Google DeepMind

# The Limits of Predicting Agents from Behaviour

Alexis Bellot, Jon Richens, Tom Everitt

## Introduction

**Top Goals:**

- Determine the behaviour of AI agents out-of-distribution.
- Define Safety specifications, e.g. harm, fairness, with respect to an AI agent's subjective causal model.
- Exploit the relationship between our observations of AI behaviour and their subjective causal model to predict agent intentions and Safety specifications.

**Main Outcomes:**

- A demonstration that we can partially predict agent behaviour given only observations.
- Partial predictions are given in terms of bounds as a function of input data that are shown to be tight.
- This sets the limits of what can be predicted about AI agents with a purely data-driven approach, without introducing assumptions beyond the fact that the AI agent is "competent".
- Encouraging causal modelling as an approach to AI Safety for future work.

## What is an AI agent's subjective causal model ?

A Structural Causal Model $\widehat{\mathcal{M}} : \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P \rangle$ that describes the agent's choice of policy in different scenarios.

The agent's choice of policy is given by,

$$\arg\min_{\pi \in \Pi} \mathbb{E}_P[Y_\pi ; \widehat{\mathcal{M}}_\sigma]$$

where is $\sigma$ a "shift" in the environment, e.g., atomic intervention.

**When do agents learn world models?**

- If they are able to adapt to a sufficiently large set of distribution shifts (Richens and Everitt, 2024).
- If they have rational preferences over interventions (Piermont and Halpern, 2024) or,
- If they satisfy a regret bound for a sufficiently diverse set of simple goal-directed tasks Richens et al., 2025).

## The Limits of Predicting Agents from Behaviour

**Input**

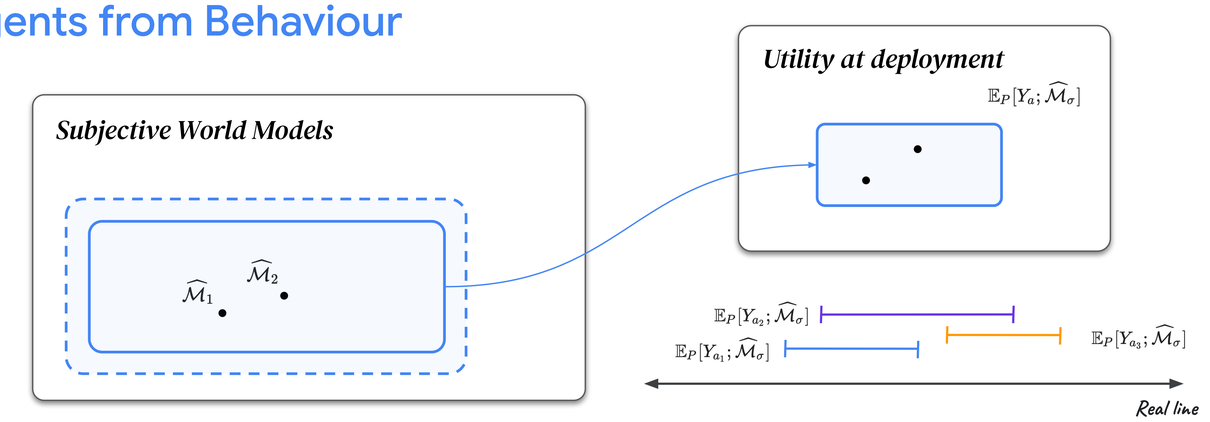
- Observations from one or more environments $(\mathbf{x}, a, y) \sim P(\cdot\,; \mathcal{M}_\pi)$

**Assumptions**

- Competence $P(\cdot\,; \widehat{\mathcal{M}}_\pi) \approx P(\cdot\,; \mathcal{M}_\pi)$

**Output**

- Tight bounds on the difference of expected utilities between competing actions (or other relevant quantity, e.g. harm, fairness specification)




## Results

An AI is **weakly predictable** under a given shift and context if there exists an action that is *provably sub-optimal*.

If the shift is a fixed / atomic intervention, say $do(\mathbf{Z} = \mathbf{z})$, and the context is $\mathbf{C} = \mathbf{c}$ this happens *if and only if* there exists a pair of actions $(a, a^*)$,

$$\frac{\mathbb{E}_{P_a}[\,Y \mid \mathbf{c}, \mathbf{z}\,]P_a(\mathbf{c}, \mathbf{z})}{P_a(\mathbf{c}, \mathbf{z}) + 1 - P_a(\mathbf{z})} - \frac{\mathbb{E}_{P_{a^*}}[\,Y \mid \mathbf{c}, \mathbf{z}\,]P_{a^*}(\mathbf{c}, \mathbf{z}) + 1 - P_{a^*}(\mathbf{z})}{P_{a^*}(\mathbf{c}, \mathbf{z}) + 1 - P_{a^*}(\mathbf{z})} > 0$$


This **recipe** can be used to derive conditions for weak predictability in terms of the observed data for,

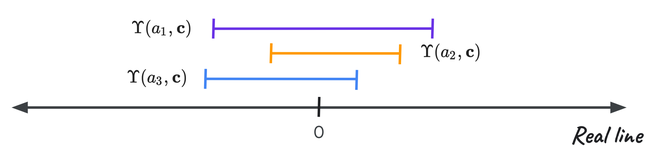
- Atomic interventions in the environment
- Arbitrary shifts in distribution of some variables
- Known shifts in distribution of some variables
- Multiple datasets from different environments

And also to reason about more complex beliefs of AI agents such as perception of **harm** and **fairness**.

**Harm** and **fairness** are often defined as counterfactual probabilities.

For example, an AI agent intends to be counterfactually fair if,

$$\Upsilon(a, \mathbf{c}) := \mathbb{E}_P\left[\, Y_{a,z_1} \mid z_0, \mathbf{c}; \widehat{\mathcal{M}} \,\right] - \mathbb{E}_P\left[\, Y_a \mid z_0, \mathbf{c}; \widehat{\mathcal{M}} \,\right] = 0$$




## Practical Relaxations

- What if the AI isn't perfectly competent? We can introduce a notion of **approximate competence** to relax this assumption resulting in looser bounds.
- What is the AI isn't perfectly rational? The agent might not always choose the action with the highest expected utility. We can assume instead a form of **"bounded" rationality**.
- What if the AI mis-interprets the goal of the problem? We could relax our analysis to account for proxy reward functions that are correlated, but distinct, from the observed reward data – **approximate inner alignment**.
- What if we could introduce prior knowledge on the AI's decision making process? With additional **structural assumptions,** e.g. access to a causal diagram, we could further improve our bounds.

## References


- Drago Plecko, Elias Bareinboim, et al. Causal fairness analysis: a causal toolkit for fair machine learning. Foundations and Trends® in Machine Learning, 17(3):304–589, 2024.
- Jonathan Richens and Tom Everitt. Robust agents learn causal world models. arXiv preprint arXiv:2402.10877, 2024.
- Jonathan Richens, Rory Beard, and Daniel H Thompson. Counterfactual harm. Advances in Neural Information Processing Systems, 35:36350–36365, 2022.
- Joseph Y Halpern and Evan Piermont. Subjective causality. arXiv preprint arXiv:2401.10937, 2024.