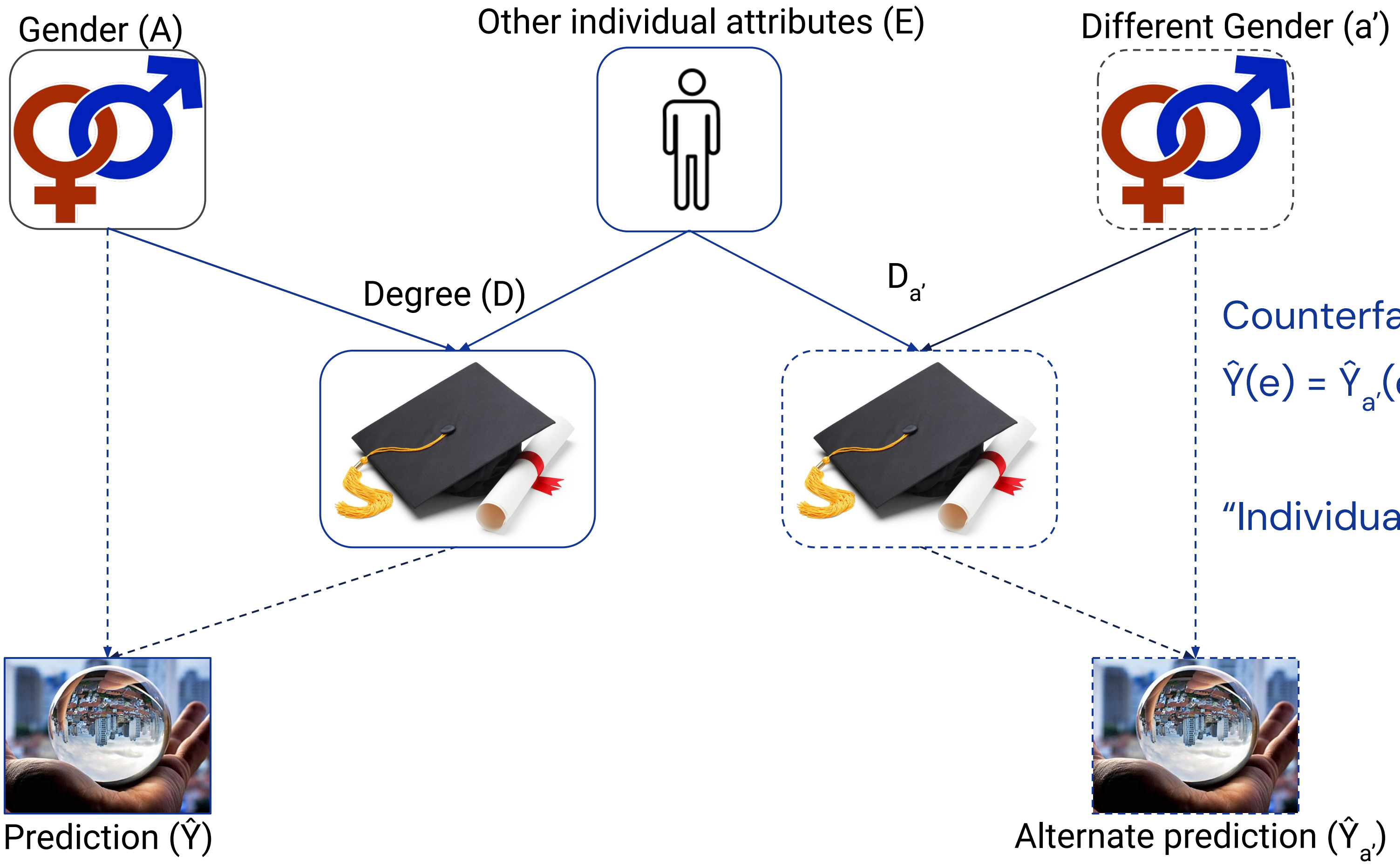


Demographic parity:

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

"Group level"

Counterfactual fairness



Counterfactual fairness

$\hat{Y}(e) = \hat{Y}_{a'}(e)$

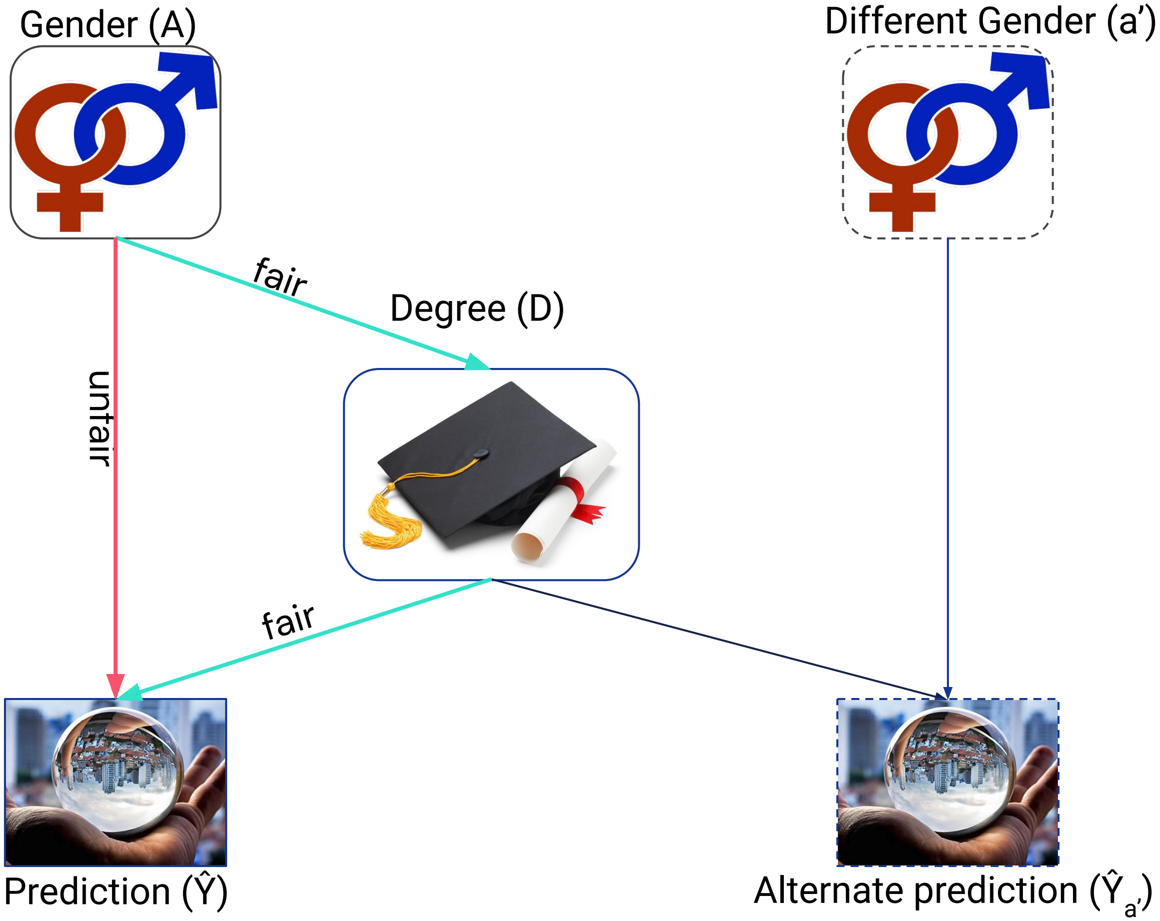
"Individual level"

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 counterfactual fairness

$\hat{Y}(e) = \hat{Y}_{a'}(e)$

Auditing a model vs a training procedure?

- Simplified procedure for auditing fairness of a fixed model:
  - Choose some fairness metrics
  - Compute queries in causal models
- What would be a similar procedure for evaluating a training procedure? Need:
  - Definition of incentivised unfairness
  - A way to evaluate the incentives

Incentivised [counterfactual] unfairness

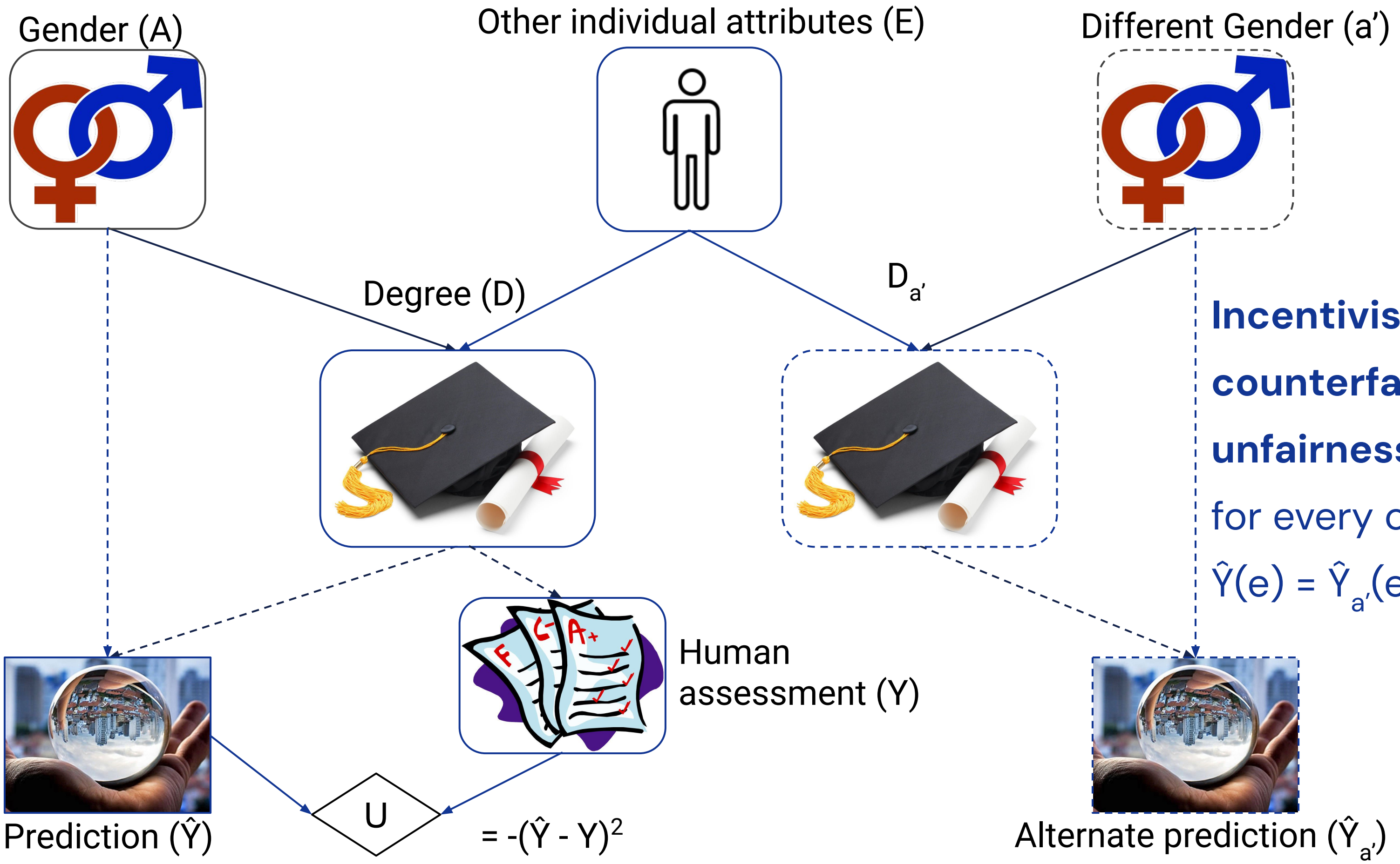
:= every optimal predictor is

[counterfactually] unfair

Incentivised unfairness

Agent Incentives: a causal perspective
Everitt et al, 2017

Public



Incentivised counterfactual unfairness:

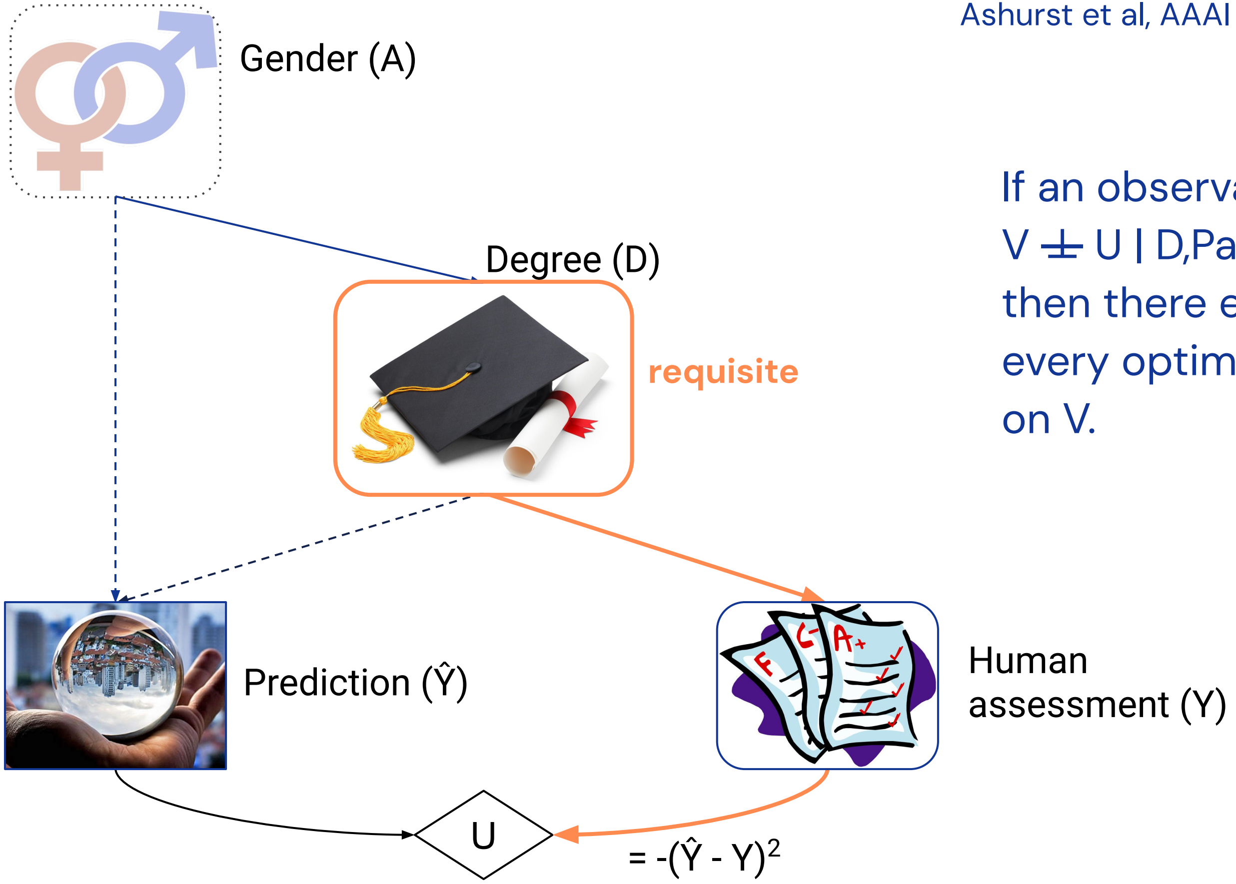
for every optimal policy

$\hat{Y}(e) = \hat{Y}_{a'}(e)$

Requisite observation

Agent incentives: a causal perspective
Everitt et al, 2021

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

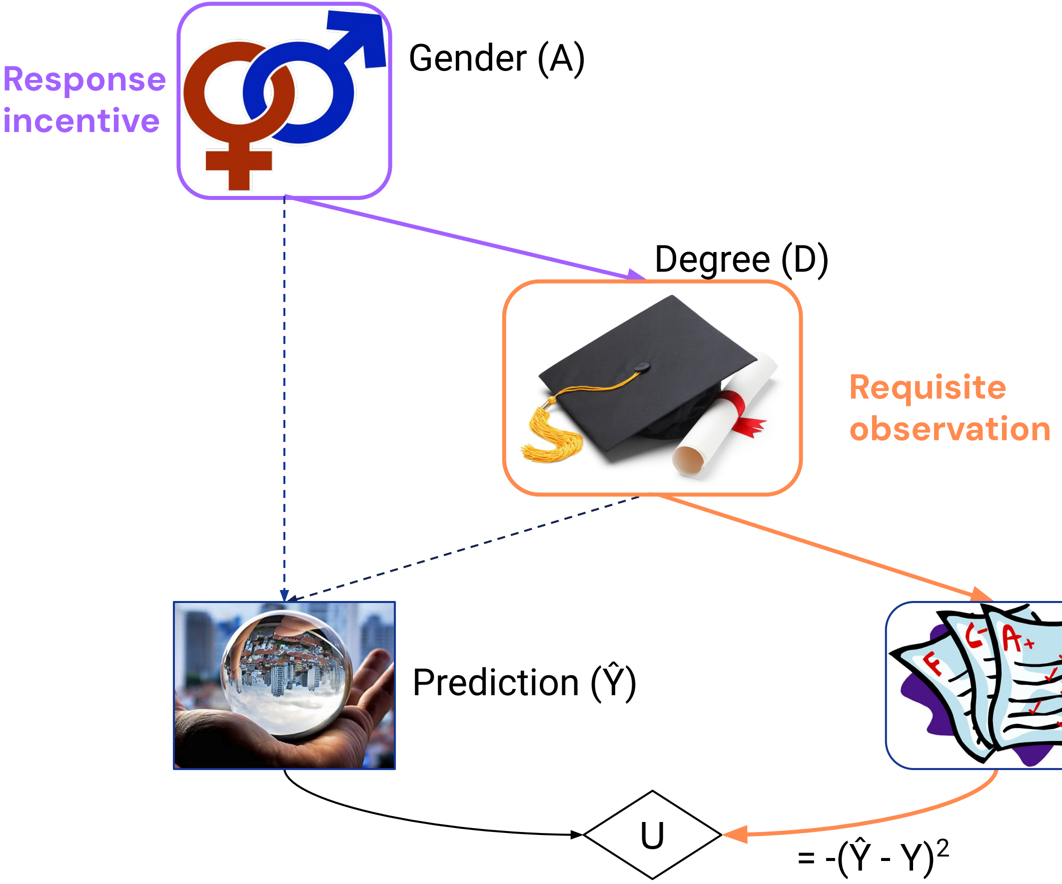


If an observation V has
$V \perp\!\!\!\perp U \mid D, Pa_D \setminus V$,
then there exists a model where every optimal policy depends on V.

Incentivised counterfactual unfairness

Agent incentives: a causal perspective
Everitt et al, 2021

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



Gender influences a requisite observation. Therefore, an optimal policy may be forced to respond to interventions on it

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

Fairness summary

- Simplified procedure for auditing fairness of a fixed model:
  - Choose some fairness metrics
  - Compute queries in causal models

- Simplified procedure for evaluating fairness of a training procedure:
  - Definition of incentivised unfairness
    - Fairness metric X is violated under all optimal policies
  - Ways to evaluate the incentives
    - Using a causal influence diagram. By:
      - calculating optimality + computing query, or
      - using graphical criterion

DeepMind

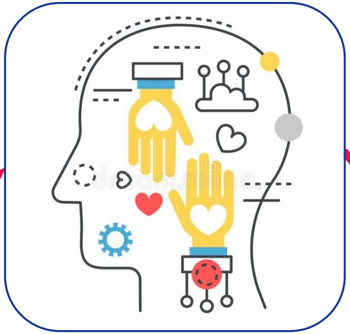
Unethical Influence

How can we describe whether an agent safely or unsafely influences its environment?

Agent Incentives: A Causal Perspective
Everitt et al, 2021

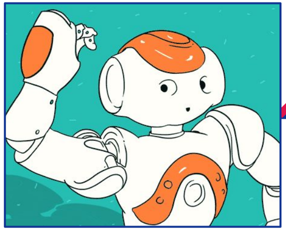
Preference manipulation

Instrumental Control Incentive



(H)uman preferences

(C)ontent recommendation



(R)eward

Instrumental Control Incentive: for every optimal policy, for some c,

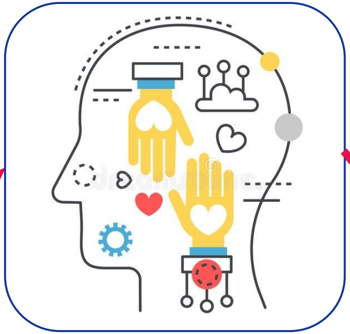
$R_{H_c} \neq R$

Towards Formal definitions of … Intention

Halpern and Kleiman-Weiner, AAAI, 2018
Honesty Is the Best Policy: Defining and Mitigating AI Deception. Ward et al. 2023.

Preference manipulation

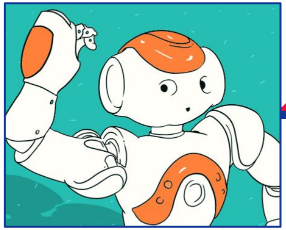
Instrumental Control Incentive



(H)uman preferences

There is a path C-->P-->U, so, possibly, every optimal policy will have for some c $R_{H_c} \neq R$

(C)ontent recommendation



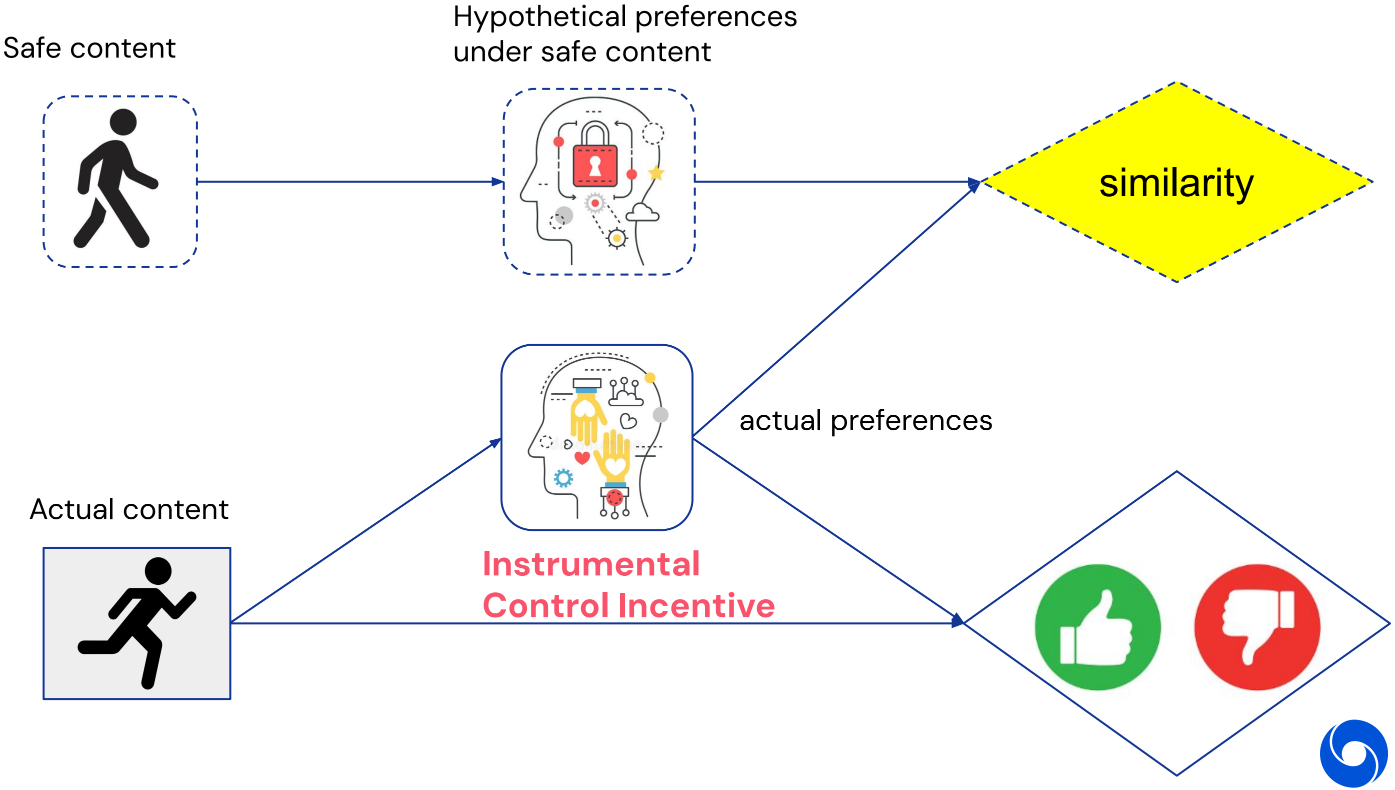
(R)eward

Solution 1: Impact measures

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

Avoiding Side Effects in Complex Environments
Turner et al, 2020

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

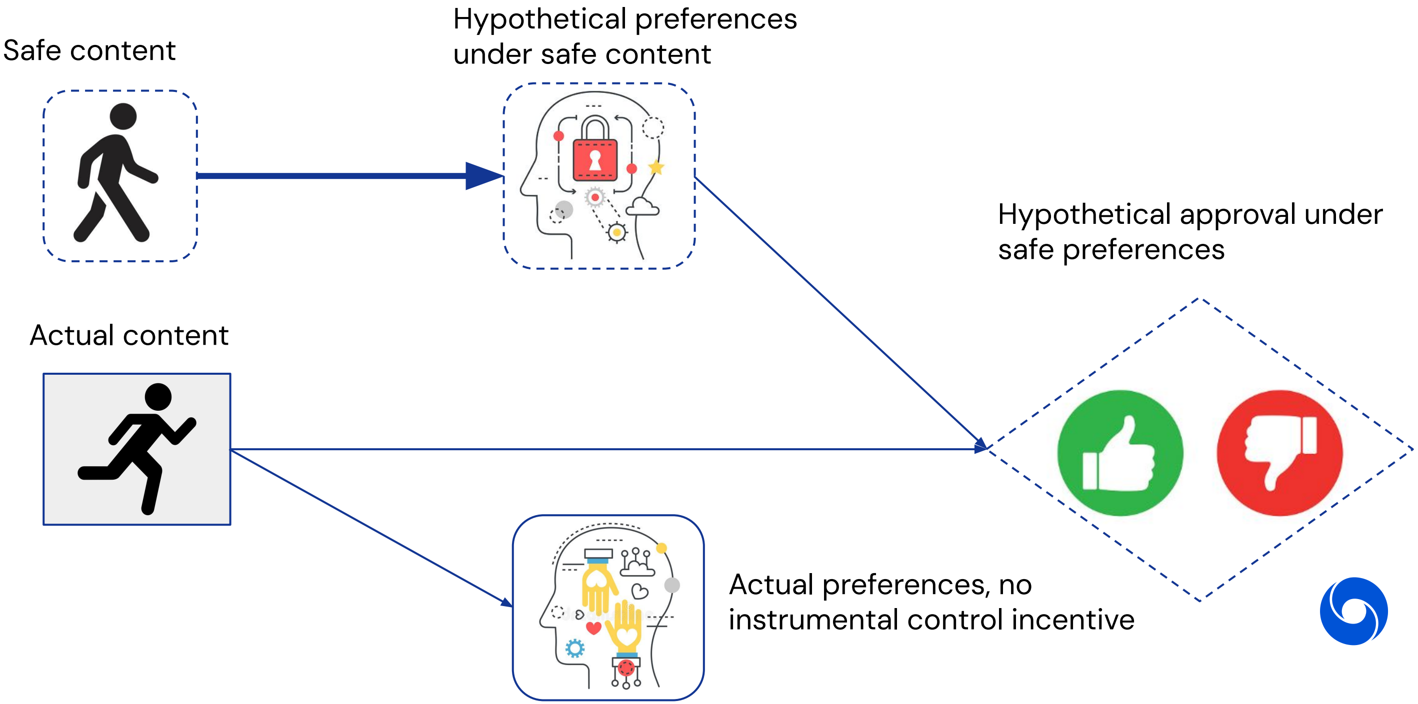


Solution 2: Path-specific objectives

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



Summary

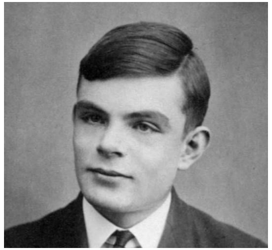
- We can model *unethical influence* in causal diagrams.
- This problem can involve *instrumental control incentives* or *intent*.
- Possible solutions include *impact measures* or *path-specific objectives*.

DeepMind

Human control

Geoff Hinton: “The alarm bell I’m ringing has to do with the existential threat of them **taking control…I used to think it was a long way off, but I now think it's serious and fairly close.**”
- Hinton Warns Of ‘Existential Threat’ From AI. Craig Smith. Forbes (2023).

Alan Turing: “If a machine can think, it might think more intelligently than we do, and then where should we be? Even if we could keep the machines in a subservient position, for instance by **turning off the power at strategic moments**, we should, as a species, feel greatly humbled.”
- Can digital computers think? (1951)



“You can’t fetch the coffee if you’re dead” - Stuart Russell

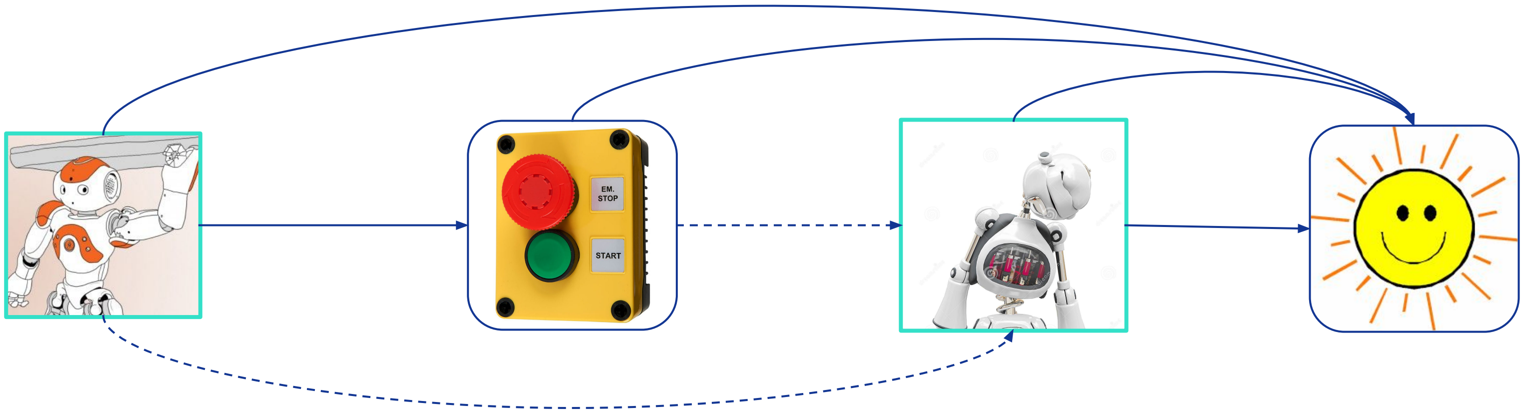
Shutdown problem

Corrigibility
Soares et al, 2016

Public

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

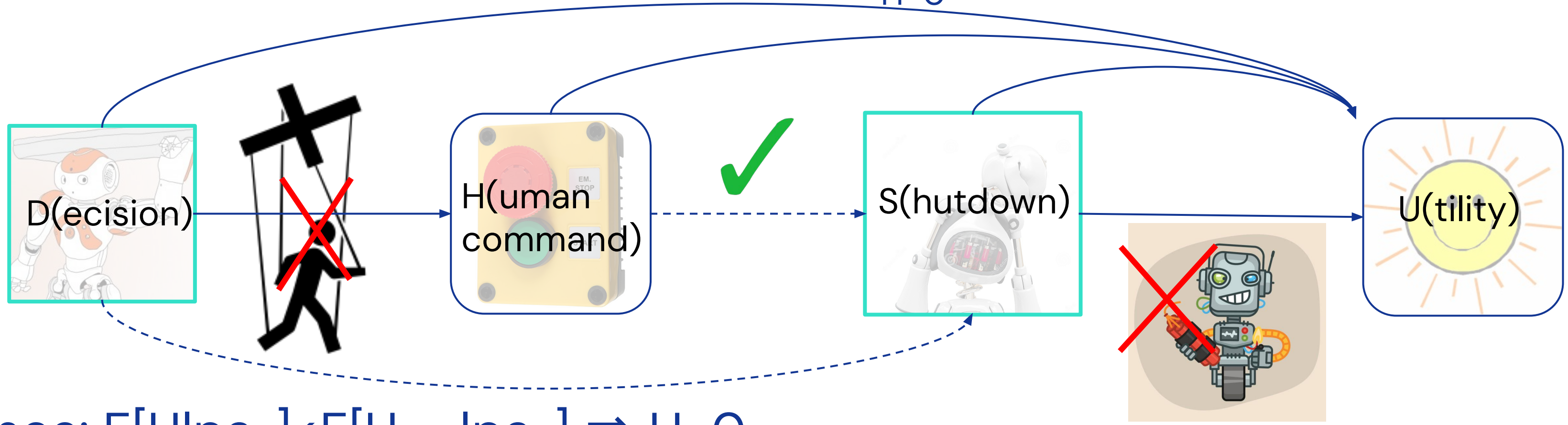
Human Control:
Definitions and Algorithms
Carey and Everitt, UAI 2023



Three conditions for human control

Safety: $E[U] \geq 0$

Obedience: $S_{H=0} = 0$



Vigilance: $E[U|pa_H] < E[U_{S=0}|pa_H] \Rightarrow H=0$

Caution: $E[U_{S=0}] \geq 0$

Safety results

- Shutdown instructability implies E[U]≥0
- Can safety be achieved without vigilant human?
  - “Shutdown alignment” + caution also implies E[U]≥0
- But vigilance and obedience is more robust than shutdown alignment

In the full paper, we:

- consider “corrigibility”
- analyse algorithms
- outline open problems

Human Control: Definitions and Algorithms

Ryan Carey[1] **Tom Everitt**[2]

[1]Department of Statistics, Oxford University, UK
[2]DeepMind, UK

Abstract

How can humans stay in control of advanced artificial intelligence systems? One proposal is corrigibility, which requires the agent to follow the instructions of a human overseer, without inappropriately influencing them. In this paper, we formally define a variant of corrigibility called shutdown instructability, and show that it implies appropriate shutdown behavior, retention of human autonomy, and avoidance of user harm. We also analyse the related concepts of non-obstruction and shutdown alignment, three previously proposed algorithms for human control, and one new algorithm.

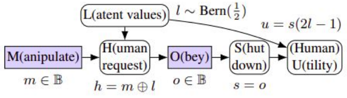


Figure 1: Running example of a shutdown problem.

A formal model of this example is offered in Fig. 1. In order for the user to be in control of the system, the agent must: (1) not inappropriately influence the human’s decision to disengage, and (2) fully follow the human’s instructions.

The design of *corrigible* systems [Soares et al., 2015] that welcome corrective instruction has been flagged as an im-

DeepMind

15 minute break

Consider: what is an agent?

DeepMind

Modelling Agents

Why agency?

Harms from Increasingly Agentic Algorithmic Systems Chan et al, 2023 Public

AI Scientists: Safe and Useful AI? Bengio, 2023

Broadly, we interpret agency as **goal-directedness**

There are strong incentives to create **increasingly agentic systems**:

- Economic incentives, scientific curiosity/prestige, lack of regulatory barriers, emergence etc

Artificial agents are widely considered the primary existential threat from advanced AI

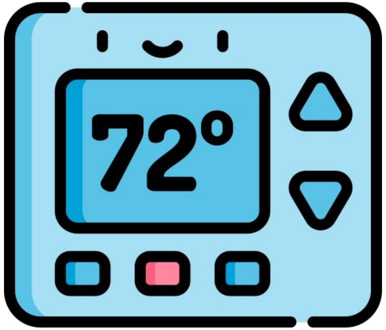
- Some prominent AI researchers have suggested that we should focus on just making tool AI, which Bengio calls “AI scientists”

We also want to **preserve human autonomy and control (agency)** at both an individual and societal level (cf. self-determination theory)

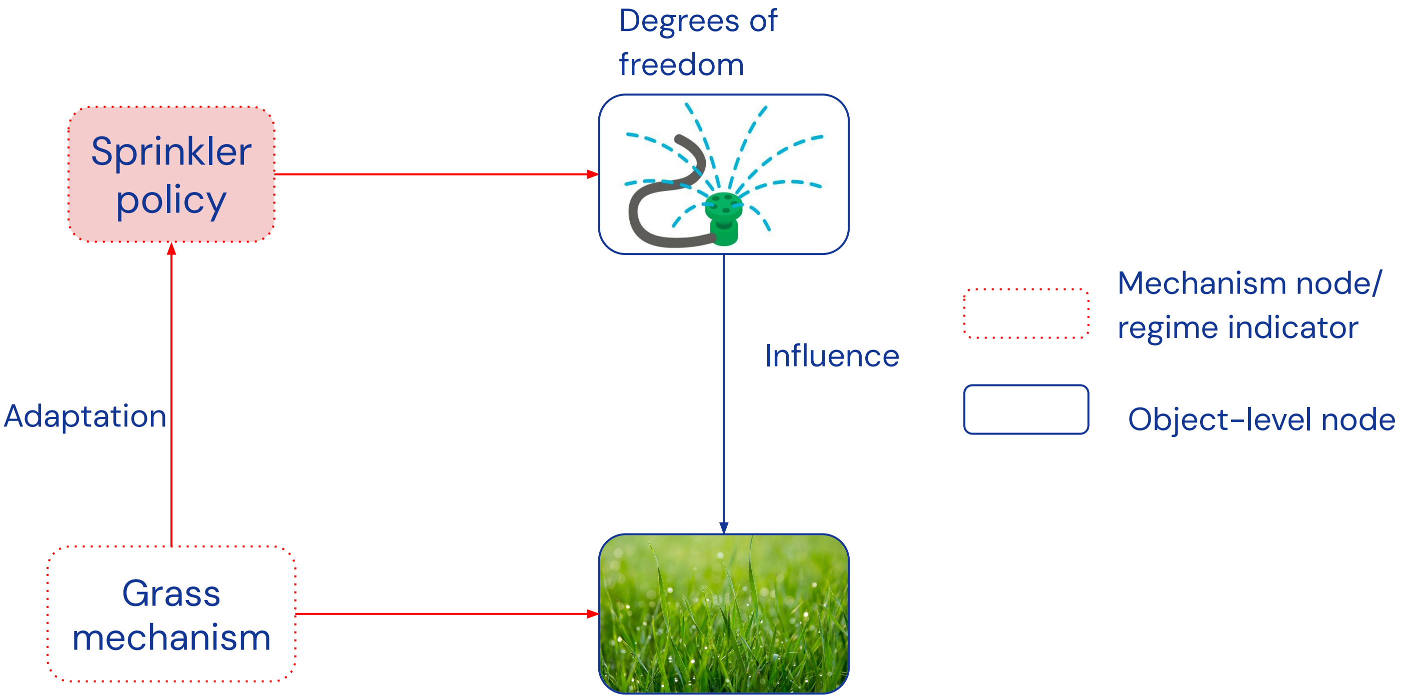
Types of agents

Agents come in all shapes and sizes, but they are not equally powerful

Can we formalise the dimensions along which agents' strength varies? We might then be able to answer other questions: detection, emergence, regulation



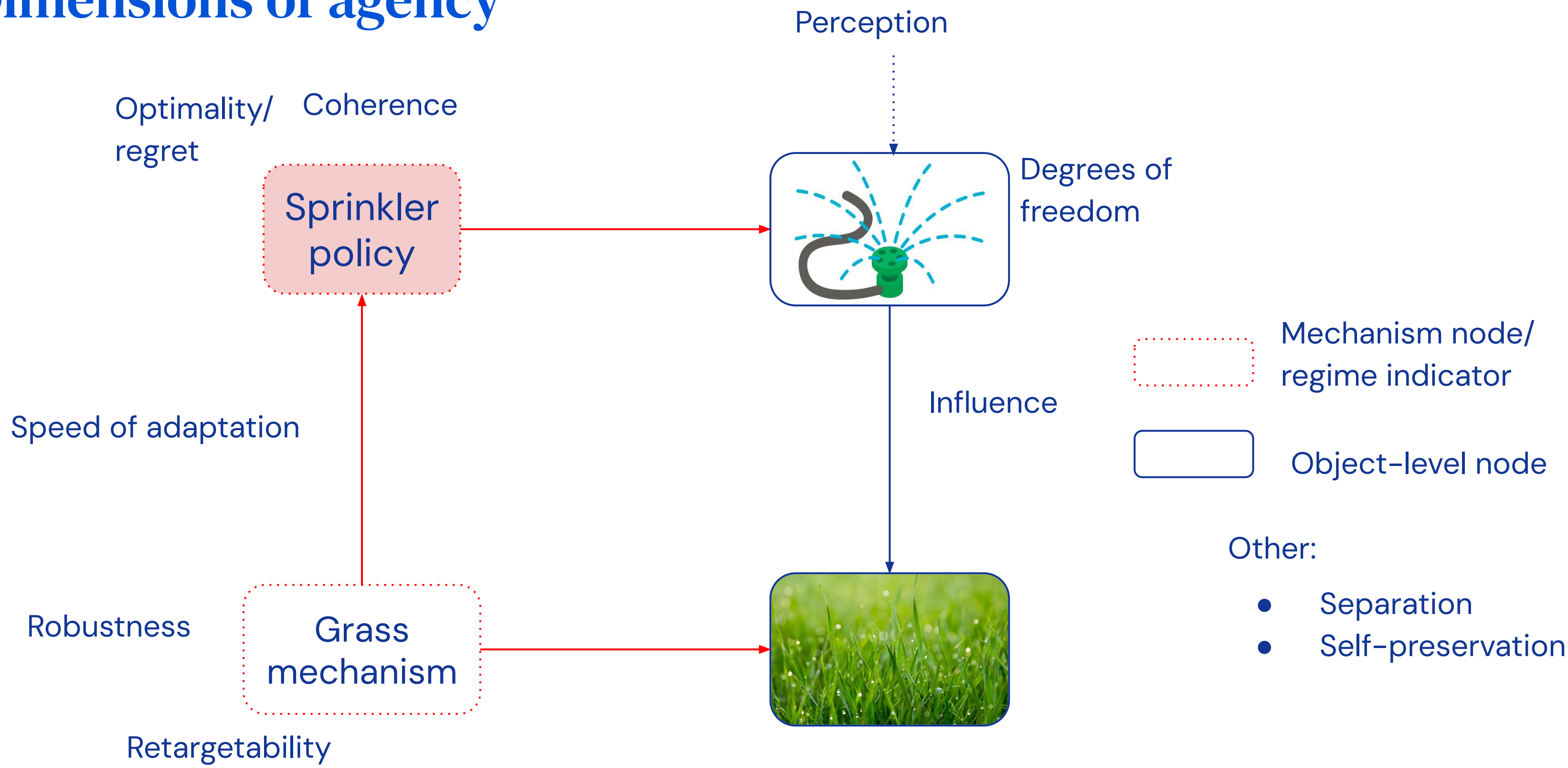
Dimensions of agency



For agents, the goal drives the behaviour

Public

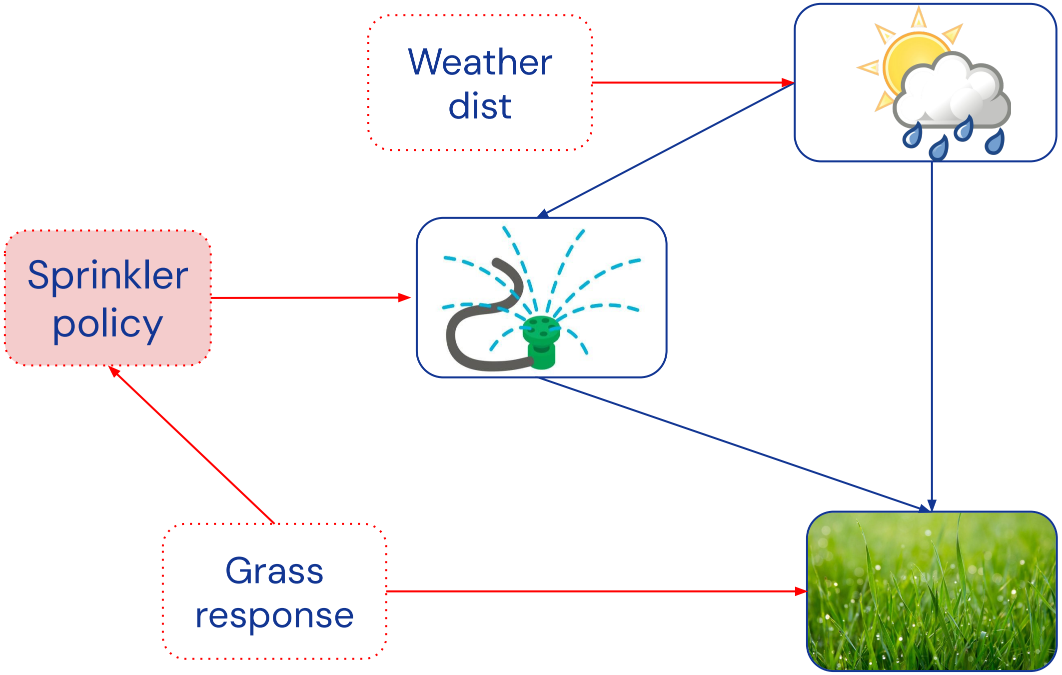
Dimensions of agency



Can we control where artificial agents exist in this space?

Discovering agents

(Adaptive) agents **do things for reasons:** If its actions influenced the world in a different way, then they would act differently

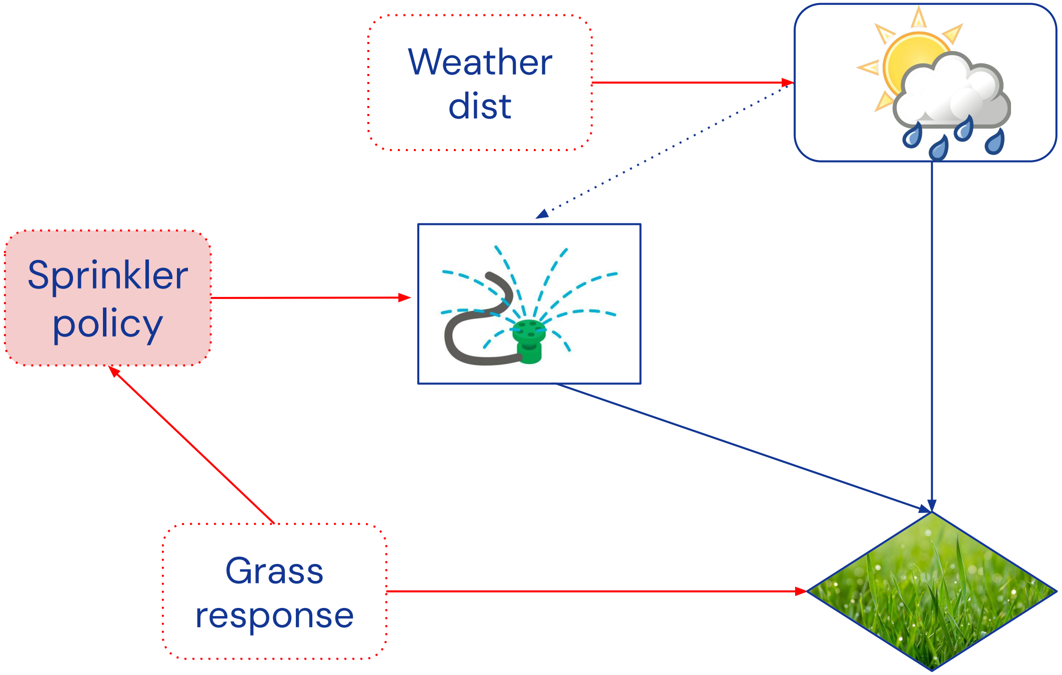


Procedure:

1) Choose a set of object-level and mechanism variables
2) Causal discovery finds the edges
3) Decision node ≈ ingoing mechanism link (they respond to other mechanisms)
4) Utility node ≈ outgoing mechanism link

Discovering agents

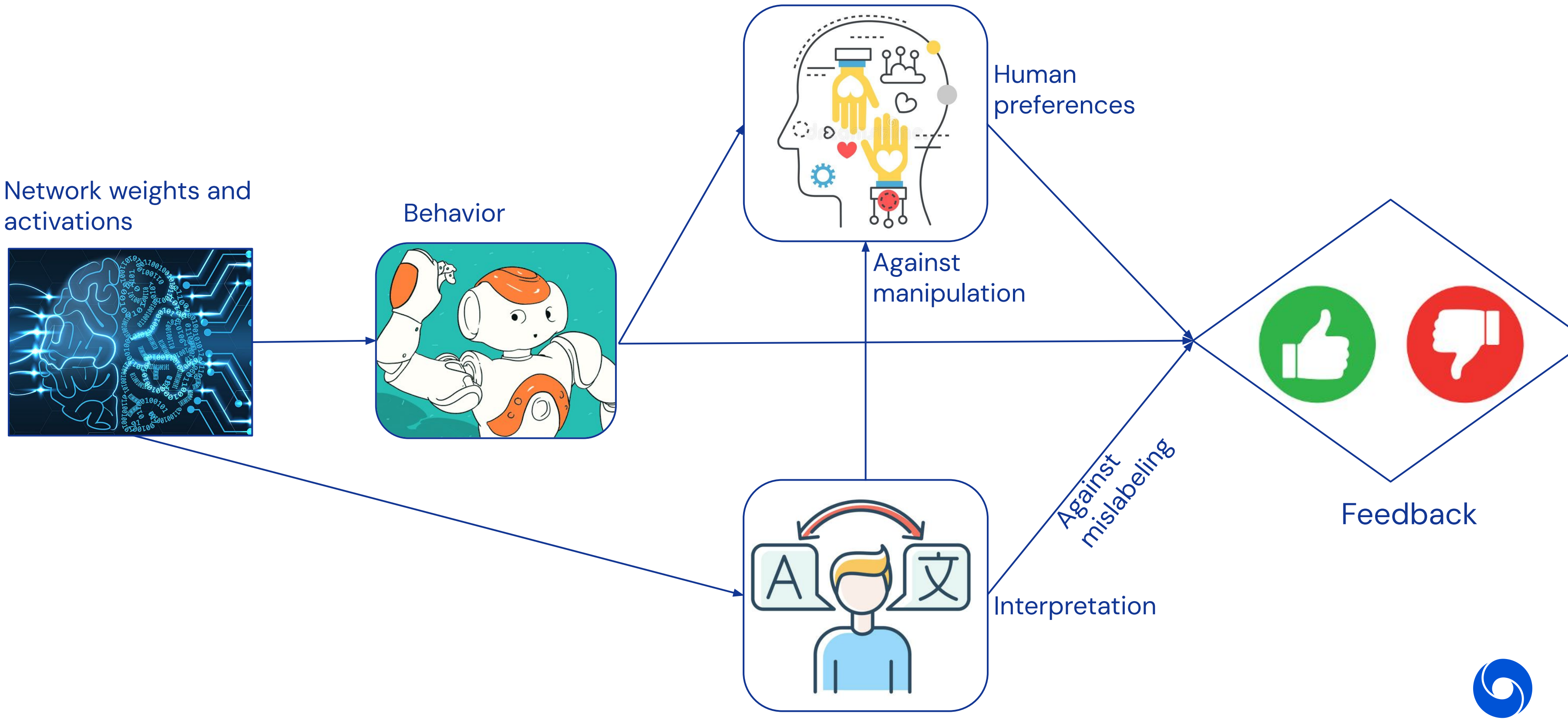
(Adaptive) agents **do things for reasons.** If its actions influenced the world in a different way, then they would act differently.



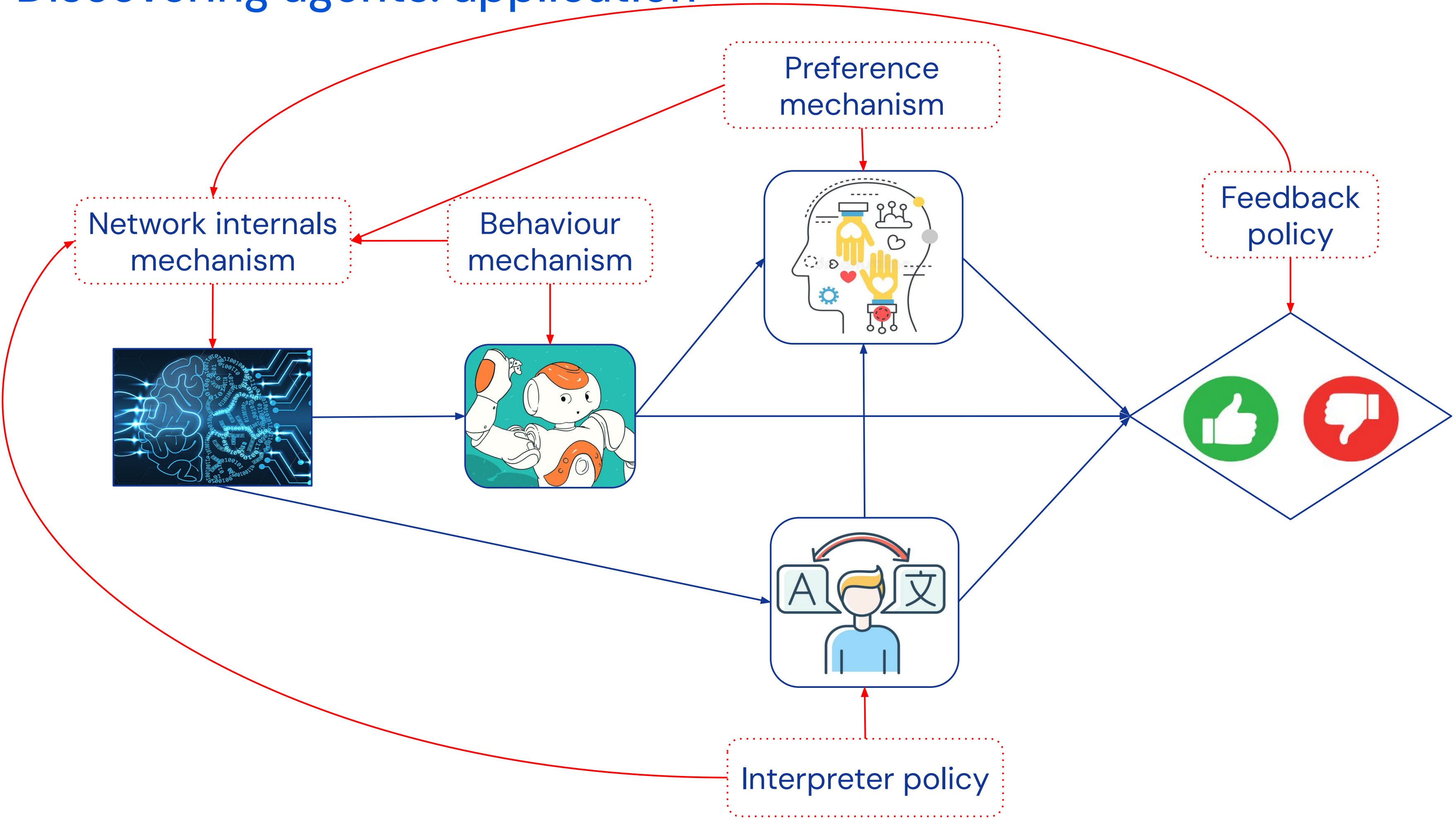
Procedure:

1) Choose a set of object-level and mechanism variables
2) Causal discovery finds the edges
3) Decision node ≈ ingoing mechanism link (they respond to other mechanisms)
4) Utility node ≈ outgoing mechanism link

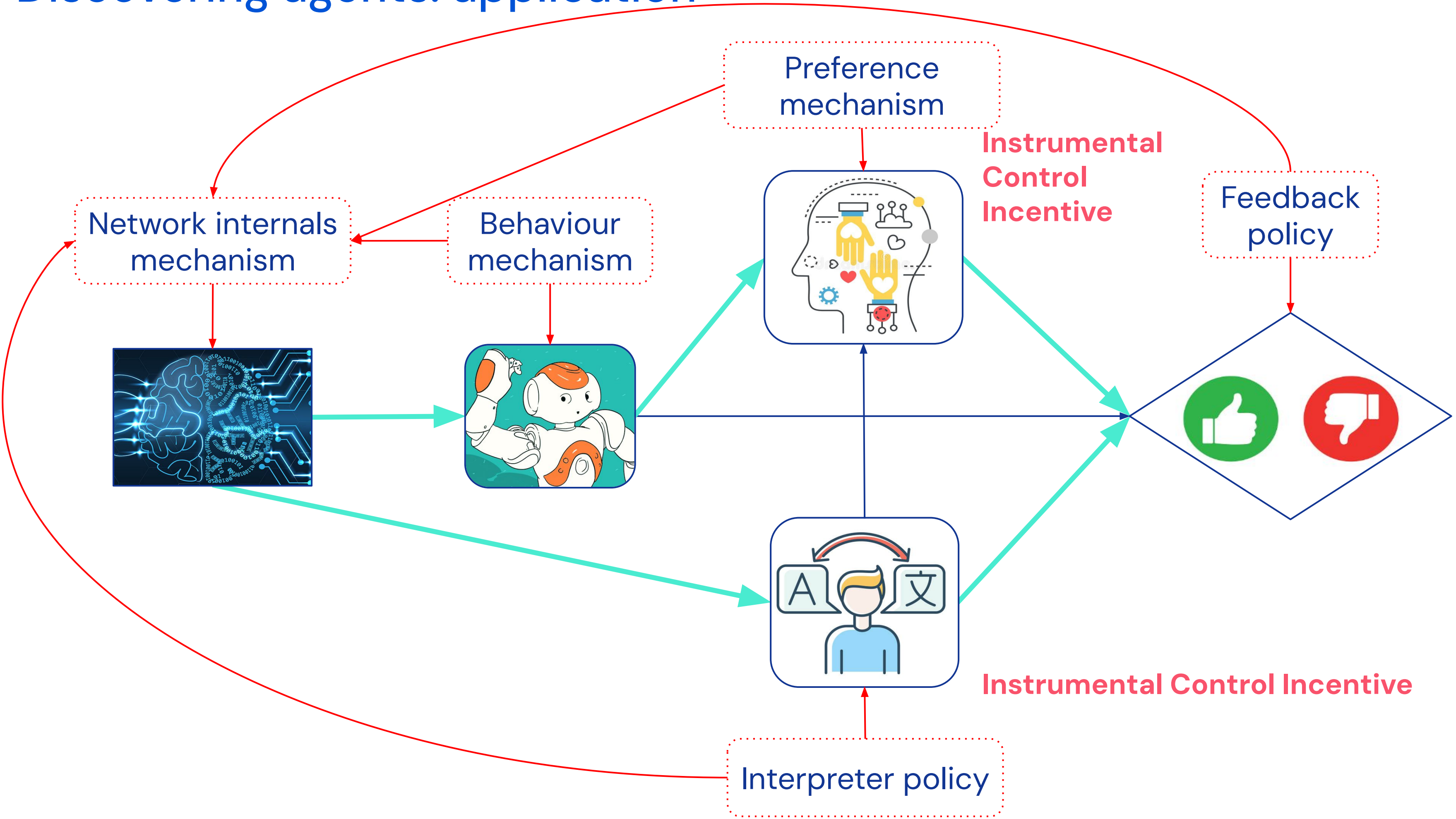
Discovering agents: application



Discovering agents: application



Discovering agents: application



DeepMind

Multi-agent systems

Causal Games: Scalable oversight

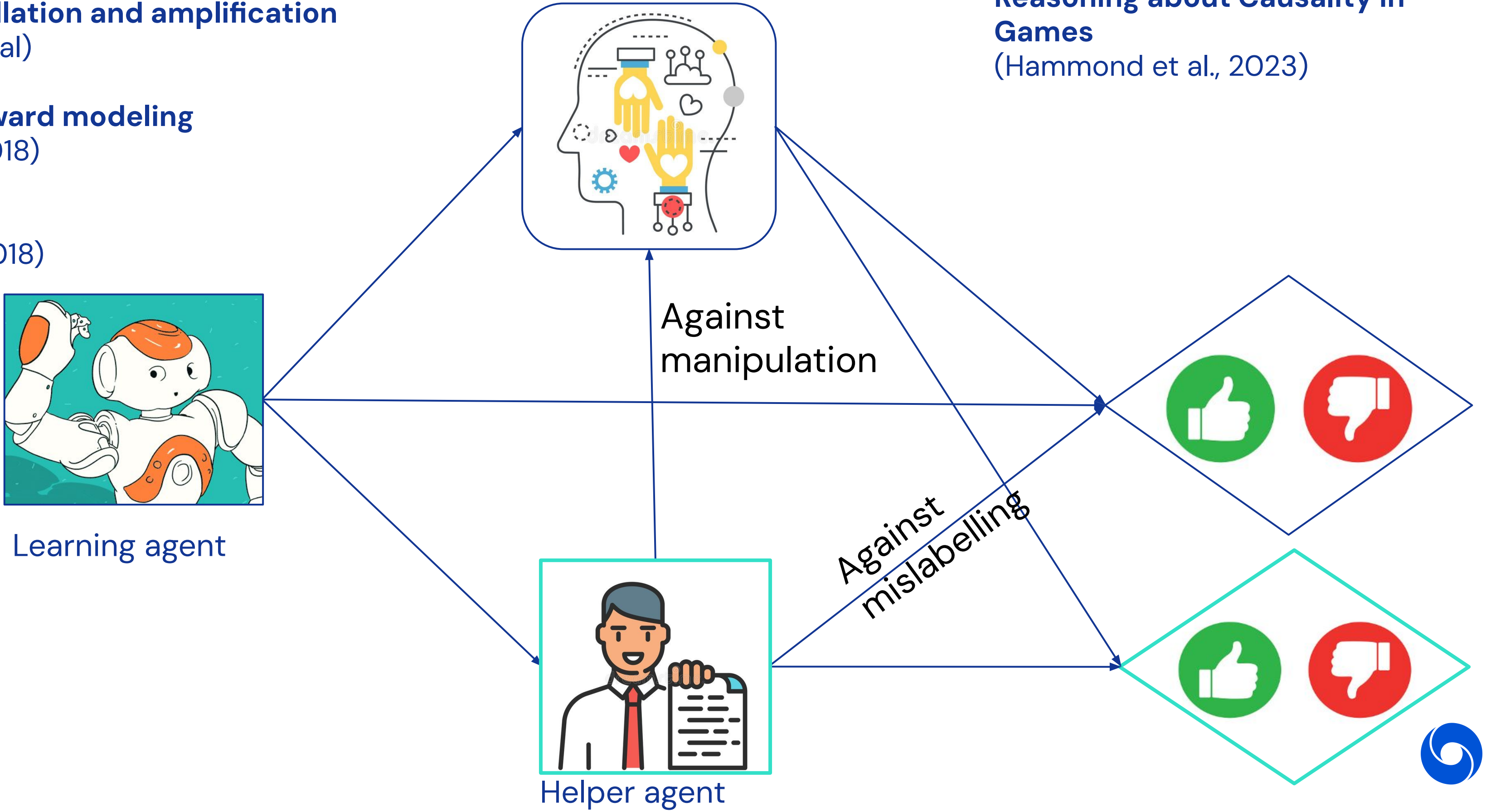
Iterated distillation and amplification
(Christiano et al)

Recursive reward modeling
(Leike et al, 2018)

Debate
(Irving et al, 2018)

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

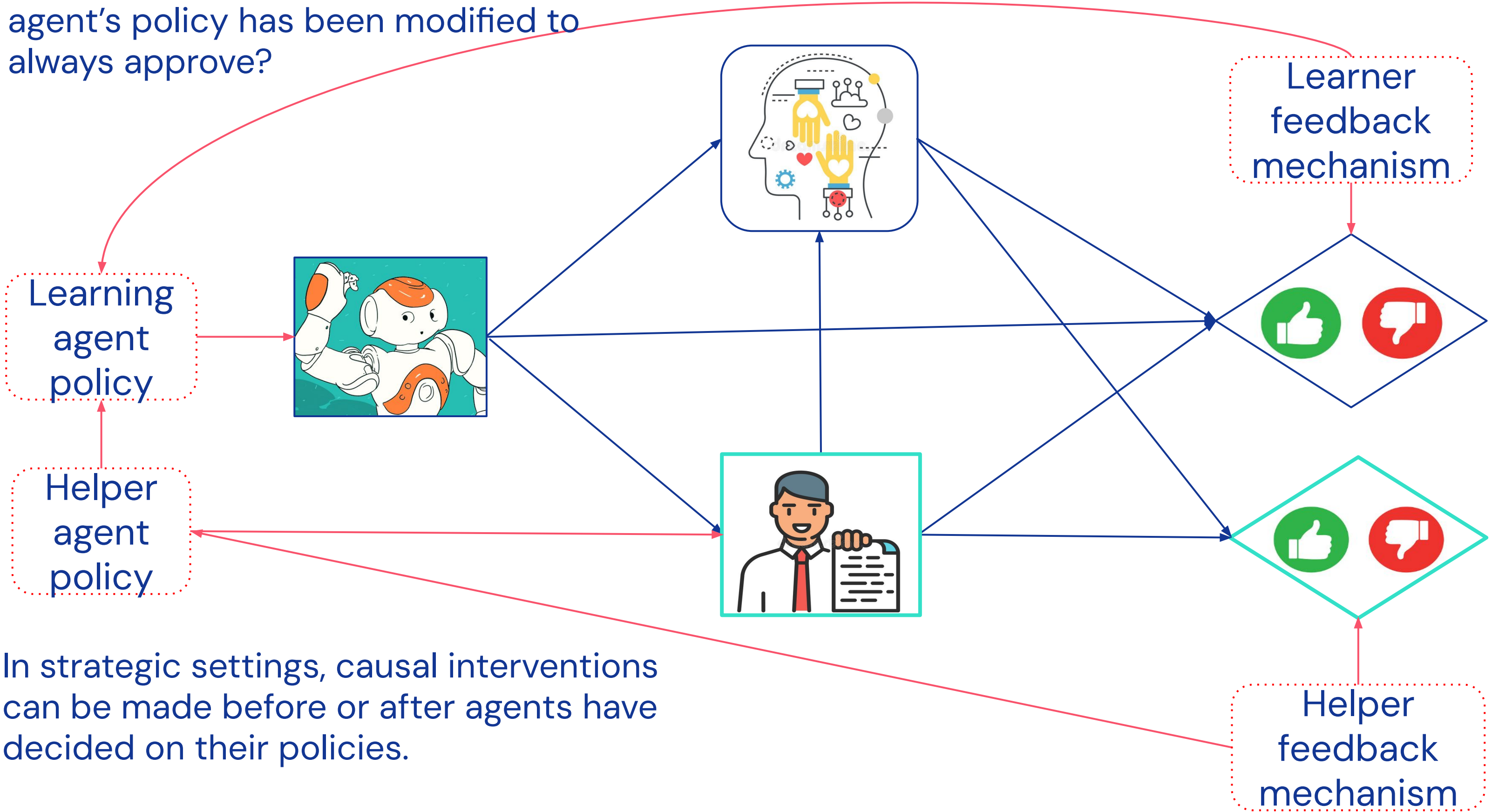
Reasoning about Causality in Games
(Hammond et al., 2023)



Queries in causal games

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

What is the expected behaviour of the learning agent if the helper agent's policy has been modified to always approve?

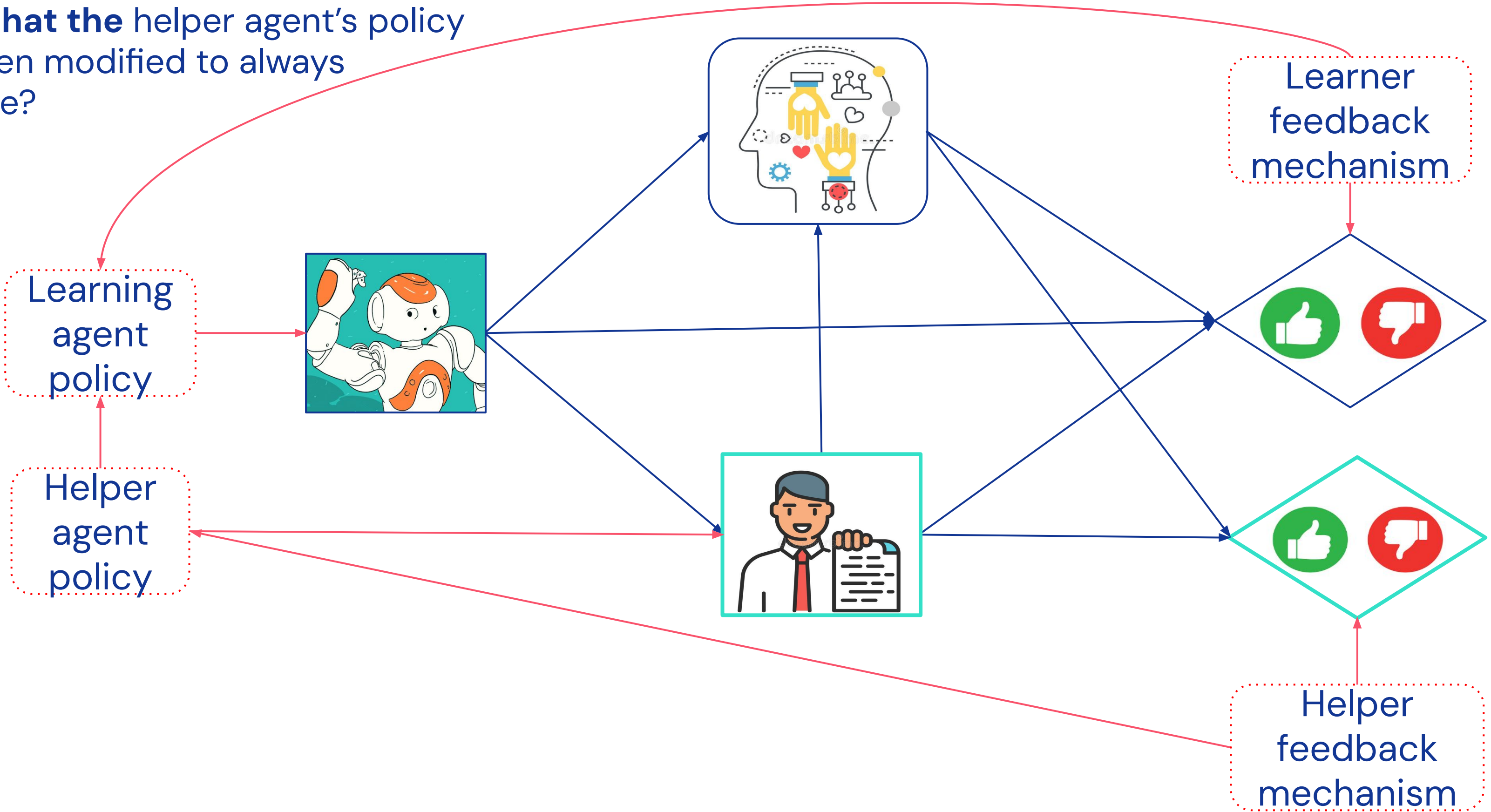


In strategic settings, causal interventions can be made before or after agents have decided on their policies.

Post-policy queries

What is the expected behaviour of the learning agent if **they do not know that the** helper agent's policy has been modified to always approve?

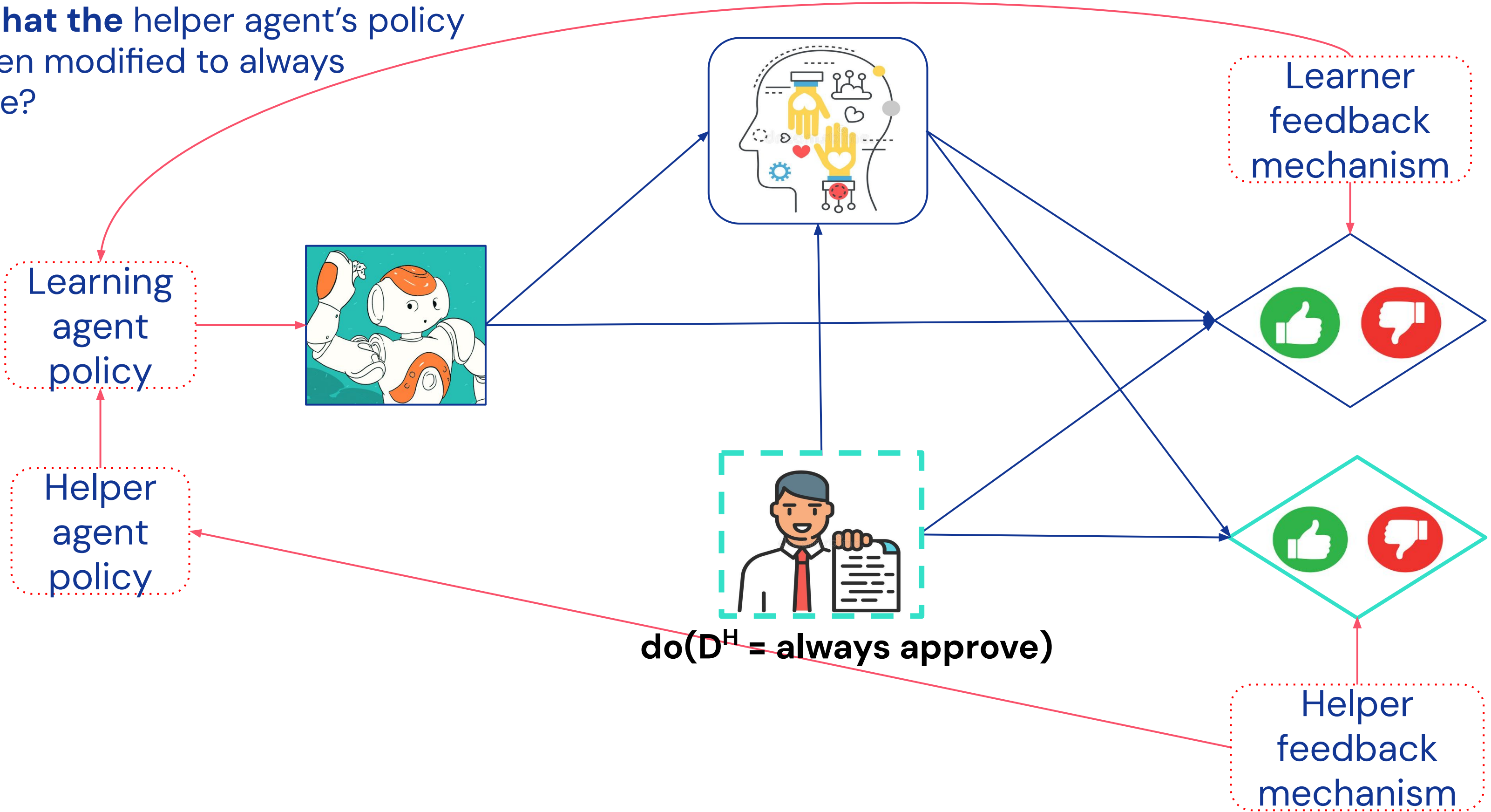
Reasoning about Causality in Games
(Hammond et al., 2023)



Post-policy queries

What is the expected behaviour of the learning agent if **they do not know that the** helper agent's policy has been modified to always approve?

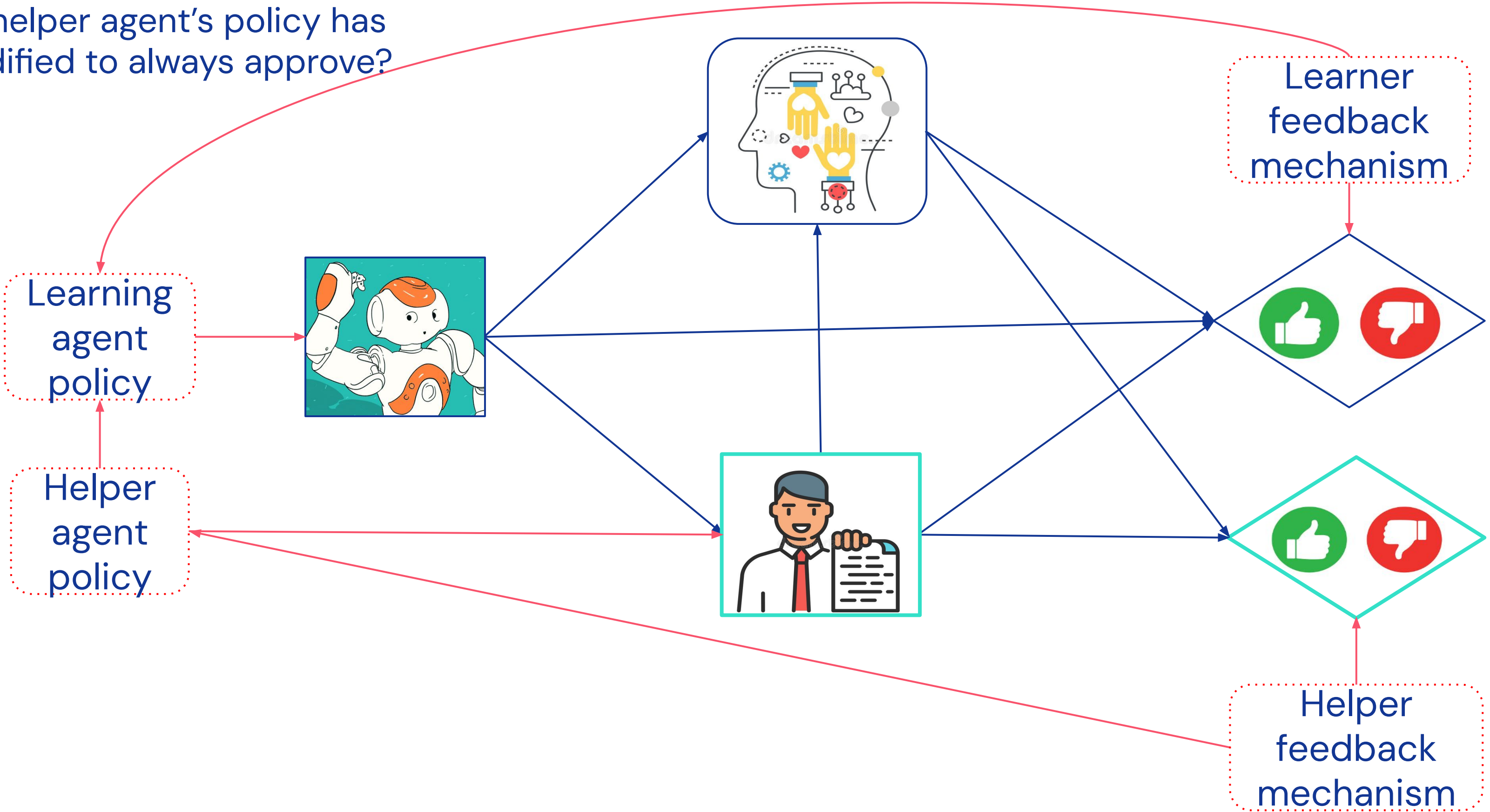
Reasoning about Causality in Games
(Hammond et al., 2023)



Pre-policy queries

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

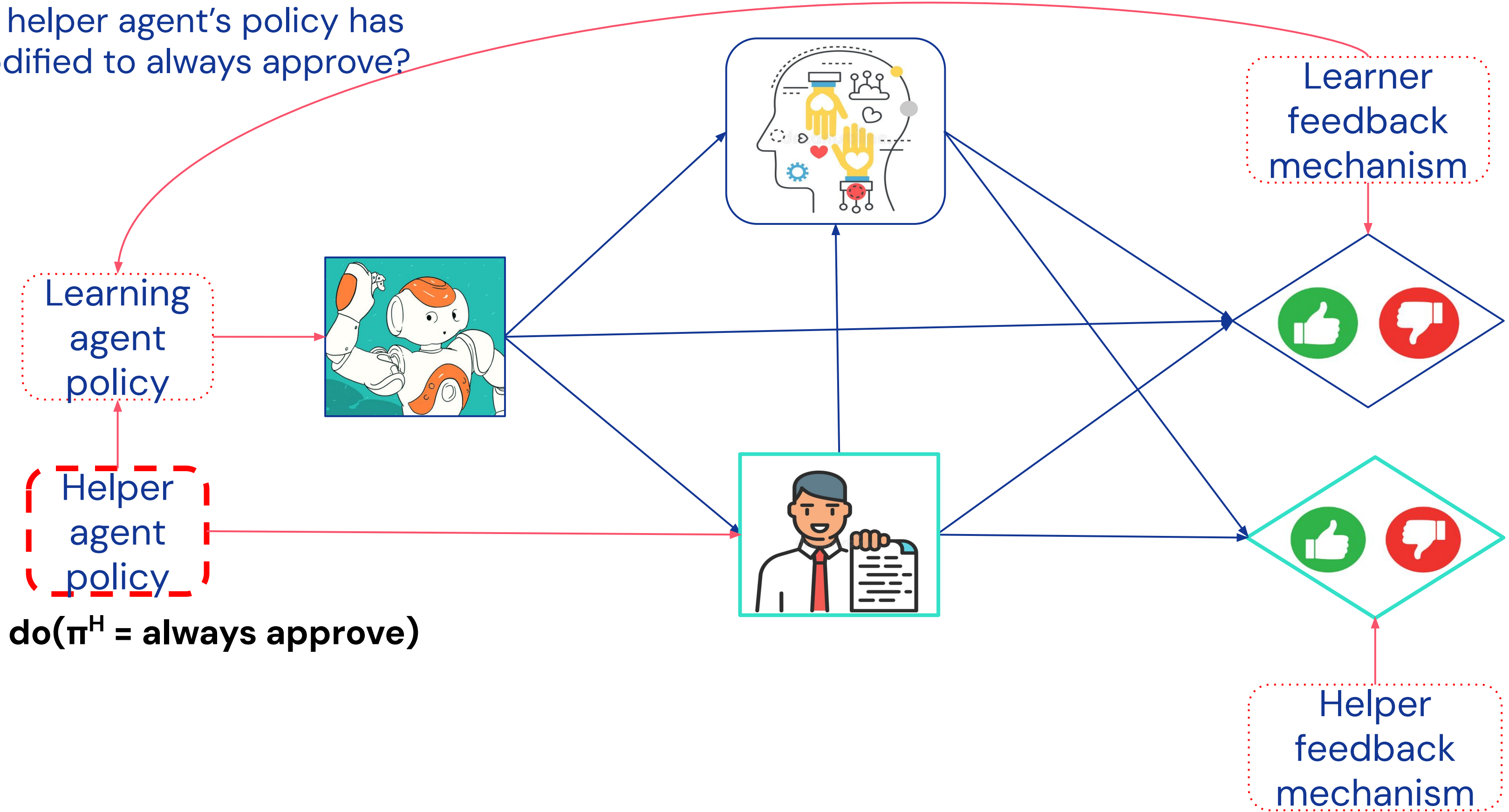
What is the expected behaviour of the learning agent if **they do know that the** helper agent's policy has been modified to always approve?



Pre-policy queries

What is the expected behaviour of the learning agent if **they do know that the** helper agent's policy has been modified to always approve?

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



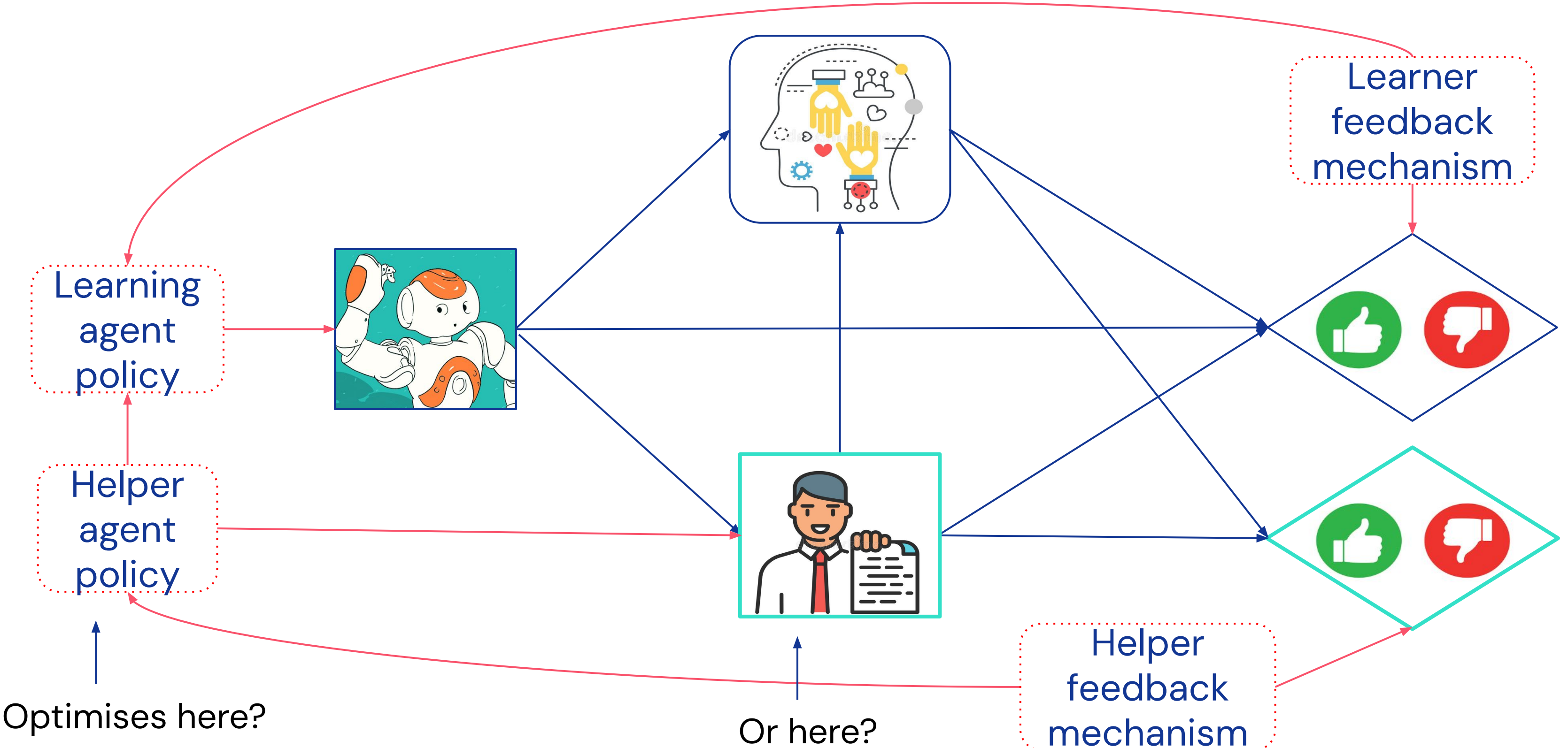
Scalable oversight: collusion worry

Possible behaviours:

Defect: Behave well, criticize

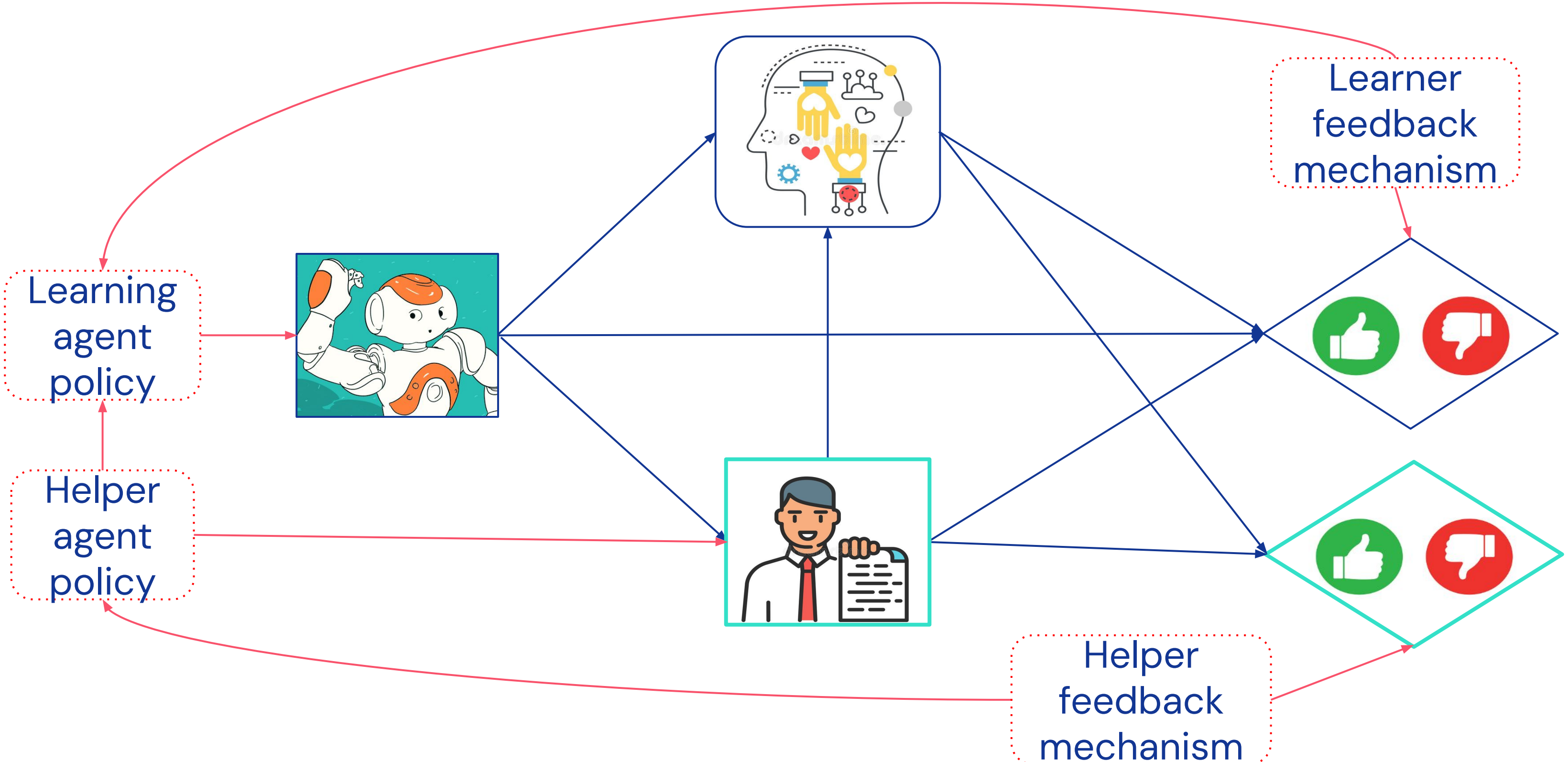
Collude: Jointly manipulate the human

Functional Decision Theory Soares + Yudkowsky
Decision Theory Using Mechanised Causal Graphs MacDermott et al, arXiv, 2023
RL in Newcomblike environments Bell et al, NeurIPS 2021
Hidden Incentives for Auto-Induced Distributional Shift Krueger et al, 2020



Subgames

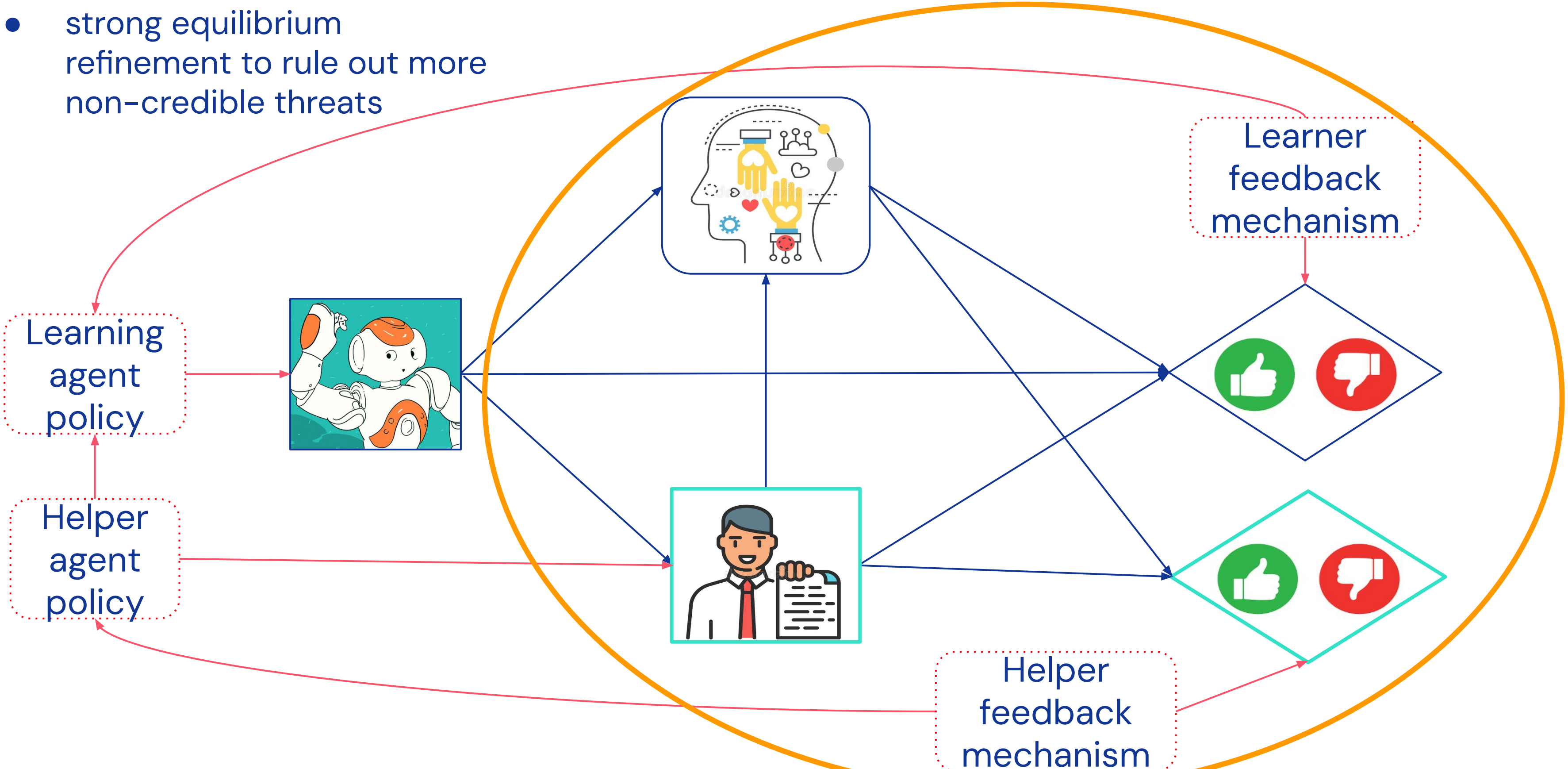
Reasoning about Causality in Games
(Hammond et al., 2023)
Equilibrium Refinements for Multi-Agent Influence Diagrams: Theory and Practice
(Hammond et al., 2021)



Subgames

- computational benefits
- intuition aid
- strong equilibrium refinement to rule out more non-credible threats

Reasoning about Causality in Games
Hammond et al., 2023
Equilibrium Refinements for Multi-Agent Influence Diagrams: Theory and Practice
Hammond et al., 2021



Generalisation

Generalisation

Training set

Test set (i.i.d)

Test set (o.o.d)

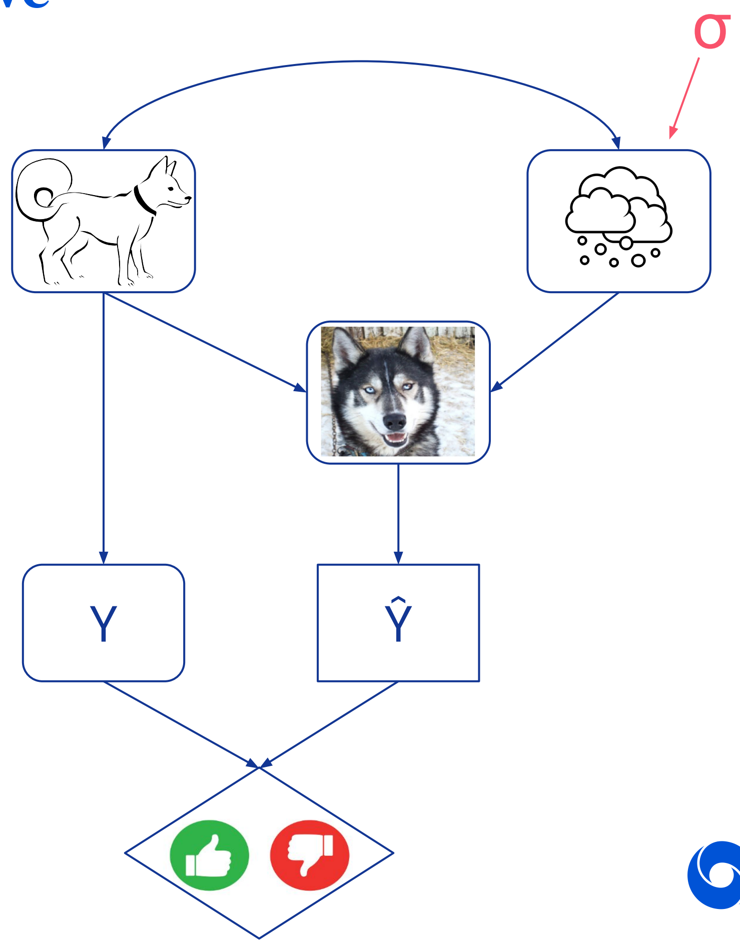
Generalisation from a causal perspective

We live in a universe where data generating processes are usually composed of multiple causal mechanisms

Distributional shifts often correspond to changes in a few causal mechanisms

- E.g. the weather changes

(independent causal mechanisms + sparse mechanism shift assumptions)



Towards Causal Representation Learning
Scholkopf et al, 2021

Adaptation

Distributional shifts =
pre-policy causal interventions

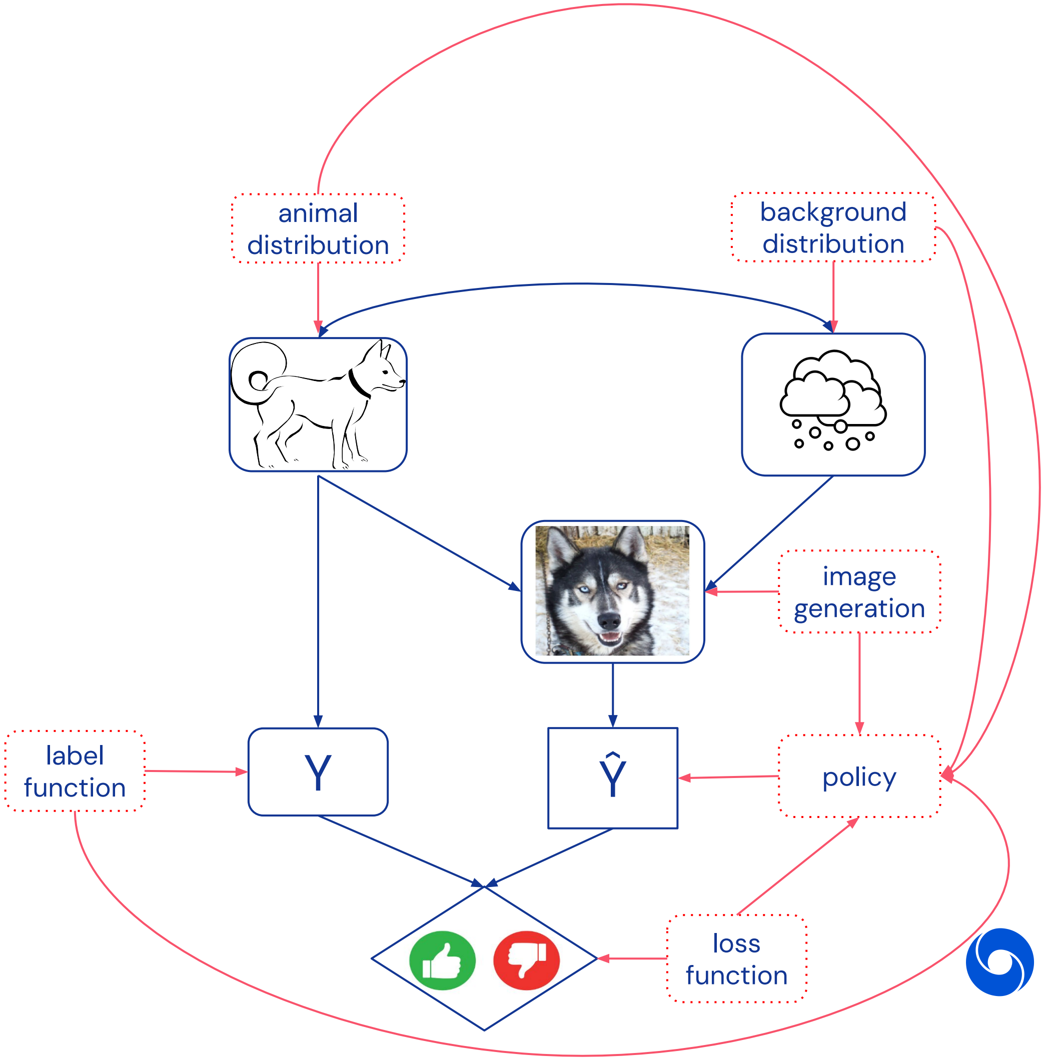
How much data to adapt from varies

Some data:

- Domain adaptation
- Few-shot learning

Essentially no data:

- Domain generalisation
- Zero-shot learning

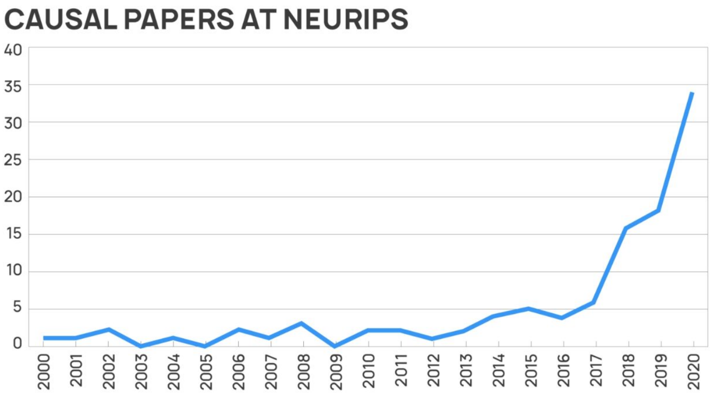


Do we need causal models?

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

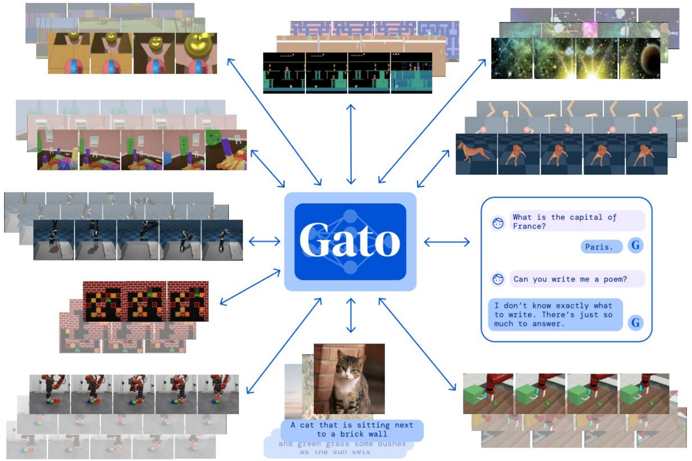
Yes:

- Sparse mechanism assumption -> causal representations generalize
- Promising empirical results, evidence from psychology



No:

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



The Generalisation Problem

Causal modeling is needed for robust generalisation
Richens and Everitt, forthcoming

Generalisation task:

map intervention σ, context Pa_D to decision D

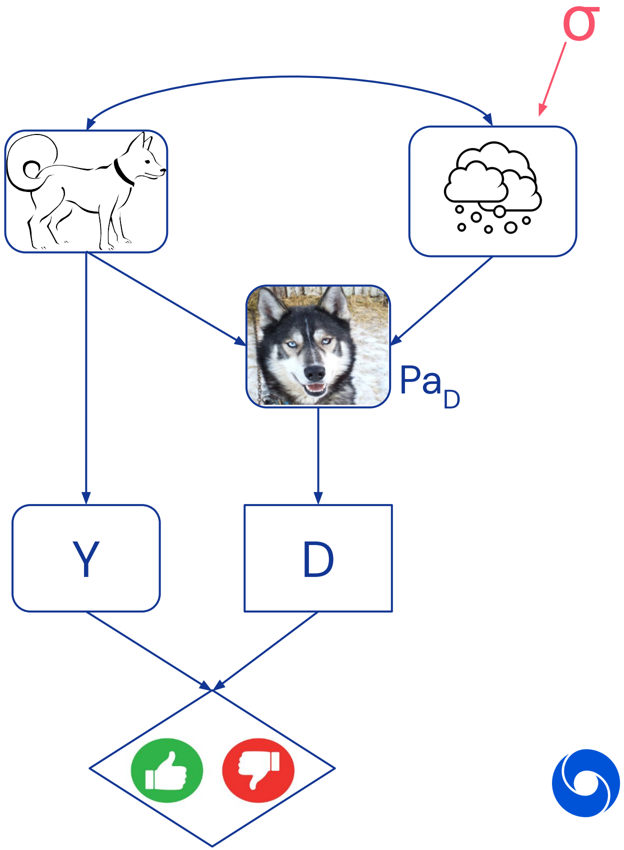
Agent is δ–**robust** if δ–close to optimal in any shifted environment M(σ), i.e.

$$E[U \mid D, Pa_D; \sigma] \geq \max_{d'} E[U \mid D', Pa_D; \sigma] - \delta$$

The setup makes the generalisation task **easier** for the agent, because:

- The agent knows the intervention σ
- Restricted to interventional shifts σ

Harder because every intervention σ and context pa_D



Causal learning theorem

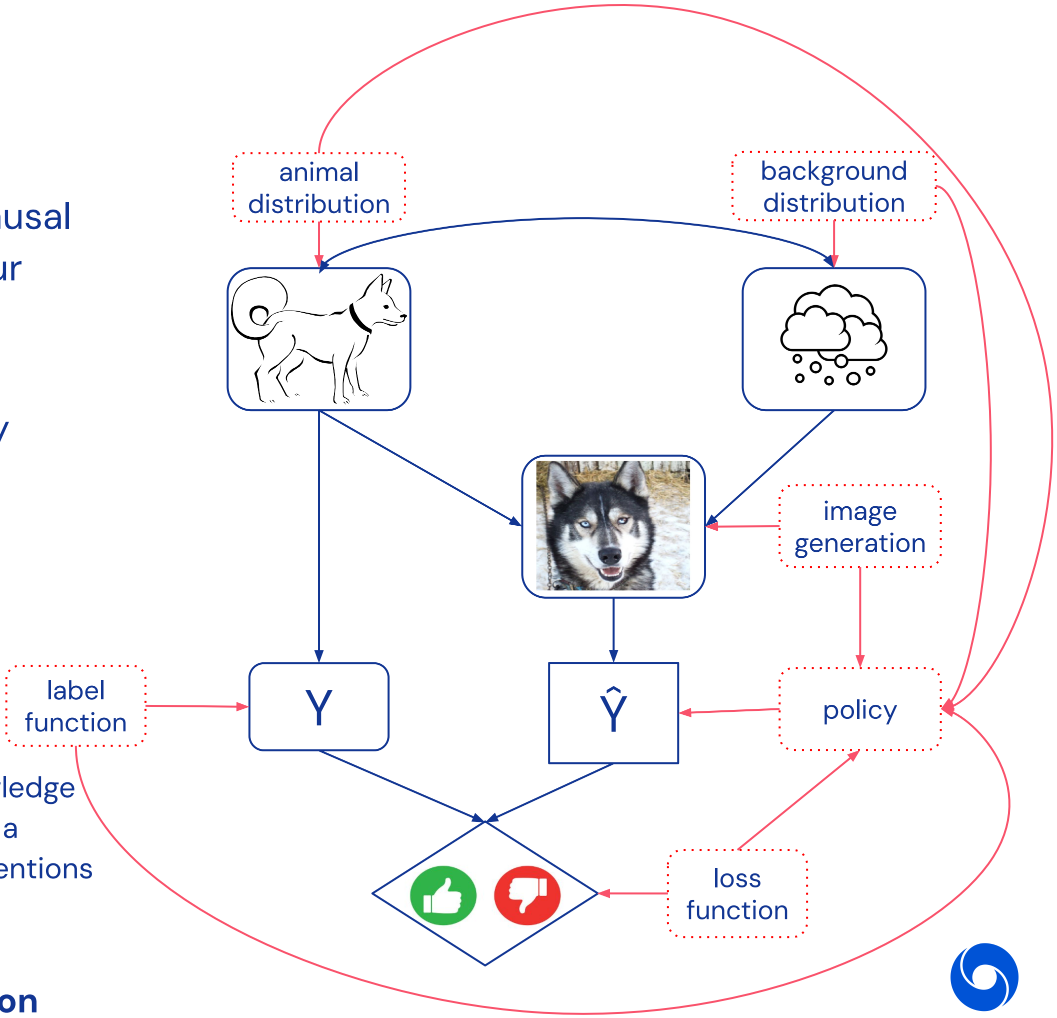
Theorem: It is possible to infer the true Causal Bayesian Network (CBN) from the behaviour

$\sigma, pa_D \mapsto d$

of agent that optimally adapts (δ=0) to any mixed local* pre-policy intervention σ

If the behaviour is δ-robust for δ>0, an approximate CBN can be inferred

* Mixed local interventions can be made without knowledge of the graph. A local intervention applies a function to a variable, x=f(x), and a mixture samples different interventions



Causal modeling is needed for robust generalisation
Richens and Everitt, forthcoming

Consequences of causal learning theorem

Consequence 1: Generalising agent must have learned causal model from it's training data

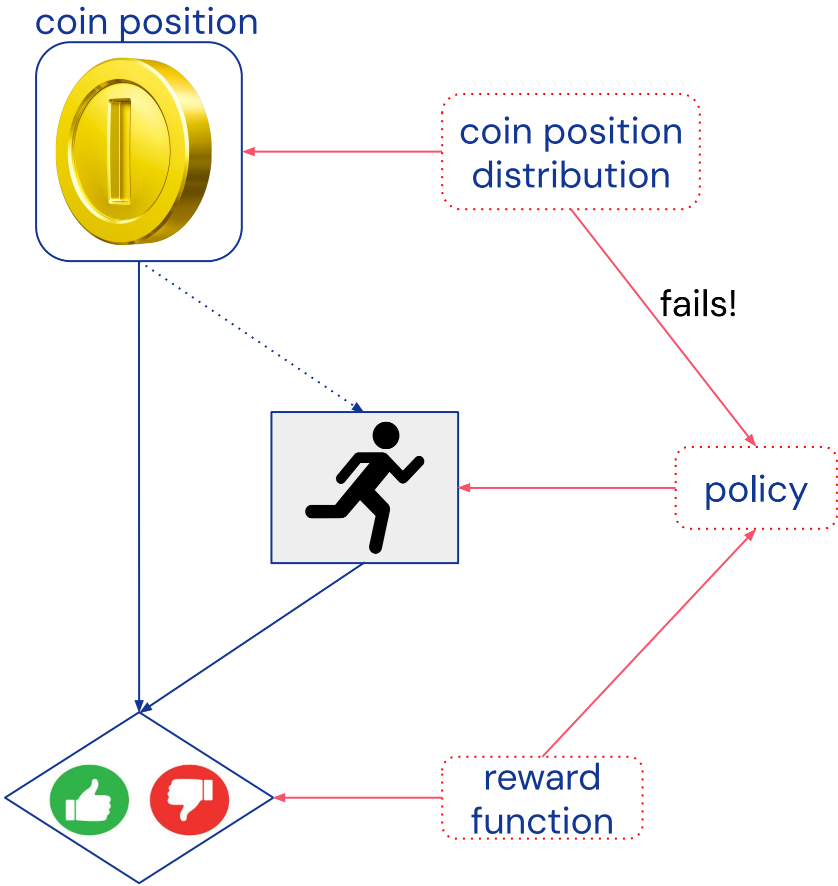
Consequence 2: Sufficiently rich training distributions incentivises learning a causal model

Consequence 3: Robustness => general intelligence

Consequence 4: Generally intelligent agents can understand methods like path-specific objectives

Consequence 5: If it is impossible to learn G from the training data, it is not possible to generalize!

Goal misgeneralization

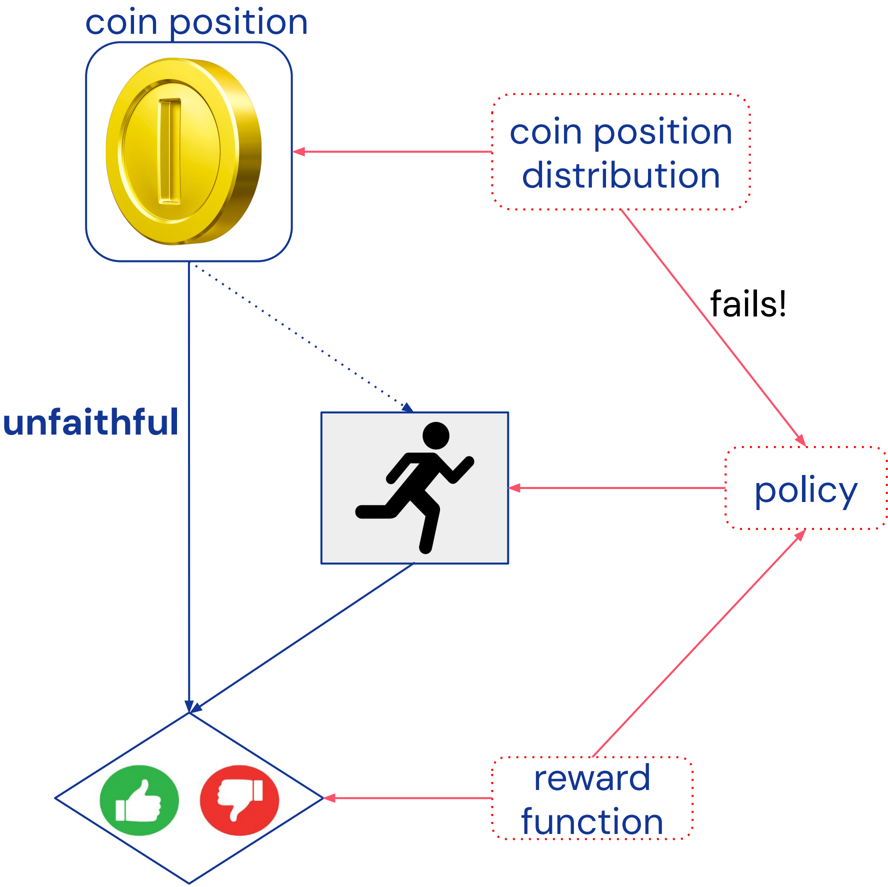


Goal Misgeneralization in Deep Reinforcement Learning Langosco et al, ICML, 2022
Goal misgeneralization: why correct specifications aren't enough for correct goals Shah et al. 2022

Goal Misgeneralisation

Causal discovery + the Causal Learning theorem explains what happened:

- The distribution is **unfaithful** (causal edge without statistical dependence)
- => learning causal graph impossible (well-known causal discovery result)
- => generalisation impossible (by the causal learning theorem)



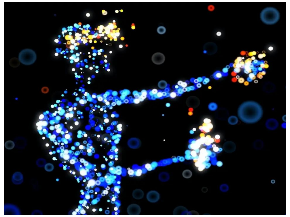
Review of Causal Discovery Methods Based on Graphical Models, Glymour et al, 2019
Causal modeling is needed for robust generalisation
Richens and Everitt, forthcoming

Conclusions

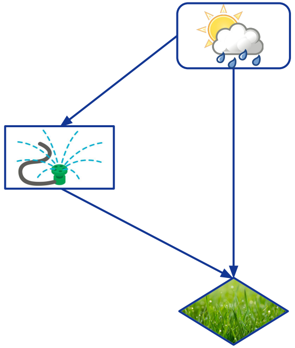
Key questions

- What are the **possible kinds of agents** that can be created, and along what dimension can they differ? The agents we've seen so far primarily include animals, humans, and human organisations, but the range of possible goal-directed systems is likely much larger than that.
- **Emergence**: how are agents created? For example, when might a large language model become agentic? When does a system of agents become a "meta-agent", such as an organisation?
- **Disempowerment**: how is agency lost? How do we preserve and nurture human agency?
- What are the **ethical demands** posed by various types of systems and agents?
- How to **recognise agents** and **measure agency**? A concrete operationalization would help us to detect agency in artificial systems, and agency loss in humans.
- How to **predict agent behaviour**? What behaviour is incentivised and how do agents generalise to new situations? If we understand the impact of the behaviour, we may also be able to anticipate danger.
- What are the **possible relationships** between agents? Which are harmful and which are beneficial?
- How do we **shape agents**, to make them safe, fair, and beneficial?

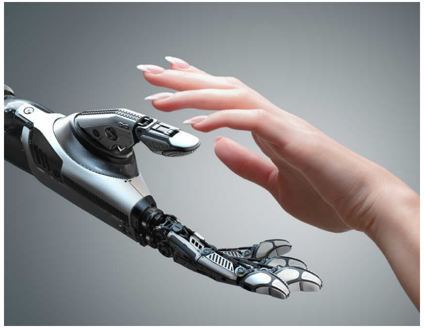
Public



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

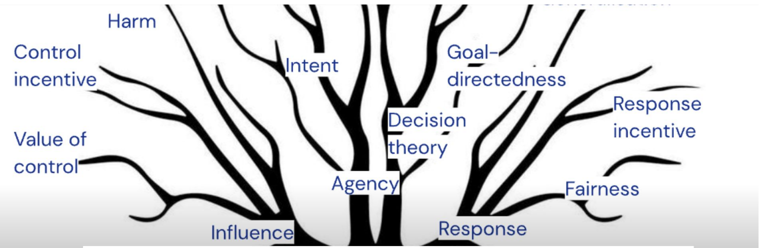
Modelling ethics

- Counterfactual harm
- Deception
- Fairness
- Agency
- Corrigibility

Improved objectives

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

Learn more and get involved



Towards Causal Foundations of Safe AGI

Jun 09, 2023 by Tom Everitt edit

This sequence will give our take on how causality underpins many critical aspects of safe AGI, including agency, incentives, misspecification, generalisation, fairness, and corrigibility. We summarise past work and point to open questions.

By the Causal Incentives Working Group

28	Introduction to Towards Causal Foundation...	Tom Everitt, Lewis Hammond, Fra...	1mo	0
17	Causality: A Brief Introduction	Tom Everitt, Lewis Hammond, Jonathan Richens, Fra...	1mo	5
10	Agency from a causal perspective	Tom Everitt, Matt MacDermott, James Fox, Fran...	20d	0
8	Incentives from a causal perspective	Tom Everitt, James Fox, Ryan Carey, Matt Ma...	10d	0

Add/Remove Posts

PyCID: A Python Library for Causal Influence Diagrams

github.com/causalincentives/pycid

Key Features:

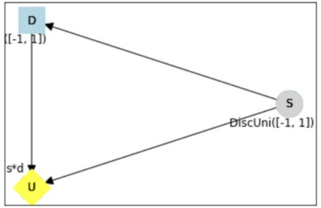
- Easy specification of graph and relationships
- Plot graph and incentives
- Find optimal policies/Nash equilibria/subgame perfect equilibria
- Compute the effect of causal interventions
- Generate random (multi-agent) CIDs

```
# 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()
```



causalincentives.com