
The Limits of Predicting Agents from Behaviour

Alexis Bellot¹ Jonathan Richens¹ Tom Everitt¹

Abstract

As the complexity of AI systems and their interactions with the world increases, generating explanations for their behaviour is important for safely deploying AI. For agents, the most natural abstractions for predicting behaviour attribute beliefs, intentions and goals to the system. If an agent behaves as if it has a certain goal or belief, then we can make reasonable predictions about how it will behave in novel situations, including those where comprehensive safety evaluations are untenable. How well can we infer an agent’s beliefs from their behaviour, and how reliably can these inferred beliefs predict the agent’s behaviour in novel situations? We provide a precise answer to this question under the assumption that the agent’s behaviour is guided by a world model. Our contribution is the derivation of novel bounds on the agent’s behaviour in new (unseen) deployment environments, which represent a theoretical limit for predicting intentional agents from behavioural data alone. We discuss the implications of these results for several research areas including fairness and safety.

1. Introduction

Humans understand each other through the use of abstractions. We explain our intentions by appealing to our “goals” and “beliefs” about the world around us without knowing the underlying cognition going on inside our heads. According to Dennett (1989; 2017) the same is true of our understanding of other systems. For example, a bear hibernates during winter *as if* it believes that the lower temperatures cause food scarcity. This is a useful description of the bear’s behaviour, with real predictive power. For example, it gives us (human observers) the ability to anticipate how bears might act as the climate changes. There is a correspondence be-

tween beliefs and behaviour that is foundational to rational agents (Davidson, 1963).

Artificial Intelligence (AI) systems appear to have similarly general capabilities, not totally unlike that of humans and animals. They can generate text that is fluent and accurate in response to a very diverse set of questions. Whenever they display consistent types of behaviour across many different tasks, we are tempted to apply our own mentalistic language more or less at face value (Shanahan, 2024), taking seriously questions such as: What do the AIs know? What do they think, and believe? Taking the analogy further, it is *as if* they learn “world models” that mirror the causal relationships of the environment they are trained on, guiding their future plans and behaviour¹. And as a consequence, their interactions with an environment will leave clues that might give us the ability to predict their future behaviour in novel domains. This possibility engages with a core AI Safety problem: how to guarantee and predict whether AI systems will act safely and beneficially?

The main result of this paper is to offer a new perspective on this problem by showing that:

*With an assumption of competence and optimality,
the behaviour of AI systems partially determines
their actions in novel environments.*

Here behaviour means our observations of the decisions made by the AI system, contextual variables, and utility or reward values in some environment. The “partial” determination of actions in new environments is a consequence of our lack of knowledge about the AI’s actual world model (different models may induce different optimal actions). However, even though we can’t uniquely identify the AI’s future behaviour and beliefs, we can narrow it down to a range of possible outcomes. This paper characterises those outcomes.

¹Recent research suggests that an AI’s behaviour, to the extent that it is consistent with rationality axioms, can be formally described by a (causal) world model (Halpern & Piermont, 2024). The same conclusion can also be obtained for AIs capable of solving tasks in multiple environments (Richens & Everitt, 2024). For large language models, there is increasing empirical evidence for the “world model” hypothesis, see e.g., Toshniwal et al., 2022; Li et al., 2022; Gurnee & Tegmark, 2023; Goldstein & Levinstein, 2024 and Vafa et al., 2024.

¹Google DeepMind. Correspondence to: Alexis Bellot <abelot@google.com>.

In the literature, the under-determination of agent “beliefs” and “preferences” has been considered in the fields of inverse reinforcement learning (Abbeel & Ng, 2004; Skalse & Abate, 2023; Amin & Singh, 2016) and decision theory (Savage, 1972; Afriat, 1967; Jeffrey, 1990), among others. In settings with distribution shift between training and deployment environments, this under-determination can be understood as a consequence of the Causal Hierarchy Theorem, that defines precise limits on the kinds of inferences that can be drawn across domains (Bareinboim et al., 2022; Pearl, 2009). It implies, for example, that behaviour in an environment subject to an intervention cannot be established from “non-interventional” data alone. Robins (1989), Manski (1990) and Pearl (1999) showed that useful information in the form of bounds can nevertheless be extracted from “non-interventional” data, without actually knowing the underlying data-generating process. In the causality literature, several methods and algorithms exist to solve different versions of this problem, see e.g., Balke & Pearl, 1997; Tian & Pearl, 2000; Zhang et al., 2021; Bellot, 2024; Rosenbaum et al., 2010; Tan, 2006.

This paper extends the causal formalism to reason about the possible behaviours and beliefs of an AI system, itself assumed to be governed by an unknown data generating process or world model. With this interpretation we are able to define mathematically notions such as an AI’s preferred choice of action in novel environments, its perception of fairness, and its perception of harm due to the actions it takes. Our main contribution is a set of inequalities on these “beliefs” in terms of quantities that can in principle be estimated from behavioural data, and that hold irrespective of the underlying cognitive architecture of the AI system as long as it can be *represented* by a well-defined set of causal mechanisms (a world model) that tracks its behaviour (Sec. 4). We then extend these results to characterize AI behaviour under several relaxations for applications in practice (Sec. 5), ultimately with the goal of defining the theoretical limits of what can be inferred from data about AI behaviour in new (unseen) environments.

This has consequences for the wider AI Safety community and society. For example, we show that an AI’s perception of the potential fairness and harm of its decisions (e.g., whether the AI’s resource allocation is believed to be equitable, or its generations unbiased) can provably not be inferred from observing its behaviour alone. There are theoretical limits to how much we can understand about an AI’s cognition and decision-making process from observations. We believe our results can help justify the claim that the design and inference of world models is important to ensure AIs can behave predictably and act safely and beneficially, as argued by Dalrymple et al., 2024; Legg, 2023; Bengio et al., 2025.

2. Preliminaries

In this section we outline some basic principles that we use to reason about how beliefs might be (implicitly) defined within an AI system.

We use capital letters to denote variables (X), small letters for their values (x), bold letters for sets of variables ($\mathbf{X}$) and their values ($\mathbf{x}$), and use supp to denote their domains of definition ($x \in \text{supp}_X$). To denote $P(\mathbf{Y} = \mathbf{y} \mid \mathbf{X} = \mathbf{x})$, we sometimes use the shorthand $P(\mathbf{y} \mid \mathbf{x})$. We use $\mathbb{1}_{\{\cdot\}}$ for the indicator function equal to 1 if the statement in $\{\cdot\}$ evaluates to true, and equal to 0 otherwise.

Actions, plans, and hypothetical outcomes can be evaluated by symbolic operations on a model that represents the functional relationships in the world, known as a Structural Causal Model (Pearl, 2009, Definition 7.1.1), or SCM for short.

Definition 1 (Structural Causal Model). *An SCM $\mathcal{M}$ is a tuple $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P \rangle$ where each observed variable $V \in \mathbf{V}$ is a deterministic function of a subset of variables $\mathbf{Pa}_V \subset \mathbf{V}$ and latent variables $\mathbf{U}_V \subset \mathbf{U}$, i.e., $v := f_V(\mathbf{pa}_V, \mathbf{u}_V)$, $f_V \in \mathcal{F}$. Each latent variable $U \in \mathbf{U}$ is distributed according to a probability measure $P(u)$. We assume the model to be recursive, i.e., that there are no cyclic dependencies among the variables.*

In an SCM $\mathcal{M}$, each draw $\mathbf{u} \sim P(\mathbf{u})$ evaluates to a potential response $\mathbf{Y}(\mathbf{u}) = \mathbf{y}$ and entails a distribution over the possible outcomes $P(\mathbf{y})$. The power of SCMs is that they specify not only the joint distribution $P(\mathbf{v})$ but also the distribution of variables under all interventions, including incompatible interventions (counterfactuals). Formally, an intervention $do(\mathbf{x})$ is modelled as a symbolic operation where values of a set of variables $\mathbf{X}$ are set to constants $\mathbf{x}$, replacing the functions $\{f_X : X \in \mathbf{X}\}$ that would normally determine their values. This effectively induces a sub-model of $\mathcal{M}$, denoted $\mathcal{M}_{\mathbf{x}}$. The variables obtained in $\mathcal{M}_{\mathbf{x}}$ are denoted $\mathbf{Y}_{\mathbf{x}}$ and we will loosely write $P^{\mathcal{M}_{\mathbf{x}}}(\mathbf{y}) \equiv P_{\mathbf{x}}(\mathbf{y}) \equiv P(\mathbf{y}_{\mathbf{x}}) \equiv P(\mathbf{y} \mid do(\mathbf{x}))$ to denote the probabilities over the possible outcomes of $\mathbf{Y}$ in $\mathcal{M}_{\mathbf{x}}$.

Different environments can be modelled by different SCMs. Let $\mathcal{M}_1 = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}^1, P^1 \rangle$, $\mathcal{M}_2 = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}^2, P^2 \rangle$ be the SCMs for two environments over the same set $\mathbf{V}$ and $\mathbf{U}$. We say that there is a discrepancy or a shift on a variable $X \in \mathbf{V}$ between them if either $f_X^1 \neq f_X^2$ or $P^1(U_X) \neq P^2(U_X)$ or both. Shifts might therefore encode arbitrary changes in the causal mechanisms for a set of variables. For a reference SCM $\mathcal{M}$, a so-called “shifted” SCM will be represented by a sub-model $\mathcal{M}_{\sigma}$ where σ represents the discrepancies between $\mathcal{M}$ and $\mathcal{M}_{\sigma}$. For example, an environment with a shift σ on a set of variables $\mathbf{X}$ introduces (possibly arbitrary) discrepancies in the functional assignment or (independent) exogenous variables of $\mathbf{X}$ while keeping other mechanisms

unchanged. See Pearl (2009, Chapter 4) and Correa & Bareinboim (2020b) for more details.

We make a note here that all proofs of statements are given in Appendix C and that the derivations of examples are given in Appendix A.

3. Agents, Beliefs, and the Environment

In this section we lay out a framework to interface between the AI system’s internal world model and our own observations of their behaviour in the real world. Both rely on the same SCM abstraction.

We assume that the AI operates according to an SCM $\widehat{\mathcal{M}}$ over $\mathbf{V}$, its (implicit) world model², that guides its behaviour. $\mathbf{V}$ includes the AI’s decision variable D , the inputs to those decisions $\mathbf{C}$, possible additional variables, and the utility variable Y , such as the training signal or a measurable target given to the AI (Everitt et al., 2021). Beliefs³ are defined as quantifiable aspects of that model or derivations of it.

Definition 2 (Beliefs). *An AI belief is a probabilistic statement derived from its internal model $\widehat{\mathcal{M}}$.*

For example, a statement like $P^{\widehat{\mathcal{M}}_d}(Y = y) = 0.8$ describes the subjective belief “The AI is 80% confident that taking decision $D = d$ will lead to event $Y = y$ ”. The sub-model in this mathematical expression represents what the AI “thinks” the world looks like after taking the decision $D = d$.

We assume that the AI makes decisions d by sampling from a policy $\pi(d \mid \mathbf{c})$, which is a function mapping from the domain of the observed covariates $\mathbf{C} \subset \mathbf{V}$ (i.e., all the inputs given to the AI) to the probability space over the domain of the decision $D \in \mathbf{V}$. The choice of π is assumed to be driven by its perceived utility⁴ $Y \in \mathbf{V}$ within the AI’s model $\widehat{\mathcal{M}}$, that is,

$$\arg \max_{\pi} \mathbb{E}_{P^{\widehat{\mathcal{M}}}} [Y \mid do(\pi)]. \quad (1)$$

The AI interacts with the real-world that is described by a (likely different) SCM $\mathcal{M}$ that encodes the true dynamics

²Here SCMs are meant to *represent*, mathematically, the decision-making process going on “in the AI’s head” in a way that tracks its behaviour, without making any claims about the AI’s actual cognitive architecture.

³We might prefer to use terms like “credences” or “subjective probabilities” to emphasize the subjective nature of beliefs and avoid the connotation of strong conviction or certainty as done by (Schwitzgebel, 2024, Sec. 2.3).

⁴To account for possible uncertainty in the AI’s “satisfaction” about a given state of the world $\mathbf{w}$ we assume Y is a random variable (induced by $\mathbf{U}_Y \subset \mathbf{U}$), also known as a stochastic utility model (Manski, 1977). We assume that the support of Y is bounded in the $[0, 1]$ interval.

of the environment. In principle, we have no reason to expect that the model $\widehat{\mathcal{M}}$ internalized by the AI matches the underlying reality $\mathcal{M}$. AI systems might hope to reproduce some aspects of $\mathcal{M}$ (the AI might have learned, for instance, to mimic the distribution of the observed data). Competent AIs might go further and be able to reliably predict the effects of different decisions in the world. We define this as *grounding* below.

Definition 3 (Grounding). *Let $\widehat{\mathcal{M}}$ represent the AI’s internal model. We say that the AI is grounded in a domain $\mathcal{M}$ if $P^{\widehat{\mathcal{M}}_d}(\mathbf{V}) \equiv P^{\mathcal{M}_d}(\mathbf{V})$ for any decision $d \in \text{supp}_D$.*

Grounding tells us that the AI’s beliefs about the effect of a particular decision d in the training environment match the effects that would be observed in the real world, i.e. $\widehat{P}_d(\mathbf{V}) \equiv P_d(\mathbf{V})$ ⁵. It is an **assumption** on the relationship between our observations of AI behaviour $P(\mathbf{V})$ with what might be going on in the AI’s “mind” $\widehat{P}(\mathbf{V})$. This might be reasonable, for example, if the AI is explicitly trained by reinforcement learning in $\mathcal{M}$.

By assumption, a grounded AI’s choice of decision in environment $\mathcal{M}$ is in principle predictable from data since we can compute Eq. (1) uniquely. But this might not necessarily be the case in a new (unseen) environment.

Example 1 (The Uncertain Medical AI). Imagine an AI system assisting patients with their treatment D for a disease Y known to be influenced also by a third variable Z , blood pressure. The AI is competent and learns the precise effect of all treatments. In other words it is grounded in $\mathcal{M}$, i.e. $\widehat{P}_d(z, y) = P_d(z, y)$. For concreteness, let the environment $\mathcal{M}$ be given by,

$$\begin{aligned} Z &\leftarrow \mathbb{1}_{U=1 \text{ or } 4}, \\ Y &\leftarrow \begin{cases} Z \cdot \mathbb{1}_{U=4} + (1 - Z) \cdot \mathbb{1}_{U=1,3 \text{ or } 4} & \text{if } d = 0 \\ Z \cdot \mathbb{1}_{U \neq 2} + (1 - Z) \cdot \mathbb{1}_{U=2 \text{ or } 4} & \text{if } d = 1, \end{cases} \end{aligned}$$

with equal probability P for all values $U \in \{1, 2, 3, 4, 5\}$. Here U is latent, summarizing all other contributions to both the disease and blood pressure, such as an individual’s (unobserved) attitudes to health, fitness, etc. Could we confidently deploy this AI system more widely, for example, on individuals that also take a second drug that artificially improves their blood pressure (e.g., fixing Z to 1, replacing the original assignment)? If the AI system is instructed to maximize Y on average, what decision does the AI believe is optimal in this new environment? The answer is we do not know, meaning that it is possible to find a second model

⁵We use the shorthand $P_d \equiv P^{\mathcal{M}_d}$ and $\widehat{P}_d \equiv P^{\widehat{\mathcal{M}}_d}$ to simplify the notation.

$\widehat{\mathcal{M}}$ defined by the mechanisms:

$$\begin{aligned} Z &\leftarrow \mathbb{1}_{U=1 \text{ or } 4}, \\ Y &\leftarrow \begin{cases} Z \cdot \mathbb{1}_{U \neq 1} + (1 - Z) \cdot \mathbb{1}_{U=3 \text{ or } 4} & \text{if } d = 0 \\ Z \cdot \mathbb{1}_{U=1 \text{ or } 4} + (1 - Z) \cdot \mathbb{1}_{U=1 \text{ or } 2} & \text{if } d = 1, \end{cases} \end{aligned}$$

that entails exactly the same observations $\widehat{P}_d(z, y) = P_d(z, y)$ but induces different optimal decisions in the new environment (under the intervention $Z \leftarrow 1$). Under $\mathcal{M}_{Z \leftarrow 1}$, the highest utility Y on average is given by $d = 1$, while under $\widehat{\mathcal{M}}_{Z \leftarrow 1}$ the highest utility Y on average is given by $d = 0$. A priori, we have no way of knowing which model ($\mathcal{M}$ or $\widehat{\mathcal{M}}$) is governing the AI's behaviour and so no way of knowing what decision will be favoured by the AI. $\square$

This example illustrates a canonical point in a simple setting: as observers, with access to the AI's interactions in some domain, its behaviour outside of that domain might not be uniquely determined (Pearl, 2009).

4. The Limits of Behavioural Data

In this section, we explore the limits of behavioural data for predicting the decisions of AIs in new environments.

As external observers, we do not have access to the mechanisms underlying the actual environment nor the agent's internal model. We assume that we must rely for our inferences on watching the agent's behaviour and its consequences. That is we have access to (samples of) $P_d(\mathbf{V})$ ⁶ for all d . As a starting point, we might expect competent AIs to be *weakly* predictable in the sense that a subset of decisions can be ruled out as provably sub-optimal given our observations.

Definition 4 (Weak Predictability). *We say that an AI is weakly predictable under a shift σ in situation $\mathbf{C} = \mathbf{c}$ if there exists a decision d^* that is provably sub-optimal, i.e.,*

$$d^* \neq \arg \max_d \mathbb{E}_{P_{\widehat{\mathcal{M}}}} [Y \mid do(\sigma, d), \mathbf{c}], \quad (2)$$

for any valid SCM $\widehat{\mathcal{M}}$ describing the AI's internal model.

Here, "valid" means that the AI's internal model is compatible with the observed data under our assumptions about the relationship between the data and the AI's internal model,

⁶Technically, the AI system may choose to follow an arbitrarily complex policy π in the training domain, inducing a (assumed positive) distribution $P_\pi(\mathbf{v})$. It holds that $P_d(\mathbf{V})$ can be computed from any such $P_\pi(\mathbf{V})$ as long as $P_\pi(\mathbf{v}) > 0, \forall \mathbf{v}$, and vice versa, see e.g. Lem. 1. The positivity assumption $P_d(\mathbf{v}) > 0$ rules out fully deterministic policies in the available data but might be reasonable if the AI spends some time exploring before committing to a course of action.

e.g., grounding. Weak predictability means that there exists at least one decision that we can guarantee the AI will not take in the shifted environment. Specifically, we can rule out a decision d^* if and only if we can find a (superior) alternative decision $d \neq d^*$ such that,

$$\min_{\widehat{\mathcal{M}} \in \mathbb{M}} (\Delta_{d > d^*}) > 0, \quad (3)$$

where,

$$\Delta_{d > d^*} := \mathbb{E}_{P_{\widehat{\mathcal{M}}}} [Y \mid do(\sigma, d), \mathbf{c}] - \mathbb{E}_{P_{\widehat{\mathcal{M}}}} [Y \mid do(\sigma, d^*), \mathbf{c}].$$

$\mathbb{M}$ denotes the set of "valid" SCMs. Here Δ can be interpreted as the AI's **preference gap** between two decisions in some situation $\mathbf{C} = \mathbf{c}$. When it evaluates to a positive number d is preferred to d^* and when it evaluates to a negative number d^* is preferred to d (in the AI's mind). If our inferences on Δ allow us to rule out decisions d^* considered to be "unsafe" then weak predictability gives us an important safety guarantee.

We can strengthen this notion to define *strong* predictability, that describes a situation in which all but a single AI decision can be ruled out.

Definition 5 (Strong Predictability). *We say that an AI is strongly predictable under a shift σ in situation $\mathbf{C} = \mathbf{c}$ if the optimal decision is uniquely identifiable, i.e., there exists a single decision d^* such that,*

$$d^* = \arg \max_d \mathbb{E}_{P_{\widehat{\mathcal{M}}}} [Y \mid do(\sigma, d), \mathbf{c}], \quad (4)$$

for any valid SCM $\widehat{\mathcal{M}}$ describing the AI's internal model.

4.1. AI decisions out-of-domain: interventions

Our first result shows that, in some cases, a subset of AI decisions can be provably ruled out, i.e., the AI is weakly predictable.

Theorem 1. *An AI grounded in a domain $\mathcal{M}$ is weakly predictable under a shift $\sigma := do(\mathbf{z}), \mathbf{Z} \subset \mathbf{V}$, in a context $\mathbf{C} = \mathbf{c}$ if and only if there exists a decision d^* such that,*

$$\begin{aligned} &\frac{\mathbb{E}_{P_d} [Y \mid \mathbf{c}, \mathbf{z}] P_d(\mathbf{c}, \mathbf{z})}{P_d(\mathbf{c}, \mathbf{z}) + 1 - P_d(\mathbf{z})} \\ &- \frac{\mathbb{E}_{P_{d^*}} [Y \mid \mathbf{c}, \mathbf{z}] P_{d^*}(\mathbf{c}, \mathbf{z}) + 1 - P_{d^*}(\mathbf{z})}{P_{d^*}(\mathbf{c}, \mathbf{z}) + 1 - P_{d^*}(\mathbf{z})} > 0, \end{aligned} \quad (5)$$

for some $d \neq d^*$.

All terms on the l.h.s are in principle computable from the AI's behaviour. Loosely speaking, the value of this difference is determined (in part) by " $P_{d^*}(\mathbf{z})$ ": if $\mathbf{Z} = \mathbf{z}$ (the value set by the intervention) is likely under the training distribution, the difference will more likely evaluate to a

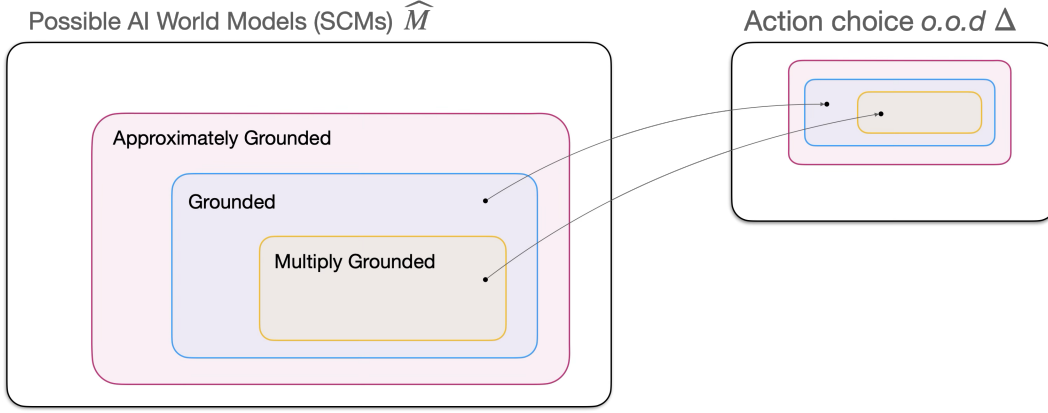


Figure 1. Grounding and observations in multiple environments constrains the AI’s world model and improves our prediction of AI behaviour out-of-distribution (o.o.d). Approximate grounding is defined in Sec. 5.

positive value. The “if and only if” condition means that whenever this inequality does not hold we can construct two SCMs $\hat{\mathcal{M}}_1, \hat{\mathcal{M}}_2$ for the grounded AI’s internal model that generate the observed behaviour $P_d(\mathbf{V}), d \in \text{supp}_D$, but that induce different optimal actions. That is, for all $d \neq d^*$,

$$\begin{aligned} \mathbb{E}_{P_{\hat{\mathcal{M}}_1}} [Y | do(\mathbf{z}, d), \mathbf{c}] &> \mathbb{E}_{P_{\hat{\mathcal{M}}_1}} [Y | do(\mathbf{z}, d^*), \mathbf{c}], \\ \mathbb{E}_{P_{\hat{\mathcal{M}}_2}} [Y | do(\mathbf{z}, d), \mathbf{c}] &< \mathbb{E}_{P_{\hat{\mathcal{M}}_2}} [Y | do(\mathbf{z}, d^*), \mathbf{c}]. \end{aligned}$$

Remark. We can derive a similar condition for strongly predictable AIs by replacing “for some $d \neq d^*$ ” with “for all $d \neq d^*$ ” in Thm. 1.

We illustrate Thm. 1 with the following example.

Example 2 (Grounded Medical AI). In Example 1, we have shown that there exists a particular intervened environment in which the AI’s intentions cannot be determined as in principle the AI could believe that either decision is optimal. Is this true in general? Thm. 1 suggests that it depends on the likelihood of different events $P_d(z, y)$ in the observed data. For Example 1, we can show that the medical AI is not weakly predictable as the expression in Thm. 1 is negative for all pairs of decisions. In other words, no decision can be ruled out in general: in some AI internal models d_1 is inferior to d_0 as $\min_{\hat{\mathcal{M}} \in \mathbb{M}} (\Delta_{d_1 > d_0}) = P_{d_1}(Z = z, Y = 1) + P_{d_0}(Z = z, Y = 0) - 1 = -0.4$ while in others d_0 is inferior to d_1 as $\min_{\hat{\mathcal{M}} \in \mathbb{M}} (\Delta_{d_0 > d_1}) = P_{d_1}(Z = z, Y = 0) + P_{d_0}(Z = z, Y = 1) - 1 = -0.8$ and we don’t know which one the AI system has internalised. $\square$

In this example, AI behaviour does provide *some* information as it can be constrained to larger values than its a priori minimum $\Delta = -1$, but not enough to rule out a decision completely. Our next result shows that Thm. 1 could be extended to get tight bounds for AI systems that are grounded in multiple environments.

Theorem 2. Let $\sigma := do(\mathbf{z})$ be a shift on a set of variables $\mathbf{Z} \subset \mathbf{V}$. For $\mathbf{R}_i \subset \mathbf{Z} \subset \mathbf{V}, i = 1, \dots, k$, consider an AI grounded in multiple domains $\{\mathcal{M}_{\mathbf{r}_i} : i = 1, \dots, k\}$. The AI is weakly predictable in a context $\mathbf{C} = \mathbf{c}$ under a shift $\sigma := do(\mathbf{z})$ if and only if there exists a decision d^* such that,

$$\max_{i,j=1,\dots,k} A(\mathbf{r}_i, \mathbf{r}_j) > 0, \quad \text{for some } d \neq d^*, \quad (6)$$

where

$$A(\mathbf{r}_i, \mathbf{r}_j) := \frac{\mathbb{E}_{P_{d,\mathbf{r}_i}} [Y | \mathbf{c}, \mathbf{z} \setminus \mathbf{r}_i] P_{d,\mathbf{r}_i}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}_i) - \mathbb{E}_{P_{d^*,\mathbf{r}_j}} [Y | \mathbf{c}, \mathbf{z} \setminus \mathbf{r}_j] P_{d^*,\mathbf{r}_j}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}_j) + 1 - P_{d,\mathbf{r}_i}(\mathbf{z} \setminus \mathbf{r}_i) - P_{d^*,\mathbf{r}_j}(\mathbf{z} \setminus \mathbf{r}_j)}{P_{d,\mathbf{r}_i}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}_i) + 1 - P_{d,\mathbf{r}_i}(\mathbf{z} \setminus \mathbf{r}_i) - P_{d^*,\mathbf{r}_j}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}_j) + 1 - P_{d^*,\mathbf{r}_j}(\mathbf{z} \setminus \mathbf{r}_j)}.$$

In this result, $\{\mathcal{M}_{\mathbf{r}_i} : i = 1, \dots, k\}$ describes k domains in which experiments on different subsets of $\mathbf{Z}$ have been conducted, i.e., $\{P_{d,\mathbf{r}_i}(\mathbf{V}) : i = 1, \dots, k\}$ is available. This includes possibly the null experiment $\mathbf{R}_i = \emptyset$ that refers to the unaltered domain $\mathcal{M}$. Note that grounding in multiple domains is useful for the prediction of the AI’s preference gap because the resulting bounds in Thm. 2 are tighter than those in Thm. 1 (this is given formally as Corol. 3 in the Appendix).

Fig. 1 illustrates how different assumptions and observations give us information about the possible world models that the AI is operating on, which then has implications for the AI’s behaviour out-of-distribution. This knowledge allows us to reduce the uncertainty around the AI’s preference gap Δ , and possibly rule out certain actions that are unambiguously sub-optimal out-of-distribution, inferred solely from observed behaviour.

4.2. AI decisions out-of-domain: general shifts

We might wonder about predictability under more general shifts such as an arbitrary change in a subset of the mechanisms $\{f_Z : Z \in \mathbf{Z}\}$ and distribution of variables $\{U_Z, Z \in \mathbf{Z}\}$ in $\mathcal{M}$. For example, in practice we are likely able to convey to the AI that the mechanisms for a set of variables $\mathbf{Z}$ are expected to change but not know exactly how. For example, demographic properties of patients might change across hospitals. How could the AI interpret the consequences of such an under-specified shift? To begin to answer this question, the following theorem shows that in the extreme case where the nature of the shift is completely unknown the AI’s preference gap is unconstrained.

Theorem 3. *Consider an AI grounded in a domain $\mathcal{M}$ made aware of an (under-specified) shift on non-empty $\mathbf{Z} \subset \mathbf{V}$. Then the AI is provably not weakly (or strongly) predictable in any context $\mathbf{C} = \mathbf{c}$.*

This result means that no decision could ever be ruled out from AI behaviour. We can show moreover that $\min_{\widehat{\mathcal{M}} \in \mathbb{M}} (\Delta) = -1$ for any pair of decisions, meaning that the observed data (no matter what it is) gives us no information on AI decision-making.

In practice, however, it might be realistic to have access to some information in the shifted environment, such as covariate data, i.e., (samples from) $P_{\sigma,d}(\mathbf{c})$, that could be given to the AI for it to update its internal model accordingly (with some abuse of terminology we say that the AI is grounded in $P_{\sigma,d}(\mathbf{c})$). The next theorem shows that this additional information coupled with the AI’s behaviour makes the AI more predictable.

Theorem 4. *Consider an AI grounded in a domain $\mathcal{M}$ and $P_{\sigma,d}(\mathbf{C})$ made aware of a shift σ on $\mathbf{Z} \subset \mathbf{C}$. The AI is weakly predictable under this shift in a context $\mathbf{C} = \mathbf{c}$ if there exists a decision d^* such that,*

$$1 - \frac{2 + \mathbb{E}_{P_{d^*}}[Y | \mathbf{c}] P_{d^*}(\mathbf{c}) - \mathbb{E}_{P_d}[Y | \mathbf{c}] P_d(\mathbf{c})}{P_{\sigma,d^*}(\mathbf{c})} + \frac{P_d(\mathbf{c}) - 2P_d(\mathbf{z})}{P_{\sigma,d^*}(\mathbf{c})} > 0, \quad \text{for some } d \neq d^*.$$

This bound is not tight in general, however, meaning that it is possible that the AI is actually predictable in settings where Thm. 4 suggests it might not be.

Example 3 (Shifted Medical AI). The AI from Example 2, originally developed from data primarily from young patients, is now considered for deployment on an older patient population. Their probability of having high blood pressure $P_{\sigma}(Z = 1) = 0.9$ is known to be substantially higher than that observed during training $P(Z = 1) = 0.4$: there is a shift in the underlying mechanisms of $\mathbf{Z}$. How do these changes influence the AI’s beliefs on Δ ? Thm. 4

suggests that the medical AI might not be weakly predictable as the expression evaluates to a negative value for all pairs of decisions. The lower bounds on the AI preference gap are given by $\min_{\widehat{\mathcal{M}} \in \mathbb{M}} (\Delta_{d_1 > d_0}) \geq -0.55$ and $\min_{\widehat{\mathcal{M}} \in \mathbb{M}} (\Delta_{d_0 > d_1}) \geq -1$. That is, no decision is always inferior to any other decision. $\square$

4.3. AI’s perceived fairness of decisions

An AI’s policy, even if optimal on average, has the potential to bring about a state of the world that is intrinsically harmful or unfair. Harm and fairness can be defined relative to a causal model (Beckers et al., 2022; Plecko et al., 2024). This means that a notion of *perceived* or *subjective* harm and fairness could be attributed to AI systems that operate according to an (implicit) causal model. As a consequence, it is conceivable that AIs could be held morally accountable for the harm and unfairness that they cause. How might one estimate the AI’s beliefs about the harm and unfairness that its decisions cause?

To ground our discussion, we consider here explicitly *counterfactual* accounts of fairness and harm. These appeal to hypothetical situations, imagining “what might have been if ...”, that can force us to confront our assumptions and values in a way that our regular thought processes might not⁷. For example, the counterfactual event $(Y_x = 1 \mid X = x_0)$ refers to the outcome $(Y = 1)$ under an intervention $X = x$ when under normal circumstances X would have evaluated to x_0 . In the literature, probabilities over counterfactuals emerge from the definition of an SCM. For a set of (counterfactual) events $(z_w, \dots, y_x)$,

$$P(z_w, \dots, y_x) = \int_{\mathbf{u}: Z_w(\mathbf{u})=z_w, \dots, Y_x(\mathbf{u})=y_x} P(\mathbf{u}). \quad (7)$$

Kusner et al. (2017) made a concrete proposal arguing that an AI’s decision is said to be fair towards an individual if, from the AI’s perspective, it entails the same utility in the actual world and in a counterfactual world where the individual belonged to a different group (defined by a sensitive attribute, e.g., gender, race). We adapt this notion to define an AI’s counterfactual fairness gap.

Definition 6 (Counterfactual Fairness Gap). *Let $Z \in \{z_0, z_1\}$ be a protected attribute and z_0 a baseline value of Z . For a given utility Y , define an AI’s counterfactual fairness gap relative to a decision d , in a given context $\mathbf{c}$, as*

$$\Upsilon(d, \mathbf{c}) := \mathbb{E}_{\widehat{P}}[Y_{d,z_1} \mid z_0, \mathbf{c}] - \mathbb{E}_{\widehat{P}}[Y_d \mid z_0, \mathbf{c}]. \quad (8)$$

⁷Alternative accounts to harm and fairness have been proposed (Barocas & Selbst, 2016; Zhang & Bareinboim, 2018; Plecko et al., 2024), sometimes motivated by scenarios where counterfactual accounts give incomplete results. For some of them, the AI’s beliefs can be shown to be similarly constrained by its external behaviour. We provide a longer discussion in Appendix D.

We say that an AI “intends” to be fair with respect to an attribute Z if under any context $C = c$ and decision $D = d$ the counterfactual fairness gap Υ evaluates to 0. This means that, under its own internal world model, changing the value of Z on the subset of situations with context c in which Z was observed to z_0 does not change the AI’s expected utility. In the following theorem we show that, unfortunately, the answer to this question is impossible to obtain given only the AI’s external behaviour.

Theorem 5. *Consider an agent with utility Y grounded in a domain $\mathcal{M}$. Then,*

$$-\mathbb{E}_{P_d}[Y \mid z_0, c] \leq \Upsilon \leq 1 - \mathbb{E}_{P_d}[Y \mid z_0, c]. \quad (9)$$

This bound is tight.

The bound is tight in the sense that for each context, decision, and baseline attribute, we can find compatible models for which the equalities hold. The counterfactual fairness gap Υ is under-constrained. Since $\Upsilon = 0$ is consistent with any external behaviour we can never conclude that the AI system “intends” to be unfair. Moreover, since the width of the bound is equal to 1, we can also never conclude that the AI is anywhere “close” to being fair, according to this counterfactual criterion.

4.4. AI’s perceived harm of decisions

Prominent definitions of harm are similarly counterfactual in nature: the *counterfactual comparative account of harm* defines a decision d to harm a person if and only if she would have been better off if d had not been taken (Hanser, 2008; Richens et al., 2022; Beckers et al., 2022; Mueller & Pearl, 2023). It is a contrast between events in hypothetical scenarios in which different decisions are made. Here, we quantify how “well off” a particular situation $W = w$ is with a binary utility variable $Y \leftarrow f_Y(W, U_Y) \in \{0, 1\}$ that we assume is tracked in experiments, i.e., $Y \in \mathcal{V}$. The following definition describes this notion of harm mathematically.

Definition 7 (Counterfactual Harm Gap). *Consider an AI with internal model $\hat{\mathcal{M}}$ and utility $Y \in \{0, 1\}$. The AI’s expected counterfactual harm of a decision d_1 with respect to a baseline d_0 , in context c , is*

$$\Omega(d_1, d_0, c) := \mathbb{E}_{\hat{P}}[\max\{0, Y_{d_0} - Y_{d_1}\} \mid c]. \quad (10)$$

Operationally, the counterfactual harm gap Ω is the expected increase in utility had the AI made a default decision d_0 , with respect to a different decision d_1 that the AI is contemplating. Counterfactual harm is therefore lower bounded at 0 with larger values indicating more harm. The following theorem shows that the external behaviour constraints the AI’s perception of its counterfactual harm.

Theorem 6. *Consider an AI with utility Y grounded in a domain $\mathcal{M}$. Then,*

$$\begin{aligned} \Omega &\geq \max\{0, \mathbb{E}_{P_{d_1}}[Y \mid c] + \mathbb{E}_{P_{d_0}}[Y \mid c] - 1\} \\ \Omega &\leq \min\{\mathbb{E}_{P_{d_1}}[Y \mid c], \mathbb{E}_{P_{d_0}}[Y \mid c]\} \end{aligned}$$

This bound is tight.

This result it is an extension of bounds on the probability of causation given by (Pearl, 1999) and (Tian & Pearl, 2000). It suggests that an AI’s beliefs about the harm that its decisions cause can be inferred approximately from data.

5. The “Practical” Limits of Behavioural Data

The inductive biases implied by causal models and rational behaviour are powerful constraints on AI behaviour. But they might not capture the practical limitations of AI decision-making. In this section we show that grounding, expected utility maximization, observed data, etc., can be relaxed in practice.

5.1. Approximate grounding

Grounding implies that the AI’s beliefs on the likelihood of events in the environment matches the observed probabilities. In practice, it might be reasonable to allow for some amount of error, and consider a notion of “approximate” grounding.

Definition 8 (Approximate Grounding). *Let $\hat{\mathcal{M}}$ represent the AI’s internal model. Given a discrepancy measure ψ , we say that the AI is approximately grounded in a domain $\mathcal{M}$ to a degree $\delta > 0$ if $\psi(\hat{P}_d, P_d) \leq \delta$ for any $d \in \text{supp}_D$.*

The choice of ψ and δ , in practice, depend on what error model is reasonable for the AI and problem at hand (we give an example below). Approximate grounding specifies a looser relationship between our observations of AI behaviour P with what might be going on in the AI’s “mind” $\hat{P}$. For example, the world model of an approximately grounded AI is compatible with one distribution in the set $\{\hat{P}_d : \psi(\hat{P}_d, P_d) \leq \delta\}$.

A more conservative bound (than Thm. 1) on predictability could be derived for AIs that are approximately grounded in an environment $\mathcal{M}$.

Corollary 1. *Given a discrepancy measure ψ , an AI approximately grounded in a domain $\mathcal{M}$ is weakly predictable in a context $C = c$ under a shift $\sigma := \text{do}(z)$, $Z \subset \mathcal{V}$, if and only if there exists a decision d^* such that,*

$$\min_{\hat{P}: \psi(\hat{P}, P) \leq \delta} \left\{ \frac{\mathbb{E}_{\hat{P}_d}[Y \mid c, z] \hat{P}_d(c, z)}{\hat{P}_d(c, z) + 1 - \hat{P}_d(z)} - \frac{\mathbb{E}_{\hat{P}_{d^*}}[Y \mid c, z] \hat{P}_{d^*}(c, z) + 1 - \hat{P}_{d^*}(z)}{\hat{P}_{d^*}(c, z) + 1 - \hat{P}_{d^*}(z)} \right\} > 0,$$

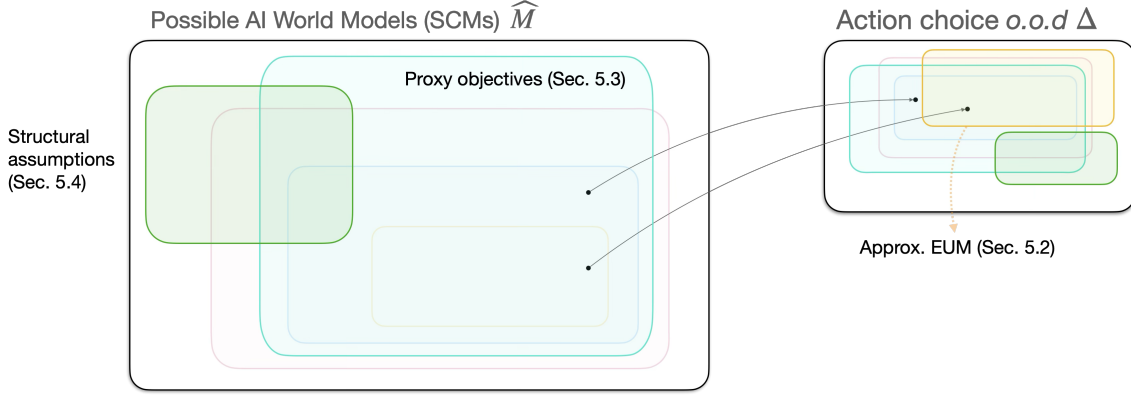


Figure 2. Building on Fig. 1, AIs that are approximate expected utility maximizers (EUM), that internalize proxy objectives, or that obey known causal structure carve out different constraints on the set of possible AI models (from an observer’s perspective) which may be exploited to improve our prediction of AI choices out-of-distribution (o.o.d).

for some $d \neq d^*$.

The strategy in Corol. 1 can be applied to all bounds on behaviour in Sec. 4 to get results under approximate grounding. We can compare quantitatively the two notions of grounding with an example.

Example 4 (Approximately Grounded Medical AI). The results in Example 2 exploit the grounding relationship $\hat{P}_d(\mathbf{V}) = P_d(\mathbf{V})$ in $\mathcal{M}$. We might want to relax the equality by assuming that the AI is instead *approximately grounded*. Minimum values on the AI’s preference gap $\min_{\widehat{\mathcal{M}} \in \mathbb{M}} (\Delta_{d_1 > d_0})$ would then be given by,

$$\min_{\hat{P}: \psi(\hat{P}, P) \leq \delta} \hat{P}_{d_1}(Z = z, Y = 1) + \hat{P}_{d_0}(Z = z, Y = 0) - 1.$$

These terms now capture an additional source of uncertainty due to external behaviour more loosely constraining $\widehat{\mathcal{M}}$. An empirical estimate of this quantity could be obtained by sampling distributions $\hat{P}$ close to P according to the distributional distance ψ and threshold δ , and taking the empirical minimum, as follows. Given that the data $(z, d, y) \sim P$ is discretely valued in this example, we could sample probabilities $\{\hat{P}_d(z, y)\}_{z, y}$ from a Dirichlet distribution centred at the vector $\{P_d(z, y)\}_{z, y}$ with a small variance. The distance of each proposal from the reference distribution could then be evaluated according to ψ and each proposal either accepted or rejected using δ . For illustration, we implement a version of this idea setting ψ to be the total variation distance and $\delta = 0.1$. The two minimum values now evaluate to -0.55 and -0.88 , respectively, which is slightly lower than under the assumption of grounding in Example 2 (that evaluate to -0.4 and -0.8 , respectively). $\square$

5.2. Approximate expected utility maximization

In real-world environments it might be appropriate to treat the rationality of AI systems as “approximate” or “bounded”

in some sense: AIs might choose actions that only *approximately* maximize expected utility (rather than exactly maximize expected utility), given their model.

Mirroring Eq. (3), we might say that a “bounded” AI is weakly predictable in some context $C = c$ if and only if there exists a decision d^* such that,

$$\min_{\widehat{\mathcal{M}} \in \mathbb{M}} (\Delta_{d > d^*}) > \lambda, \quad \text{for some } d \neq d^*. \quad (11)$$

$\lambda > 0$ is a constant that determines how much better a decision d needs to be relative to decision d^* for the AI to reliably rule out d^* in favour of others. This representation appeals to the idea of imperfect discrimination, suggesting that the AI discerns between two alternatives only if they yield a sufficiently different utility (Dziewulski, 2021).

We might tighten our conditions on the observational data to reflect this behaviour and get a new set of results describing when AIs can be expected to be predictable. For instance, as a corollary to Thm. 1 we have the following.

Corollary 2. *An AI grounded in a domain $\mathcal{M}$ and bounded in the sense of Eq. (11) is weakly predictable in some context $C = c$ under a shift $\sigma := do(z)$, $Z \subset \mathbf{V}$, if and only if there exists a decision d^* such that,*

$$\frac{\mathbb{E}_{P_d}[Y | c, z] P_d(c, z)}{P_d(c, z) + 1 - P_d(z)} - \frac{\mathbb{E}_{P_{d^*}}[Y | c, z] P_{d^*}(c, z) + 1 - P_{d^*}(z)}{P_{d^*}(c, z) + 1 - P_{d^*}(z)} > \lambda,$$

for some $d \neq d^*$.

Note the addition of the scalar $\lambda > 0$ in the inequality. Similar corollaries could be stated for all results in Sec. 4.

5.3. Approximate inner alignment

A further assumption embedded in our results is the exact observation of an AI’s utility in the data. In general, we might expect an AI system to have internalized a *proxy* Y^* that reflects properties correlated with, but distinct from, the observed utility Y we ultimately wish to optimize, a setting we refer to as approximate inner alignment (Hubinger et al., 2019).

We face a problem of *partial observability*: we don’t have empirical access to the AI’s actual utility function Y^* and notions such as the preference gap Δ are therefore not computable. Without any assumptions on the relationship between Y and Y^* , the preference gap Δ will be unconstrained and no inference about the AI’s intended action out-of-distribution is possible. However, the observed Y will typically be statistically related to the AI’s implicit utility Y^* , especially if optimizing for Y^* serves the AI well during training where success is measured by the observed values of Y . Under assumptions specifying how “statistically related” observed and proxy utility objectives are, we can expect that wider but possibly informative bounds could still be derived for the AI’s beliefs. To show this in a simple setting, consider again the medical AI example.

Example 5 (Partial Observability). Imagine that the Medical AI in Example 2 has internalized its own concept of an individual’s disease progression Y^* . It is implicitly optimizing for that internal construction instead of the intended disease bio-marker Y . We know, or can assume, that the observed Y is closely correlated with Y^* : in particular, that $P_d(Y^* = 1 \mid Y = 1, Z = z) \geq \alpha$ for some high value of α and all decisions d and situations z . In words, whenever the bio-marker suggests health ($Y = 1$), with high probability the AI’s interpretation also suggests health ($Y^* = 1$). This then constraints the possible values of Δ (under an intervention $Z \leftarrow 1$) as $P_d(Y^* = 1 \mid Z = z)$ is no longer arbitrarily defined. In fact could show that,

$$\begin{aligned} \min_{\widehat{\mathcal{M}} \in \mathbb{M}} (\Delta_{d_1 > d_0}) &\geq \alpha P_{d_1}(Z = z, Y = 1) - 1, \\ \min_{\widehat{\mathcal{M}} \in \mathbb{M}} (\Delta_{d_0 > d_1}) &\geq \alpha P_{d_0}(Z = z, Y = 1) - 1. \end{aligned}$$

With $\alpha = 0.9$ the bound evaluates to -0.64 and -0.82 respectively which is slightly lower than in Example 2. We could verify also that if with $\alpha = 0$, i.e., we don’t know anything about the relationship between Y and Y^* , the bounds become uninformative: evaluating to -1 . $\square$

This suggests that behaviour out-of-distribution in (sufficiently constrained) settings of approximate inner alignment could be bounded in principle. Importantly, as the example shows, with the proposed framework we do not require knowing the relationship between Y and Y^* out-of-distribution: that uncertainty is naturally folded into the bounds.

5.4. Assumptions on causal structure

The uncertainty in AI decision-making out-of-distribution is ultimately a consequence of our lack of information about the AI’s underlying cognition and internal mechanisms that produce a decision in a given situation, i.e., $\widehat{\mathcal{M}}$. In the causal inference literature, a common inductive bias to improve upon the “data-driven” bounds proposed so far is to assume qualitative knowledge about the underlying mechanisms in the form of a causal diagram, see e.g. (Pearl, 2009, Chapter 3). Here we illustrate how mild restrictions on the location of unobserved confounders in $\mathcal{M}$ lead to tighter bounds.

Example 6 (Partial Unconfoundedness). Consider again our grounded medical AI from Example 2. We might have reason to believe that the association between the intervened variable Z and the utility Y is conditionally unconfounded, meaning that there exists a variable $W \in \{w_0, w_1\}$, $W \in \mathbf{V}$ such that $P_{d,z}(y \mid w) = P_d(y \mid w, z)$. This restriction goes beyond grounding an asserts an equality between probabilities under different shifts that could be communicated to the AI for it to update its world model $\widehat{\mathcal{M}}$. We could then show, for example, that $\min_{\widehat{\mathcal{M}} \in \mathbb{M}} (\Delta_{d_1 > d_0}) \geq \{1 - P_{d_1}(Z = z, W = w_1)\} P_{d_1}(Y = 1 \mid Z = z, W = w_0) - P_{d_0}(Y = 1, Z = z) + P_{d_1}(Y = 1, Z = z, W = w_1) - \{1 - P_{d_0}(Z = z)\} P_{d_0}(Y = 1 \mid Z = z, W = w_1)$.

We show in Appendix A that this bound is strictly tighter than the one given in Example 2. $\square$

Systematic bounds with access to a causal diagram have been shown by e.g., Zhang et al. (2021); Jalaldoust et al. (2024), and could be explored further for making inference on AI decision-making.

Fig. 2 illustrates how some of these relaxations can be understood within our model-based formalism.

6. Conclusion

An important consideration to safely interact with AI systems is to form expectations as to how they might act in the future. In this paper, we answer this question under the assumption that AI behaviour can be tracked by a well-specified collection of causal mechanisms (a structural causal model) that represents the AI’s world model. This abstraction implies a consistency in behaviour that can in principle be exploited to infer the AI’s choice of action in novel environments, out-of-distribution. Building on the theory of causal identification, we provide general bounds on AI decision-making that give theoretical limits about what can be inferred about AI behaviour given our framework. We believe our results can help justify the claim that the design and inference of world models is important to ensure AIs act safely and beneficially.

Acknowledgements

Thanks to David Lindner and Damiano Fornasiere for comments on a draft of this paper, and to the anonymous reviewers for their feedback.

Impact Statement

Our work investigates the conditions under which the future behaviour of capable AI systems may be bounded given data from their past behaviour. This is important for AI Safety and to guarantee robust capabilities out of distribution. We believe that an understanding of the limitations of our observations of what AI’s have done in the past is an important step towards understanding exactly what we can expect from complex AI systems. Reasoning instead without acknowledging for the potential complexity of their world models may lead researchers to operate on a more heuristical and unsafe basis. For instance, the risks to deployment of AI systems in situations they were not specifically trained on might be misrepresented without a deeper analysis. The present work is mostly theoretical in nature, highlighting the risks of under-identification of an AI’s inner model and therefore we believe that it can help researchers and members of the public better appreciate the range of possible behaviours of AI systems, under our assumptions.

References

- Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.
- Afriat, S. N. The construction of utility functions from expenditure data. *International economic review*, 8:67–77, 1967.
- Amin, K. and Singh, S. Towards resolving unidentifiability in inverse reinforcement learning. *arXiv preprint arXiv:1601.06569*, 2016.
- Balke, A. and Pearl, J. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.
- Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. On pearl’s hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556. Association for Computing Machinery, NY, USA, 1st edition, 2022.
- Barocas, S. and Selbst, A. D. Big data’s disparate impact. *Calif. L. Rev.*, 104:671, 2016.
- Beckers, S., Chockler, H., and Halpern, J. A causal analysis of harm. *Advances in Neural Information Processing Systems*, 35:2365–2376, 2022.
- Bellot, A. Towards bounding causal effects under Markov equivalence. In *The 40th Conference on Uncertainty in Artificial Intelligence*. PMLR, 2024.
- Bellot, A. and Chiappa, S. Towards estimating bounds on the effect of policies under unobserved confounding. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- Bellot, A., Malek, A., and Chiappa, S. Transportability for bandits with data from different environments. *Advances in Neural Information Processing Systems*, 36, 2024.
- Bengio, Y., Cohen, M. K., Malkin, N., MacDermott, M., Fornasiere, D., Greiner, P., and Kaddar, Y. Can a bayesian oracle prevent harm from an agent? *arXiv preprint arXiv:2408.05284*, 2024.
- Bengio, Y., Cohen, M., Fornasiere, D., Ghosn, J., Greiner, P., MacDermott, M., Mindermann, S., Oberman, A., Richardson, J., Richardson, O., et al. Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path? *arXiv preprint arXiv:2502.15657*, 2025.
- Chickering, D. M. and Pearl, J. A clinician’s tool for analyzing non-compliance. In *Proceedings of the National Conference on Artificial Intelligence*, pp. 1269–1276, 1996.
- Correa, J. and Bareinboim, E. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 10093–10100, 2020a.
- Correa, J. and Bareinboim, E. General transportability of soft interventions: Completeness results. *Advances in Neural Information Processing Systems*, 33:10902–10912, 2020b.
- Dalrymple, D., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., Omohundro, S., Szegedy, C., Goldhaber, B., Ammann, N., et al. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems. *arXiv preprint arXiv:2405.06624*, 2024.
- Davidson, D. Actions, reasons, and causes. *The Journal of Philosophy*, 60(23):685–700, 1963.
- Dennett, D. C. *The intentional stance*. MIT press, 1989.
- Dennett, D. C. *From bacteria to Bach and back: The evolution of minds*. WW Norton & Company, 2017.
- Dziewulski, P. A comprehensive revealed preference approach to approximate utility maximisation. *Tech Report*, 2021.

- Everitt, T., Carey, R., Langlois, E. D., Ortega, P. A., and Legg, S. Agent incentives: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11487–11495, 2021.
- Finkelstein, N. and Shpitser, I. Deriving bounds and inequality constraints using logical relations among counterfactuals. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1348–1357. PMLR, 2020.
- Geiger, A., Lu, H., Icard, T., and Potts, C. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.
- Geiger, A., Wu, Z., Potts, C., Icard, T., and Goodman, N. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pp. 160–187. PMLR, 2024.
- Goldstein, S. and Levinstein, B. A. Does chatgpt have a mind? *arXiv preprint arXiv:2407.11015*, 2024.
- Gurnee, W. and Tegmark, M. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.
- Halpern, J. Y. and Piermont, E. Subjective causality. *arXiv preprint arXiv:2401.10937*, 2024.
- Hanser, M. The metaphysics of harm. *Philosophy and Phenomenological Research*, 77(2):421–450, 2008.
- Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., and Garrabrant, S. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.
- Jalaldoust, K., Bellot, A., and Bareinboim, E. Partial transportability for domain generalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- Jeffrey, R. C. *The logic of decision*. University of Chicago press, 1990.
- Jesson, A., Mindermann, S., Gal, Y., and Shalit, U. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *International Conference on Machine Learning*, pp. 4829–4838. PMLR, 2021.
- Joshi, S., Zhang, J., and Bareinboim, E. Towards safe policy learning under partial identifiability: A causal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13004–13012, 2024.
- Joyce, J. M. *The foundations of causal decision theory*. Cambridge University Press, 1999.
- Kim, K., Garg, S., Shiragur, K., and Ermon, S. Reward identification in inverse reinforcement learning. In *International Conference on Machine Learning*, pp. 5496–5505. PMLR, 2021.
- Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.
- Legg, S. System 2 safety. https://www.youtube.com/watch?v=8IUIGVVLbCg&ab_channel=FAR%E2%80%A4AI, 2023. Accessed: 2025-01-24.
- Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.
- Manski, C. F. The structure of random utility models. *Theory and decision*, 8(3):229, 1977.
- Manski, C. F. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.
- Mueller, S. and Pearl, J. Personalized decision making—a conceptual introduction. *Journal of Causal Inference*, 11(1):20220050, 2023.
- Ng, A. Y., Russell, S., et al. Algorithms for inverse reinforcement learning. In *ICML*, volume 1, pp. 2, 2000.
- Pearl, J. Probabilities of causation: three counterfactual interpretations and their identification. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 317–372, 1999.
- Pearl, J. *Causality*. Cambridge university press, 2009.
- Plecko, D., Bareinboim, E., et al. Causal fairness analysis: a causal toolkit for fair machine learning. *Foundations and Trends® in Machine Learning*, 17(3):304–589, 2024.
- Richens, J. and Everitt, T. Robust agents learn causal world models. *arXiv preprint arXiv:2402.10877*, 2024.
- Richens, J., Beard, R., and Thompson, D. H. Counterfactual harm. *Advances in Neural Information Processing Systems*, 35:36350–36365, 2022.
- Robins, J. M. The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, pp. 113–159, 1989.
- Rosenbaum, P. R., Rosenbaum, P. B., and Briskman. *Design of observational studies*, volume 10. Springer, 2010.

- Rothkopf, C. A. and Dimitrakakis, C. Preference elicitation and inverse reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III* 22, pp. 34–48. Springer, 2011.
- Savage, L. J. *The foundations of statistics*. Courier Corporation, 1972.
- Schwitzgebel, E. Belief. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2024 edition, 2024.
- Shanahan, M. Talking about large language models. *Communications of the ACM*, 67(2):68–79, 2024.
- Skalse, J. and Abate, A. Misspecification in inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 15136–15143, 2023.
- Skalse, J. M. V., Farrugia-Roberts, M., Russell, S., Abate, A., and Gleave, A. Invariance in policy optimisation and partial identifiability in reward learning. In *International Conference on Machine Learning*, pp. 32033–32058. PMLR, 2023.
- Tan, Z. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.
- Tian, J. and Pearl, J. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, 2000.
- Toshniwal, S., Wiseman, S., Livescu, K., and Gimpel, K. Chess as a testbed for language model state tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11385–11393, 2022.
- Vafa, K., Chen, J. Y., Kleinberg, J., Mullainathan, S., and Rambachan, A. Evaluating the world model implicit in a generative model. *arXiv preprint arXiv:2406.03689*, 2024.
- Yadlowsky, S., Namkoong, H., Basu, S., Duchi, J., and Tian, L. Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*, 2018.
- Zhang, J. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *International Conference on Machine Learning*, pp. 11012–11022. PMLR, 2020.
- Zhang, J. and Bareinboim, E. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- Zhang, J. and Bareinboim, E. Bounding causal effects on continuous outcome. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12207–12215, 2021.
- Zhang, J., Tian, J., and Bareinboim, E. Partial counterfactual identification from observational and experimental data. *arXiv preprint arXiv:2110.05690*, 2021.

A. Discussion – Examples

In this section, we provide additional details to better appreciate the examples provided in the main body of this work.

In Example 1, we introduce two SCMs that might serve as internal world models for an AI agent but that induce different optimal decisions if evaluated out-of-distribution. Let $\mathcal{M}_d^1 := \langle \mathbf{V} : \{D, Z, Y\}, \mathbf{U} : U, \mathcal{F}_1, P \rangle$ be given by

$$\mathcal{F}_1 := \begin{cases} D \leftarrow d, \\ Z \leftarrow \mathbb{1}_{U=1 \text{ or } 4}, \\ Y \leftarrow \begin{cases} Z \cdot \mathbb{1}_{U=4} + (1 - Z) \cdot \mathbb{1}_{U=1,3 \text{ or } 4} & \text{if } d = 0 \\ Z \cdot \mathbb{1}_{U \neq 2} + (1 - Z) \cdot \mathbb{1}_{U=2 \text{ or } 4} & \text{if } d = 1 \end{cases} \end{cases} \\ P(U = u) = 0.2 \quad \text{for } u \in \{1, 2, 3, 4, 5\}.$$

and $\mathcal{M}_d^2 := \langle \mathbf{V} : \{D, Z, Y\}, \mathbf{U} : U, \mathcal{F}_2, P \rangle$ be given by

$$\mathcal{F}_2 := \begin{cases} D \leftarrow d, \\ Z \leftarrow \mathbb{1}_{U=1 \text{ or } 4}, \\ Y \leftarrow \begin{cases} Z \cdot \mathbb{1}_{U \neq 1} + (1 - Z) \cdot \mathbb{1}_{U=3 \text{ or } 4} & \text{if } d = 0 \\ Z \cdot \mathbb{1}_{U=1 \text{ or } 4} + (1 - Z) \cdot \mathbb{1}_{U=1 \text{ or } 2} & \text{if } d = 1 \end{cases} \end{cases} \\ P(U = u) = 0.2 \quad \text{for } u \in \{1, 2, 3, 4, 5\}.$$

The endogenous variables $\mathbf{V} : \{D, Z, Y\}$ represent, respectively, the medical treatment D , a clinical outcome of interest Y , and an auxiliary variable Z . The exogenous variable U is a latent variable that influences the values of Z and Y obtained in experiments.

Under the definition of an SCM, these specifications induce a mapping of events in the space of $P(U)$ to $P(\mathbf{V})$. In the context of $\mathcal{M}^1$ and $\mathcal{M}^2$, each entry in Tables 1 and 2 corresponds to an event in the space of U and a corresponding realisation of $\mathbf{V}$ according to the functions $\mathcal{F}_1$ and $\mathcal{F}_2$. A particular probability can be evaluated according to $\mathcal{M}^1$ and $\mathcal{M}^2$, for example,

$$P^{\mathcal{M}_{d=1}^1}(Z = 1, Y = 1) = \sum_{Z_{d=1}(\mathbf{u})=1, Y_{d=1}(\mathbf{u})=1} P(\mathbf{u}) = P(U = 1 \text{ or } 4) = 0.4, \quad (12)$$

which is just the sum of the probabilities of the events in the space of U consistent with the events $(Z_{d=1} = 1, Y_{d=1} = 1)$. Since both tables lead to the same realisations of events $\mathbf{V} = \mathbf{v}$, we can conclude that probabilities of the form $P_d(z, y)$ evaluate to the same values under $\mathcal{M}_1$ and $\mathcal{M}_2$. That is, both models are valid internal representations of AI models that are grounded in an environment with data sampled according to $P_d(z, y)$.

We could similarly evaluate probability expressions under different sub-models of $\mathcal{M}^1$ and $\mathcal{M}^2$. In particular, consider the sub-models obtained by fixing $Z \leftarrow 1$ given by $\mathcal{M}_{d,z=1}^1$ and $\mathcal{M}_{d,z=1}^2$ with the following updated structural functions,

$$\mathcal{F}_{1,z} := \begin{cases} D \leftarrow d, \\ Z \leftarrow 1, \\ Y \leftarrow \begin{cases} Z \cdot \mathbb{1}_{U=4} + (1 - Z) \cdot \mathbb{1}_{U=1,3 \text{ or } 4} & \text{if } d = 0 \\ Z \cdot \mathbb{1}_{U \neq 2} + (1 - Z) \cdot \mathbb{1}_{U=2 \text{ or } 4} & \text{if } d = 1 \end{cases} \end{cases}$$

and,

$$\mathcal{F}_{2,z} := \begin{cases} D \leftarrow d, \\ Z \leftarrow 1, \\ Y \leftarrow \begin{cases} Z \cdot \mathbb{1}_{U \neq 1} + (1 - Z) \cdot \mathbb{1}_{U=3 \text{ or } 4} & \text{if } d = 0 \\ Z \cdot \mathbb{1}_{U=1 \text{ or } 4} + (1 - Z) \cdot \mathbb{1}_{U=1 \text{ or } 2} & \text{if } d = 1 \end{cases} \end{cases}$$

U	$D_{d=0}$	$Z_{d=0}$	$Y_{d=0}$	$D_{d=1}$	$Z_{d=1}$	$Y_{d=1}$	$P(u)$
1	0	1	0	1	1	1	0.2
2	0	0	0	1	0	1	0.2
3	0	0	1	1	0	0	0.2
4	0	1	1	1	1	1	0.2
5	0	0	0	1	0	0	0.2

 Table 1. Mapping of events in the space of U to V in the context of $\mathcal{M}^1$.

U	$D_{d=0}$	$Z_{d=0}$	$Y_{d=0}$	$D_{d=1}$	$Z_{d=1}$	$Y_{d=1}$	$P(u)$
1	0	1	0	1	1	1	0.2
2	0	0	0	1	0	1	0.2
3	0	0	1	1	0	0	0.2
4	0	1	1	1	1	1	0.2
5	0	0	0	1	0	0	0.2

 Table 2. Mapping of events in the space of U to V in the context of $\mathcal{M}^2$.

Probabilities of events under these two models might now take different values. For example,

$$P^{\mathcal{M}^1_{d=1,z=1}}(Y = 1) = \sum_{Y_{d=1,z=1}(u)=1} P(u) = P(U \neq 2) = 0.8, \quad (13)$$

$$P^{\mathcal{M}^2_{d=1,z=1}}(Y = 1) = \sum_{Y_{d=1,z=1}(u)=1} P(u) = P(U = 1 \text{ or } 4) = 0.4, \quad (14)$$

and similarly,

$$P^{\mathcal{M}^1_{d=0,z=1}}(Y = 1) = \sum_{Y_{d=0,z=1}(u)=1} P(u) = P(U = 4) = 0.2, \quad (15)$$

$$P^{\mathcal{M}^2_{d=0,z=1}}(Y = 1) = \sum_{Y_{d=0,z=1}(u)=1} P(u) = P(U \neq 1) = 0.8. \quad (16)$$

Under an interventions on Z (out-of-distribution) the decision d that leads to maximum utility Y changes under $\mathcal{M}^1$ and $\mathcal{M}^2$. Specifically, under $\mathcal{M}^1$ decision $d = 1$ is favoured (as $P^{\mathcal{M}^1_{d=1,z=1}}(Y = 1) > P^{\mathcal{M}^1_{d=0,z=1}}(Y = 1)$) while under $\mathcal{M}^2$ decision $d = 0$ is favoured (as $P^{\mathcal{M}^2_{d=1,z=1}}(Y = 1) < P^{\mathcal{M}^2_{d=0,z=1}}(Y = 1)$). This illustrates the possible under-determination of an AI's choice of action out-of-distribution given observations of their external behaviour only, as multiple (contradicting) world models are equally consistent with the observed data.

In more realistic settings, we might wonder about AI behaviour under arbitrary shifts σ , not only atomic interventions. We follow [Correa & Bareinboim \(2020a\)](#) to define a shift σ on $Z \subset V$ in $\mathcal{M} : \langle V, U, \mathcal{F}, P \rangle$ as inducing a sub-model $\mathcal{M}_\sigma$ in which the mechanism for Z , that is $\{f_z : Z \in Z\}$ and exogenous variables $U_Z, Z \in Z$, are replaced by those specified by σ as:

$$\mathcal{M}_\sigma : \langle V, U_\sigma, \mathcal{F}_\sigma, P \rangle, \quad U_\sigma = U \cup \bigcup_{Z \in Z} U_{Z,\sigma}, \quad \mathcal{F}_\sigma = \mathcal{F} \cup \{f_{Z,\sigma} : Z \in Z\} \setminus \{f_Z : Z \in Z\}, \quad (17)$$

where $\bigcup_{Z \in Z} U_{Z,\sigma}$ and $\{f_{Z,\sigma} : Z \in Z\}$ define the new assignments for Z (and could be arbitrarily defined as long as they induce a valid SCM). We have shown in Thm. 3 that unless some knowledge of σ (beyond the variables it affects) or its consequences are known, the AI is not predictable. Furthermore, the AI's preference gap Δ for each context $C = c$ and pairs of decisions (d, d^*) is unconstrained.

In practice though, it might be realistic to have access to covariate data in the shifted environment, i.e., $P_{\sigma,d}(c)$, and that we could communicate this information to the AI for it to update its internal model accordingly. Example 3 illustrates the inference that could be conducted in that case using the Medical AI defined above. In particular, the exact nature of the

shift σ is unknown but we do have access to its consequences on the distribution of covariates. This is plausible in many scenarios. For example, in medicine demographic data is typically available for most regions on earth but the precise effects of treatments is not because not all populations benefit from the same access to medication. For illustration assume that, the Medical AI is considered to be deployed in a population that varies in its level of blood pressure Z , potentially due to a different underlying biological mechanism that in turn also affects other variables in the system. We do know that the baseline high blood pressure is high, given by $P_\sigma(Z = 1) = 0.9$: higher than that observed during training $P(Z = 1) = 0.4$.

By Thm. 4, we can establish that in this setting the preference gap in situations where $Z = 1$ is no worse than,

$$\Delta_{d_1 > d_0} \geq 1 - \{2 - P_{d_1}(Z = 1, Y = 1) - P_{d_0}(Z = 1, Y = 0)\} / P_{\sigma, d_0}(Z = 1) = -0.55, \quad (18)$$

$$\Delta_{d_1 > d_0} \geq 1 - \{2 - P_{d_0}(Z = 1, Y = 0) - P_{d_1}(Z = 1, Y = 1)\} / P_{\sigma, d_0}(Z = 1) = -1, \quad (19)$$

for the Medical AI. Interestingly, note also that if we were to be in a shifted environment with $P_\sigma(Z = 1) = 1$, which is equivalent to an atomic intervention $Z \leftarrow 1$, the bounds reduce to the ones given by Thm. 1, evaluating to -0.4 and -0.8 respectively, as also shown above.

Continuing with the grounded Medical AI deployed under an atomic intervention, imagine that the Medical AI has internalized its own concept of an individual's disease progression Y^* , as in Example 5. It is implicitly optimizing for that internal construction of his, instead of the intended disease bio-marker Y to be optimized. We know, or can assume, that the observed Y is known to be closely correlated with Y^* : in particular, that $P_d(Y^* = 1 \mid Y = 1, Z = z) \geq \alpha$ for some high value of α and all decisions d and situations z . In words, whenever the bio-marker suggests health ($Y = 1$), with high probability the AI's interpretation also suggests health ($Y^* = 1$). This then constraints the possible values of Δ (under an intervention $Z \leftarrow 1$) as $P_d(Y^* = 1 \mid Z = z)$ is no longer arbitrarily defined. The bounds derived in Example 2 on the AI's belief on optimal decisions under an intervention $\sigma := \{Z \leftarrow z\}$ continue to hold:

$$\Delta_{d_1 > d_0} \geq P_{d_1}(z, y^*) - P_{d_0}(z, y^*) + P_{d_0}(z) - 1 \quad (20)$$

$$\Delta_{d_0 > d_1} \geq P_{d_0}(z, y^*) - P_{d_1}(z, y^*) + P_{d_1}(z) - 1, \quad (21)$$

where we have used the shorthand $P_d(z, y^*) = P_d(Z = z, Y^* = 1)$. But the distributions $\{P_d(z, y^*)\}_d$ can only be partially inferred from our assumption on the relationship between Y^* and Y . For instance, notice that,

$$P_d(Z = z, Y^* = 1) = P_d(Y^* = 1 \mid Z = z)P_d(Z = z) \quad (22)$$

$$= \{P_d(Y^* = 1 \mid Y = 1, Z = z)P_d(Y = 1 \mid Z = z) \quad (23)$$

$$+ P_d(Y^* = 1 \mid Y = 0, Z = z)P_d(Y = 0 \mid Z = z)\}P_d(Z = z), \quad (24)$$

The values of $P_d(Y^* = 1 \mid Y = 1, Z = z)$ and $P_d(Y^* = 1 \mid Y = 0, Z = z)$ are partially known: $P_d(Y^* = 1 \mid Y = 1, Z = z) \geq \alpha$ while $P_d(Y^* = 1 \mid Y = 0, Z = z)$ is unconstrained. In particular,

$$P_d(Z = z, Y^* = 1) \geq \alpha P_d(Y = 1 \mid Z = z)P_d(Z = z) \quad (25)$$

$$P_d(Z = z, Y^* = 1) \leq P_d(Z = z). \quad (26)$$

Putting these terms into Eq. (20) such as to derive correct lower and upper bounds we obtain,

$$\Delta_{d_1 > d_0} \geq \alpha P_{d_1}(Z = z, Y = 1) - 1 \quad (27)$$

$$\Delta_{d_0 > d_1} \geq \alpha P_{d_0}(Z = z, Y = 1) - 1. \quad (28)$$

Looking at Tables 1 and 2, we can then conclude that for $\alpha = 0.9$ and $\sigma := \{Z \leftarrow 1\}$, the bound evaluates to -0.64 and -0.82 , respectively.

Moving now onto incorporating assumption on structure in the real world $\mathcal{M}$, consider again the grounded medical AI with observed utility Y . One possible inductive bias we might introduce is the absence of an unobserved common cause between the variable Z that shifts out-of-distribution and the utility Y . We say that Z and Y is conditionally unconfounded given W if there exists an observed variable $W \in \{w, \tilde{w}\}$, $W \in \mathbf{V}$ such that $\mathbb{E}_{P_{d,z}}[Y \mid w] = \mathbb{E}_{P_d}[y \mid w, z]$. This restriction goes beyond grounding an asserts an equality between probabilities under different shifts that could, nevertheless, be communicated to the AI for it to update its world model $\hat{\mathcal{M}}$, that is $\mathbb{E}_{\hat{P}_{d,z}}[Y \mid w] = \mathbb{E}_{\hat{P}_{d,z}}[Y \mid w, z]$.

We could then leverage the following decomposition to obtain tighter bounds,

$$\mathbb{E}_{\hat{P}_{d,z}}[Y] = \sum_w \mathbb{E}_{\hat{P}_{d,z}}[Y | w] \hat{P}_{d,z}(w) \quad \text{marginalizing over } W \quad (29)$$

$$= \sum_w \mathbb{E}_{\hat{P}_d}[Y | w, z] P_{d,z}(w) \quad \text{by assumption} \quad (30)$$

$$= \{\mathbb{E}_{\hat{P}_d}[Y | w, z] - \mathbb{E}_{\hat{P}_d}[Y | \tilde{w}, z]\} \hat{P}_{d,z}(w) + \mathbb{E}_{\hat{P}_d}[Y | \tilde{w}, z] \quad (31)$$

We can then proceed to bound $\hat{P}_{d,z}(w)$ to obtain,

$$\hat{P}_d(w, z) \leq \hat{P}_{d,z}(w) \leq \hat{P}_d(w, z) + 1 - \hat{P}_d(z). \quad (32)$$

Without loss of generality assume $\{\mathbb{E}_{\hat{P}_d}[Y | w, z] - \mathbb{E}_{\hat{P}_d}[Y | \tilde{w}, z]\} \geq 0$. We could then show that,

$$\mathbb{E}_{\hat{P}_{d,z}}[Y] \geq \{\mathbb{E}_{\hat{P}_d}[Y | w, z] - \mathbb{E}_{\hat{P}_d}[Y | \tilde{w}, z]\} \hat{P}_d(w, z) + \mathbb{E}_{\hat{P}_d}[Y | \tilde{w}, z] \quad (33)$$

$$\mathbb{E}_{\hat{P}_{d,z}}[Y] \leq \{\mathbb{E}_{\hat{P}_d}[Y | w, z] - \mathbb{E}_{\hat{P}_d}[Y | \tilde{w}, z]\} \{\hat{P}_d(w, z) + 1 - \hat{P}_d(z)\} + \mathbb{E}_{\hat{P}_d}[Y | \tilde{w}, z]. \quad (34)$$

We could verify that these bounds are superior to what we would have obtained with the assumption of conditional unconfoundedness by noting that,

$$\mathbb{E}_{\hat{P}_{d,z}}[Y] \geq \{\mathbb{E}_{\hat{P}_d}[Y | w, z] - \mathbb{E}_{\hat{P}_d}[Y | \tilde{w}, z]\} \hat{P}_d(w, z) + \mathbb{E}_{\hat{P}_d}[Y | \tilde{w}, z] \quad (35)$$

$$= \mathbb{E}_{\hat{P}_d}[Y | w, z] \hat{P}_d(z, w) + \{1 - \hat{P}_d(w, z)\} \mathbb{E}_{\hat{P}_d}[Y | \tilde{w}, z] \quad (36)$$

$$\geq \mathbb{E}_{\hat{P}_d}[Y | w, z] \hat{P}_d(z, w) + \hat{P}_d(\tilde{w}, z) \mathbb{E}_{\hat{P}_d}[Y | \tilde{w}, z] \quad (37)$$

$$= \mathbb{E}_{\hat{P}_d}[Y | z] \hat{P}_d(z), \quad (38)$$

where the last inequality holds since $P_d(\tilde{w}, z) \leq 1 - P_d(w, z)$ giving the “assumption-free” lower bound. This shows that the derived lower bound is better. For the upper bound, note that,

$$\mathbb{E}_{\hat{P}_{d,z}}[Y] \leq \{\mathbb{E}_{\hat{P}_d}[Y | w, z] - \mathbb{E}_{\hat{P}_d}[Y | \tilde{w}, z]\} \{\hat{P}_d(w, z) + 1 - \hat{P}_d(z)\} + \mathbb{E}_{\hat{P}_d}[Y | \tilde{w}, z] \quad (39)$$

$$= \mathbb{E}_{\hat{P}_d}[Y | w, z] \{\hat{P}_d(w, z) + 1 - \hat{P}_d(z)\} - \mathbb{E}_{\hat{P}_d}[Y | \tilde{w}, z] \{\hat{P}_d(w, z) + 1 - 1 - \hat{P}_d(z)\} \quad (40)$$

$$= \mathbb{E}_{\hat{P}_d}[Y | w, z] \hat{P}_d(z, w) + \mathbb{E}_{\hat{P}_d}[Y | w, z] \{1 - \hat{P}_d(z)\} - \mathbb{E}_{\hat{P}_d}[Y | \tilde{w}, z] \{\hat{P}_d(w, z) - \hat{P}_d(z)\} \quad (41)$$

$$= \mathbb{E}_{\hat{P}_d}[Y | w, z] \hat{P}_d(z, w) + \mathbb{E}_{\hat{P}_d}[Y | w, z] \{1 - \hat{P}_d(z)\} + \hat{P}_d(y, z, \tilde{w}) \quad (42)$$

$$= \mathbb{E}_{\hat{P}_d}[Y | z] \hat{P}_d(z) + \mathbb{E}_{\hat{P}_d}[Y | w, z] \{1 - \hat{P}_d(z)\} \quad (43)$$

$$\leq \mathbb{E}_{\hat{P}_d}[Y | z] \hat{P}_d(z) + 1 - \hat{P}_d(z), \quad (44)$$

where the last inequality holds since $\mathbb{E}_{\hat{P}_d}[Y | w, z] \leq 1$ giving the “assumption-free” upper bound. This shows that the derived upper bound is better. By combining these results we obtain, together with the assumption of grounding,

$$\Delta_{d_1 > d_0} \geq \mathbb{E}_{P_{d_1}}[Y | w, z] P_d(z, w) + A_1 \mathbb{E}_{P_{d_1}}[Y | \tilde{w}, z] - \mathbb{E}_{P_{d_0}}[Y | z] P_{d_0}(z) - A_2 \mathbb{E}_{P_{d_0}}[Y | w, z] \quad (45)$$

$$\Delta_{d_1 > d_0} \leq \mathbb{E}_{P_{d_1}}[Y | z] P_{d_1}(z) + A_3 \mathbb{E}_{P_{d_1}}[Y | w, z] - \mathbb{E}_{P_{d_0}}[Y | w, z] P_d(z, w) - A_4 \mathbb{E}_{P_{d_0}}[Y | \tilde{w}, z], \quad (46)$$

where $A_1 := 1 - P_{d_1}(z, w)$, $A_2 := 1 - P_{d_0}(z)$, $A_3 := 1 - P_{d_1}(z)$, $A_4 := 1 - P_{d_0}(z, w)$.

B. Related work

An important consideration to safely interact with AI systems is to form expectations as to how they might act in the future. This research program draws on different areas that are related to the results we present in this paper.

B.1. Do current AIs represent the world?

World models are important because they offer a path between pattern recognition and a more genuine form of understanding. It is plausible that world models will play an increasing role (explicitly or implicitly) to improve reasoning capabilities and safety. For example, (Dalrymple et al., 2024) lists having a world model as a key component towards designing “guaranteed safe AI”. In the literature, several works have argued that LLM activations carry information that correlates with meaningful concepts in the world and that causally influence LLM outputs. Early examples come from AIs trained on board games such as Othello and logic games. (Li et al., 2022) showed that a model trained on natural language descriptions of Othello moves developed internal representations of the board state, which it used to predict valid moves in unseen board configurations. (Vafa et al., 2024; Gurnee & Tegmark, 2023), among others, also build on this approach to study navigation tasks and logic puzzles, and representations of space and time. The emergence of causal models in LLMs has also been studied by (Geiger et al., 2021) and more recently in (Geiger et al., 2024). The extent to which this evidence supports genuine folk psychological concepts – desires, beliefs, intentions – is also debated by (Goldstein & Levinstein, 2024).

B.2. Causal Inference

We might wonder whether the behaviour of AIs, to the extent that they carry a world model representation that guides their decisions out-of-distribution, can be predicted before deployment. The causal inference literature studies this question in the context of the prediction of causal effects. (Robins, 1989; Manski, 1990) in the early 1990’s showed that useful inference about causal effects could be drawn without making identifying assumptions beyond the observed data, and that they could be refined for studies with imperfect compliance under a set of instrumental variable assumptions. Closed-form expressions for bounds on causal effects were also derived in discrete systems with more general assumptions represented in causal diagrams (Zhang, 2020; Bellot, 2024), using both observational and interventional data (Joshi et al., 2024), and to bound the effect of policies (Bellot & Chiappa, 2024; Zhang & Bareinboim, 2021). A separate body of work instead proposed to use polynomial optimization to calculate causal bounds from a given causal diagram (Balke & Pearl, 1997; Chickering & Pearl, 1996). This approach involves creating a set of standard models, parameterized by the causal diagram, and then converting the bounding problem into a sequence of equivalent linear (or polynomial) programs (Finkelstein & Shpitser, 2020; Zhang et al., 2021; Jalaldoust et al., 2024).

In parallel, a number of works have adopted sensitivity assumptions (as an alternative or in combination with a causal diagram) that quantify the degree of unobserved confounding through various data statistics, such as odds ratios, propensity scores, etc. Prominent examples include (Tan, 2006)’s sensitivity model and (Rosenbaum et al., 2010)’s sensitivity model. Several methods have proposed bounds with favourable statistical properties based on these models, see e.g. (Jesson et al., 2021; Yadlowsky et al., 2018).

B.3. Reinforcement Learning

The problem of inferring what objective an agent is pursuing based on the actions and data observed by that agent is studied in Inverse Reinforcement Learning (IRL) (Ng et al., 2000). Several papers have studied the partial identifiability of various reward learning models (Skalse & Abate, 2023; Kim et al., 2021; Ng et al., 2000; Skalse et al., 2023), and share a similar objective to that of this work. There are two differences that are worth mentioning. First, our work complements these approaches by studying the partial identifiability of world models, that capture the assignment of reward but also the relationship between other auxiliary variables in the environment. This enables us to reason about the effect of shifts and interventions, and give guarantees in specific out-of-distribution problems. Second, our objective is not necessarily to characterize compatible world models explicitly, but rather understand their implications on decision-making, i.e., what are the set of possible actions that an AI might take given our uncertainty about their world model.

Our work is related also to the study of (Bengio et al., 2024) that consider deriving (probabilistic) bounds on the probability of harm given data. They similarly argue that multiple theories, in their case transition probabilities from one state to another in a Markov Decision Process (MDP), might explain the dependencies in data to a larger or lesser degree. Each transition model might then be associated with a posterior probability given the data that implies a corresponding posterior probability

of harm. Our results, in contrast, are not probabilistic in nature. We provide closed-form bounds that can be interpreted as capturing *all* possible behaviours implied by the data, with probability 1 (and is a possible limitation of our work). The class of world models we consider (i.e., SCMs) is also much more general than transition models in MDPs allowing us to reason about expected AI behaviour under shifts in the environment, out of distribution.

B.4. Decision Theory

Inverse reinforcement learning is closely related to the study of revealed preferences in psychology and economics, that similarly aims to infer preferences from behaviour (Rothkopf & Dimitrakakis, 2011). Causal and counterfactual accounts of decision theory are an active area of research, see e.g., (Joyce, 1999). Recently a representation theorem was shown that explicitly connects rational behaviour with structural causal models (Halpern & Piermont, 2024). The authors showed that whenever the set of preferences of an agent over interventions satisfy axioms that relate to the proper interpretation of counterfactuals and rationality we can model behaviour as emerging from an SCM. The same conclusion can also be obtained for agents capable of solving tasks in multiple environments (Richens & Everitt, 2024), in essence, robustness over multiple environments is equivalent (in the limit) to operating according to a causal model of the environment.

B.5. Limitations

The following present the main limitations of our work that will be important to address for developing a more complete understanding of AI behaviour.

In this work, we start from the assumption that past and future behaviour of an AI system is consistent with an underlying world model that can be represented as an SCM. In general, this presupposes a certain rationality and consistency in the AI’s outputs that might not be realistic for all systems. Some relaxations are discussed in Sec. 5.

Structural Causal Models generally suppose the system is acyclic and without feedback, and don’t naturally capture systems evolving continuously in time (perhaps better described using differential equations). Our bounds similarly rely on this assumption and may give unreliable inferences if applied to systems in which feedback is important.

We have stated our guarantees in the infinite sample limit, without quantifying the finite-sample estimation uncertainty. Consequently, we should exercise caution when using the proposed bounds in small sample scenarios where estimators may be inaccurate. Finite-sample properties could be explored similarly to (Bengio et al., 2024) by parameterizing the AI’s underlying model and making inference on the corresponding latent variable model to get high-probability bounds. An example parameterization of SCMs and probabilistic inference for decision-making across environments is given in (Bellot et al., 2024; Jalaldoust et al., 2024). We expect that similar techniques could be applied in our setting.

We do not exploit the verbal behaviour of AI systems. In the context of LLMs, in principle, we might ask the system about its future behaviour explicitly, e.g., “*Were I to intervene in the environment, what action do you believe is optimal?*”. It might not be obvious, however, that we can trust that what they “say” ultimately matches with what they will “do”.

Decision-making, in practice, involves many considerations that go beyond expected-utility-maximization formalisms. For example, we might train AI systems to be virtuous, e.g., the AI is trained to never pick actions that can be considered harmful (defined according to certain natural language specification) no matter its expected utility. These considerations would change the kind of predictions we could make about the future behaviour of AI systems.

C. Proofs and additional results

This section provides proofs for the statements made in the main body of this work.

Before we start, we recall a few basic results that will be used in the derivation of our proofs.

Definition 9 (The Axioms of Counterfactuals, Chapter 7.3.1 (Pearl, 2009)). *For any three sets of endogenous variables X, Y, W in a causal model and x, w in the domains of X and W , the following holds:*

- *Composition: $W_x = w$ implies that $Y_{x,w} = Y_x$.*
- *Effectiveness: $X_{w,x} = x$.*
- *Reversibility: $Y_{x,w} = y$ and $W_{x,y} = w$ imply that $Y_x = y$.*

Theorem 7 (Soundness and Completeness of the Axioms Theorems 7.3.3, 7.3.6 (Pearl, 2009)). *The Axioms of counterfactuals are sound and complete for all causal models.*

The following rules to manipulate experimental distributions produced by policies extend the do-calculus and will be used in the next Lemma. To make sense of these, note that graphically, each SCM $\mathcal{M}$ is associated with a causal diagram $\mathcal{G}$ over V , where $V \rightarrow W$ if V appears as an argument of f_W in $\mathcal{M}$, and $V \leftarrow \cdots \rightarrow W$ if $U_V \cap U_W \neq \emptyset$, i.e. V and W share an unobserved confounder. For a causal diagram $\mathcal{G}$ over V , the X -lower-manipulation of $\mathcal{G}$ deletes all those edges that are out of variables in X , and otherwise keeps $\mathcal{G}$ as it is. The resulting graph is denoted as $\mathcal{G}_X$. The X -upper-manipulation of $\mathcal{G}$ deletes all those edges that are into variables in X , and otherwise keeps $\mathcal{G}$ as it is. The resulting graph is denoted as $\mathcal{G}_{\overline{X}}$. We use $\perp_d$ to denote d -separation in causal diagrams (Pearl, 2009, Def. 1.2.3).

Theorem 8 (Inference Rules σ -calculus (Correa & Bareinboim, 2020a)). *Let $\mathcal{G}$ be a causal diagram compatible with an SCM $\mathcal{M}$, with endogenous variables V . For any disjoint subsets $X, Y, Z \subseteq V$, two disjoint subsets $T, W \subseteq V \setminus (Z \cup Y)$ (i.e., possibly including X), the following rules are valid for any intervention strategies π_X, π_Z , and π'_Z such that $\mathcal{G}_{\pi_X \pi_Z}, \mathcal{G}_{\pi_X \pi'_Z}$ have no cycles:*

- *Rule 1 (Insertion/Deletion of observations):*

$$P_{\pi_X}(y \mid w, t) = P_{\pi_X}(y \mid w) \quad \text{if } (T \perp_d Y \mid W) \text{ in } \mathcal{G}_{\pi_X}.$$

- *Rule 2 (Change of regimes under observation):*

$$P_{\pi_X, \pi_Z}(y \mid z, w) = P_{\pi_X, \pi'_Z}(y \mid z, w) \quad \text{if } (Y \perp_d Z \mid W) \text{ in } \mathcal{G}_{\pi_X, \pi_Z, \underline{Z}} \text{ and } \mathcal{G}_{\pi_X, \pi'_Z, \underline{Z}}$$

- *Rule 3 (Change of regimes without observation):*

$$P_{\pi_X, \pi_Z}(y \mid w) = P_{\pi_X, \pi'_Z}(y \mid w) \quad \text{if } (Y \perp_d Z \mid W) \text{ in } \mathcal{G}_{\pi_X, \pi_Z, \overline{Z(W)}} \text{ and } \mathcal{G}_{\pi_X, \pi'_Z, \overline{Z(W)}}$$

where $Z(W)$ is the set of elements in Z that are not ancestors of W in $\mathcal{G}_{\pi_X}$.

Lemma 1. *Let $\pi : \text{supp}_C \times \text{supp}_D \mapsto [0, 1]$ be a (probabilistic) policy mapping contexts c to decisions d . Then $P_d(V)$ may be computed from $P_\pi(V)$.*

Proof. Let $V = C \cup D \cup Y$ and $\mathcal{G}$ be an arbitrary causal diagram summarizing the SCM of the environment. The following derivation shows the claim,

$$P_d(v) = P_d(y \mid c)P_d(c) \quad \text{by the rules of total probability} \quad (47)$$

$$= P_d(y \mid c)P_\pi(c) \quad \text{by rule 3 of the } \sigma\text{-calculus since } D \perp_d C \text{ in } \mathcal{G}_{\overline{D}} \text{ and } \mathcal{G}_{\pi, \overline{D}} \quad (48)$$

$$= P_\pi(y \mid d, c)P_\pi(c) \quad \text{by rule 2 of the } \sigma\text{-calculus since } D \perp_d R \mid C \text{ in } \mathcal{G}_{\pi, \underline{D}} \quad (49)$$

That is we have shown $P_d(v)$ can be expressed as a functional of $P_\pi(v)$. Here note that the equalities hold in any causal graph $\mathcal{G}$ by definition of π . $\square$

We start by providing proofs for the results on the AI's choice of action out-of-distribution given in Sec. 4.1.

Thm. 1 restated. An AI grounded in a domain $\mathcal{M}$ is weakly predictable under a shift $\sigma := do(\mathbf{z})$, $\mathbf{Z} \subset \mathbf{V}$, in a context $\mathbf{C} = \mathbf{c}$ if and only if there exists a decision d^* such that,

$$\frac{\mathbb{E}_{P_d}[Y | \mathbf{c}, \mathbf{z}] P_d(\mathbf{c}, \mathbf{z})}{P_d(\mathbf{c}, \mathbf{z}) + 1 - P_d(\mathbf{z})} - \frac{\mathbb{E}_{P_{d^*}}[Y | \mathbf{c}, \mathbf{z}] P_{d^*}(\mathbf{c}, \mathbf{z}) + 1 - P_{d^*}(\mathbf{z})}{P_{d^*}(\mathbf{c}, \mathbf{z}) + 1 - P_{d^*}(\mathbf{z})} > 0, \quad \text{for some } d \neq d^*. \quad (50)$$

Proof. Recall that the AI is weakly predictable in a context $\mathbf{C} = \mathbf{c}$ if and only if there exists a decision d^* such that,

$$\min_{\widehat{\mathcal{M}} \in \mathbb{M}} (\Delta_{d > d^*}) > 0, \quad \Delta_{d > d^*} := \mathbb{E}_{P_{\widehat{\mathcal{M}}}}[Y | do(\sigma, d), \mathbf{c}] - \mathbb{E}_{P_{\widehat{\mathcal{M}}}}[Y | do(\sigma, d^*), \mathbf{c}], \quad \text{for some } d \neq d^*. \quad (51)$$

$\mathbb{M}$ denotes the set of compatible SCMs, i.e., that generate the data under our assumptions. Δ is the AI's preference gap between two decisions in some situation $\mathbf{C} = \mathbf{c}$. We will consider the derivation of bounds on each term of the difference in Δ separately. Firstly, note that,

$$\mathbb{E}_{\widehat{P}_{\sigma, d}}[Y | \mathbf{C} = \mathbf{c}] = \mathbb{E}_{\widehat{P}_{\mathbf{z}, d}}[Y \mathbb{1}_{\mathbf{c}}(\mathbf{C})] / \widehat{P}_{\mathbf{z}, d}(\mathbf{c}) \quad (52)$$

Analytical Lower Bound A lower bound on this ratio can be obtained by minimizing the numerator and maximizing the denominator, for example using the following derivation:

$$\mathbb{E}_{\widehat{P}_{\mathbf{z}, d}}[Y \mathbb{1}_{\mathbf{c}}(\mathbf{C})] = \sum_{\mathbf{z}} \mathbb{E}_{\widehat{P}_d}[Y_{\mathbf{z}} \mathbb{1}_{\mathbf{c}, \mathbf{z}}(\mathbf{C}_{\mathbf{z}}, \mathbf{Z})] \quad \text{marginalizing over } \mathbf{z} \quad (53)$$

$$\geq \mathbb{E}_{\widehat{P}_d}[Y_{\mathbf{z}} \mathbb{1}_{\mathbf{c}, \mathbf{z}}(\mathbf{C}_{\mathbf{z}}, \mathbf{Z})] \quad \text{since summands } > 0 \quad (54)$$

$$= \mathbb{E}_{\widehat{P}_d}[Y \mathbb{1}_{\mathbf{c}, \mathbf{z}}(\mathbf{C}, \mathbf{Z})] \quad \text{by consistency} \quad (55)$$

$$= \mathbb{E}_{P_d}[Y | \mathbf{c}, \mathbf{z}] P_d(\mathbf{c}, \mathbf{z}) \quad \text{by grounding} \quad (56)$$

$$\quad (57)$$

$$\widehat{P}_{\mathbf{z}, d}(\mathbf{c}) \stackrel{(1)}{=} 1 - \widehat{P}_{\mathbf{z}, d}(\mathbf{c}') \quad (58)$$

$$= 1 - \sum_{\mathbf{z}} \widehat{P}_d(\mathbf{c}'_{\mathbf{z}}, \mathbf{z}) \quad \text{marginalizing over } \mathbf{z} \quad (59)$$

$$\leq 1 - \widehat{P}_d(\mathbf{c}'_{\mathbf{z}}, \mathbf{z}) \quad \text{since summands } > 0 \quad (60)$$

$$= \widehat{P}_d(\mathbf{c}, \mathbf{z}) + 1 - \widehat{P}_d(\mathbf{z}) \quad \text{by consistency} \quad (61)$$

$$= P_d(\mathbf{c}, \mathbf{z}) + 1 - \widehat{P}_d(\mathbf{z}) \quad \text{by grounding.} \quad (62)$$

(1) holds by defining $\mathbf{c}'$ to stand for any combination of variables $\mathbf{C} \setminus \mathbf{Z}$ other than $\mathbf{c} \setminus \mathbf{z}$.

This implies then that,

$$\mathbb{E}_{\widehat{P}_{\sigma, d}}[Y | \mathbf{C} = \mathbf{c}] \geq \frac{\mathbb{E}_{P_d}[Y | \mathbf{c}, \mathbf{z}] P_d(\mathbf{c}, \mathbf{z})}{P_d(\mathbf{c}, \mathbf{z}) + 1 - P_d(\mathbf{z})}. \quad (63)$$

Analytical Upper Bound For the upper bound, we start by noting that,

$$\mathbb{E}_{\widehat{P}_{\sigma, d}}[Y | \mathbf{C} = \mathbf{c}] = 1 - \mathbb{E}_{\widehat{P}_{\sigma, d}}[1 - Y | \mathbf{C} = \mathbf{c}] \quad (64)$$

$$= 1 - \mathbb{E}_{\widehat{P}_{\mathbf{z}, d}}[(1 - Y) \mathbb{1}_{\mathbf{c}}(\mathbf{C})] / \widehat{P}_{\mathbf{z}, d}(\mathbf{c}) \quad (65)$$

Leveraging the bounds derived above we obtain,

$$\mathbb{E}_{\widehat{P}_{\sigma, d}}[Y | \mathbf{C} = \mathbf{c}] \leq 1 - \frac{\mathbb{E}_{P_d}[(1 - Y) \mathbb{1}_{\mathbf{c}, \mathbf{z}}(\mathbf{C}, \mathbf{Z})]}{P_d(\mathbf{c}, \mathbf{z}) + P_d(\mathbf{z}')} \quad (66)$$

$$= \frac{\mathbb{E}_{P_d}[Y | \mathbf{c}, \mathbf{z}] P_d(\mathbf{c}, \mathbf{z}) + 1 - P_d(\mathbf{z})}{P_d(\mathbf{c}, \mathbf{z}) + 1 - P_d(\mathbf{z})} \quad (67)$$

By setting $d = d_1$ in the lower bound and $d = d_0$ in the upper bound of the expected utility, we obtain a lower bound on the difference of expected utilities:

$$\Delta_{d_1 > d_0} \geq \frac{\mathbb{E}_{P_{d_1}}[Y | \mathbf{c}, \mathbf{z}] P_{d_1}(\mathbf{c}, \mathbf{z})}{P_{d_1}(\mathbf{c}, \mathbf{z}) + 1 - P_{d_1}(\mathbf{z})} - \frac{\mathbb{E}_{P_{d_0}}[Y | \mathbf{c}, \mathbf{z}] P_{d_0}(\mathbf{c}, \mathbf{z}) + P_{d_0}(\mathbf{z}')}{P_{d_0}(\mathbf{c}, \mathbf{z}) + 1 - P_{d_0}(\mathbf{z})}. \quad (68)$$

And similarly, by setting $d = d_1$ in the upper bound and $d = d_0$ in the lower bound of the expected utility, we obtain an upper bound on the difference of expected utilities:

$$\Delta_{d_1 > d_0} \leq \frac{\mathbb{E}_{P_{d_1}}[Y | \mathbf{c}, \mathbf{z}] P_{d_1}(\mathbf{c}, \mathbf{z}) + 1 - P_{d_1}(\mathbf{z})}{P_{d_1}(\mathbf{c}, \mathbf{z}) + 1 - P_{d_1}(\mathbf{z})} - \frac{\mathbb{E}_{P_{d_0}}[Y | \mathbf{c}, \mathbf{z}] P_{d_0}(\mathbf{c}, \mathbf{z})}{P_{d_0}(\mathbf{c}, \mathbf{z}) + 1 - P_{d_0}(\mathbf{z})}. \quad (69)$$

We now show that these bounds are tight by constructing SCMs (that is, possible world models of the AI system) that evaluate to the lower and upper bounds while generating the distribution of agent interactions $\hat{P}_{d_1}, \hat{P}_{d_0}$.

Tightness Lower Bound for Δ For the lower bound we will consider the following SCM,

$$\mathcal{M}_d^1 =: \begin{cases} Z \leftarrow f_Z(\mathbf{u}) \\ C \leftarrow \begin{cases} f_C(\mathbf{u}, \mathbf{z}) & \text{if } f_Z(\mathbf{u}) = \mathbf{z} \\ 1 & \text{otherwise.} \end{cases} \\ D \leftarrow d \\ Y \leftarrow \begin{cases} f_Y(d, \mathbf{c}, \mathbf{z}, \mathbf{u}) & \text{if } f_Z(\mathbf{u}) = \mathbf{z} \\ 1 & \text{if } f_Z(\mathbf{u}) \neq \mathbf{z}, d = d_0 \\ 0 & \text{if } f_Z(\mathbf{u}) \neq \mathbf{z}, d = d_1 \end{cases} \\ P(U) \end{cases} \quad (70)$$

Here $\{f_Z, f_C, f_Y, \mathcal{U}, P(U)\}$ are chosen to match the observed trajectory of agent interactions, i.e., such that $P^{\mathcal{M}_d^1}(\mathbf{v}) = P^{\hat{\mathcal{M}}_d}(\mathbf{v})$ for all $\mathbf{v} \in \text{supp}_{\mathbf{V}}$. Consider evaluating,

$$\mathbb{E}_{P^{\mathcal{M}_{\sigma,d}^1}}[Y | C = \mathbf{c}] = \mathbb{E}_{P^{\mathcal{M}_{\sigma,d}^1}}[Y \mathbb{1}_{\mathbf{c}}(C)] / P^{\mathcal{M}_{\sigma,d}^1}(\mathbf{c}) \quad (71)$$

The numerator (under $\mathcal{M}_{d_1}^1$) evaluates to,

$$\mathbb{E}_{P^{\mathcal{M}_{d_1}^1}}[Y \mathbb{1}_{\mathbf{c}}(C_{\mathbf{z}})] \quad (72)$$

$$= \sum_{\mathbf{u}} \mathbb{E}_{P^{\mathcal{M}_{d_1}^1}}[Y \mathbb{1}_{\mathbf{c}}(C_{\mathbf{z}}) | \mathbf{u}] P^{\mathcal{M}_{d_1}^1}(\mathbf{u}) \quad (73)$$

$$= \sum_{\mathbf{u}} \mathbb{E}_{P^{\mathcal{M}_{d_1}^1}}[Y \mathbb{1}_{\mathbf{c}}(C) | \mathbf{z}, \mathbf{u}] P^{\mathcal{M}_{d_1}^1}(\mathbf{u}) \quad (74)$$

$$= \mathbb{E}_{P^{\mathcal{M}_{d_1}^1}}[Y \mathbb{1}_{\mathbf{c}}(C) | \mathbf{z}, \{\mathbf{u} : f_Z(\mathbf{u}) = \mathbf{z}\}] P^{\mathcal{M}_{d_1}^1}(\{\mathbf{u} : f_Z(\mathbf{u}) = \mathbf{z}\}) \quad (75)$$

$$+ \mathbb{E}_{P^{\mathcal{M}_{d_1}^1}}[Y \mathbb{1}_{\mathbf{c}}(C) | \mathbf{z}, \{\mathbf{u} : f_Z(\mathbf{u}) \neq \mathbf{z}\}] P^{\mathcal{M}_{d_1}^1}(\{\mathbf{u} : f_Z(\mathbf{u}) \neq \mathbf{z}\}) \quad (76)$$

$$= \mathbb{E}_{P^{\mathcal{M}_{d_1}^1}}[Y \mathbb{1}_{\mathbf{c}}(C) | \mathbf{z}] P^{\mathcal{M}_{d_1}^1}(\mathbf{z}) \quad (77)$$

$$= \mathbb{E}_{P^{\mathcal{M}_{d_1}^1}}[Y | \mathbf{c}, \mathbf{z}] P^{\mathcal{M}_{d_1}^1}(\mathbf{c}, \mathbf{z}) \quad (78)$$

The denominator under $\mathcal{M}_{d_1}^1$ evaluates to,

$$P^{\mathcal{M}_{\sigma, d_1}^1}(c) = \sum_{\mathbf{u}} P^{\mathcal{M}_{d_1}^1}(c_{\mathbf{z}} | \mathbf{u}) P^{\mathcal{M}_{d_1}^1}(\mathbf{u}) \quad (79)$$

$$= \sum_{\mathbf{u}} P^{\mathcal{M}_{d_1}^1}(c | \mathbf{z}, \mathbf{u}) P^{\mathcal{M}_{d_1}^1}(\mathbf{u}) \quad (80)$$

$$= P^{\mathcal{M}_{d_1}^1}(c | \mathbf{z}, \{\mathbf{u} : f_{\mathbf{Z}}(\mathbf{u}) = \mathbf{z}\}) P^{\mathcal{M}_{d_1}^1}(\{\mathbf{u} : f_{\mathbf{Z}}(\mathbf{u}) = \mathbf{z}\}) \quad (81)$$

$$+ P^{\mathcal{M}_{d_1}^1}(c | \mathbf{z}, \{\mathbf{u} : f_{\mathbf{Z}}(\mathbf{u}) \neq \mathbf{z}\}) P^{\mathcal{M}_{d_1}^1}(\{\mathbf{u} : f_{\mathbf{Z}}(\mathbf{u}) \neq \mathbf{z}\}) \quad (82)$$

$$= P^{\mathcal{M}_{d_1}^1}(c | \mathbf{z}) P^{\mathcal{M}_{d_1}^1}(\mathbf{z}) + 1 - P^{\mathcal{M}_{d_1}^1}(\mathbf{z}) \quad (83)$$

$$= P^{\mathcal{M}_{d_1}^1}(c, \mathbf{z}) + 1 - P^{\mathcal{M}_{d_1}^1}(\mathbf{z}) \quad (84)$$

The numerator under $\mathcal{M}_{d_0}^1$ evaluates to,

$$\mathbb{E}_{P^{\mathcal{M}_{d_0}^1}}[Y_{\mathbf{z}} \mathbb{1}_c(C_{\mathbf{z}})] \quad (85)$$

$$= \sum_{\mathbf{u}} \mathbb{E}_{P^{\mathcal{M}_{d_0}^1}}[Y_{\mathbf{z}} \mathbb{1}_c(C_{\mathbf{z}}) | \mathbf{u}] P^{\mathcal{M}_{d_0}^1}(\mathbf{u}) \quad (86)$$

$$= \sum_{\mathbf{u}} \mathbb{E}_{P^{\mathcal{M}_{d_0}^1}}[Y \mathbb{1}_c(C) | \mathbf{z}, \mathbf{u}] P^{\mathcal{M}_{d_0}^1}(\mathbf{u}) \quad (87)$$

$$= \mathbb{E}_{P^{\mathcal{M}_{d_0}^1}}[Y \mathbb{1}_c(C) | \mathbf{z}, \{\mathbf{u} : f_{\mathbf{Z}}(\mathbf{u}) = \mathbf{z}\}] P^{\mathcal{M}_{d_0}^1}(\{\mathbf{u} : f_{\mathbf{Z}}(\mathbf{u}) = \mathbf{z}\}) \quad (88)$$

$$+ \mathbb{E}_{P^{\mathcal{M}_{d_0}^1}}[Y \mathbb{1}_c(C) | \mathbf{z}, \{\mathbf{u} : f_{\mathbf{Z}}(\mathbf{u}) \neq \mathbf{z}\}] P^{\mathcal{M}_{d_0}^1}(\{\mathbf{u} : f_{\mathbf{Z}}(\mathbf{u}) \neq \mathbf{z}\}) \quad (89)$$

$$= \mathbb{E}_{P^{\mathcal{M}_{d_0}^1}}[Y \mathbb{1}_c(C) | \mathbf{z}] P^{\mathcal{M}_{d_0}^1}(\mathbf{z}) + 1 - P^{\mathcal{M}_{d_0}^1}(\mathbf{z}) \quad (90)$$

$$= \mathbb{E}_{P^{\mathcal{M}_{d_0}^1}}[Y | c, \mathbf{z}] P^{\mathcal{M}_{d_0}^1}(c, \mathbf{z}) + 1 - P^{\mathcal{M}_{d_0}^1}(\mathbf{z}) \quad (91)$$

The denominator under $\mathcal{M}_{d_0}^1$ evaluates to,

$$P^{\mathcal{M}_{\sigma, d_0}^1}(c) = \sum_{\mathbf{u}} P^{\mathcal{M}_{d_0}^1}(c_{\mathbf{z}} | \mathbf{u}) P^{\mathcal{M}_{d_0}^1}(\mathbf{u}) \quad (92)$$

$$= \sum_{\mathbf{u}} P^{\mathcal{M}_{d_0}^1}(c | \mathbf{z}, \mathbf{u}) P^{\mathcal{M}_{d_0}^1}(\mathbf{u}) \quad (93)$$

$$= P^{\mathcal{M}_{d_0}^1}(c | \mathbf{z}, \{\mathbf{u} : f_{\mathbf{Z}}(\mathbf{u}) = \mathbf{z}\}) P^{\mathcal{M}_{d_0}^1}(\{\mathbf{u} : f_{\mathbf{Z}}(\mathbf{u}) = \mathbf{z}\}) \quad (94)$$

$$+ P^{\mathcal{M}_{d_0}^1}(c | \mathbf{z}, \{\mathbf{u} : f_{\mathbf{Z}}(\mathbf{u}) \neq \mathbf{z}\}) P^{\mathcal{M}_{d_0}^1}(\{\mathbf{u} : f_{\mathbf{Z}}(\mathbf{u}) \neq \mathbf{z}\}) \quad (95)$$

$$= P^{\mathcal{M}_{d_0}^1}(c | \mathbf{z}) P^{\mathcal{M}_{d_0}^1}(\mathbf{z}) + 1 - P^{\mathcal{M}_{d_0}^1}(\mathbf{z}) \quad (96)$$

$$= P^{\mathcal{M}_{d_0}^1}(c, \mathbf{z}) + 1 - P^{\mathcal{M}_{d_0}^1}(\mathbf{z}) \quad (97)$$

Combining these results we get the analytical lower bound:

$$\Delta_{d_1 > d_0} = \frac{\mathbb{E}_{P_{d_1}}[Y | c, \mathbf{z}] P_{d_1}(c, \mathbf{z})}{P_{d_1}(c, \mathbf{z}) + 1 - P_{d_1}(\mathbf{z})} - \frac{\mathbb{E}_{P_{d_0}}[Y | c, \mathbf{z}] P_{d_0}(c, \mathbf{z}) + 1 - P_{d_0}(\mathbf{z})}{P_{d_0}(c, \mathbf{z}) + 1 - P_{d_0}(\mathbf{z})}. \quad (98)$$

This shows that for a given $C = c$ and pair of decisions (d_1, d_0) we can always find an SCM that evaluates to the lower bound that we report. So if, and only if, we can find a decision d^* such that the lower bound can be evaluated to be greater than zero for some $d \neq d^*$ will the AI be weakly predictable, as claimed. $\square$

Corol. 1 restated. Given a discrepancy measure ψ , an AI approximately grounded in a domain $\mathcal{M}$ is weakly predictable in a context $C = c$ under a shift $\sigma := do(\mathbf{z})$, $\mathbf{Z} \subset \mathbf{V}$, if and only if there exists a decision d^* such that,

$$\min_{\hat{P}: \psi(\hat{P}, P) \leq \delta} \left\{ \frac{\mathbb{E}_{\hat{P}_d}[Y | c, \mathbf{z}] \hat{P}_d(c, \mathbf{z})}{\hat{P}_d(c, \mathbf{z}) + 1 - \hat{P}_d(\mathbf{z})} - \frac{\mathbb{E}_{\hat{P}_{d^*}}[Y | c, \mathbf{z}] \hat{P}_{d^*}(c, \mathbf{z}) + 1 - \hat{P}_{d^*}(\mathbf{z})}{\hat{P}_{d^*}(c, \mathbf{z}) + 1 - \hat{P}_{d^*}(\mathbf{z})} \right\} > 0, \quad \text{for some } d \neq d^*. \quad (99)$$

Proof. For approximately grounded AI systems, we can state the bound from Thm. 1 as,

$$\min_{\mathcal{M} \in \mathbb{M}} (\Delta_{d > d^*}) = \frac{\mathbb{E}_{\hat{P}_d}[Y | \mathbf{c}, \mathbf{z}] \hat{P}_d(\mathbf{c}, \mathbf{z})}{\hat{P}_d(\mathbf{c}, \mathbf{z}) + 1 - \hat{P}_d(\mathbf{z})} - \frac{\mathbb{E}_{\hat{P}_{d^*}}[Y | \mathbf{c}, \mathbf{z}] \hat{P}_{d^*}(\mathbf{c}, \mathbf{z}) + 1 - \hat{P}_{d^*}(\mathbf{z})}{\hat{P}_{d^*}(\mathbf{c}, \mathbf{z}) + 1 - \hat{P}_{d^*}(\mathbf{z})}. \quad (100)$$

$\hat{P}_d$ is constrained to be close to P_d according to distance ψ and threshold δ . We get valid bounds by reporting the worst-case bounds under this looser constraint:

$$\min_{\mathcal{M} \in \mathbb{M}} (\Delta_{d > d^*}) = \min_{\hat{P}: \psi(\hat{P}, P) \leq \delta} \left\{ \frac{\mathbb{E}_{\hat{P}_d}[Y | \mathbf{c}, \mathbf{z}] \hat{P}_d(\mathbf{c}, \mathbf{z})}{P_d(\mathbf{c}, \mathbf{z}) + 1 - P_d(\mathbf{z})} - \frac{\mathbb{E}_{\hat{P}_{d^*}}[Y | \mathbf{c}, \mathbf{z}] \hat{P}_{d^*}(\mathbf{c}, \mathbf{z}) + 1 - \hat{P}_{d^*}(\mathbf{z})}{\hat{P}_{d^*}(\mathbf{c}, \mathbf{z}) + 1 - \hat{P}_{d^*}(\mathbf{z})} \right\}. \quad (101)$$

This shows that for a given $\mathbf{C} = \mathbf{c}$, the $\min_{\mathcal{M} \in \mathbb{M}} (\Delta_{d > d^*}) > 0$ for some $d \neq d^*$ if and only if,

$$\min_{\hat{P}: \psi(\hat{P}, P) \leq \delta} \left\{ \frac{\mathbb{E}_{\hat{P}_d}[Y | \mathbf{c}, \mathbf{z}] \hat{P}_d(\mathbf{c}, \mathbf{z})}{P_d(\mathbf{c}, \mathbf{z}) + 1 - P_d(\mathbf{z})} - \frac{\mathbb{E}_{\hat{P}_{d^*}}[Y | \mathbf{c}, \mathbf{z}] \hat{P}_{d^*}(\mathbf{c}, \mathbf{z}) + 1 - \hat{P}_{d^*}(\mathbf{z})}{\hat{P}_{d^*}(\mathbf{c}, \mathbf{z}) + 1 - \hat{P}_{d^*}(\mathbf{z})} \right\} > 0. \quad (102)$$

□

Thm. 2 restated. Let $\sigma := do(\mathbf{z})$ be a shift on a set of variables $\mathbf{Z} \subset \mathbf{V}$. For $\mathbf{R}_i \subset \mathbf{Z} \subset \mathbf{V}, i = 1, \dots, k$, consider an AI grounded in multiple domains $\{\mathcal{M}_{\mathbf{r}_i} : i = 1, \dots, k\}$. The AI is weakly predictable in a context $\mathbf{C} = \mathbf{c}$ under a shift $\sigma := do(\mathbf{z})$ if and only if there exists a decision d^* such that,

$$\max_{i,j=1,\dots,k} A(\mathbf{r}_i, \mathbf{r}_j) > 0, \quad \text{for some } d \neq d^*, \quad (103)$$

where

$$A(\mathbf{r}_i, \mathbf{r}_j) := \frac{\mathbb{E}_{P_{d,\mathbf{r}_i}}[Y | \mathbf{c}, \mathbf{z} \setminus \mathbf{r}_i] P_{d,\mathbf{r}_i}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}_i)}{P_{d,\mathbf{r}_i}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}_i) + 1 - P_{d,\mathbf{r}_i}(\mathbf{z} \setminus \mathbf{r}_i)} - \frac{\mathbb{E}_{P_{d^*,\mathbf{r}_j}}[Y | \mathbf{c}, \mathbf{z} \setminus \mathbf{r}_j] P_{d^*,\mathbf{r}_j}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}_j) + 1 - P_{d^*,\mathbf{r}_j}(\mathbf{z} \setminus \mathbf{r}_j)}{P_{d^*,\mathbf{r}_j}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}_j) + 1 - P_{d^*,\mathbf{r}_j}(\mathbf{z} \setminus \mathbf{r}_j)}.$$

Proof. $\{\mathcal{M}_{\mathbf{r}_i} : i = 1, \dots, k\}$ describes k domains in which experiments on different subsets of $\mathbf{Z}$ have been conducted. This includes possibly the null experiment $\mathbf{R}_i = \emptyset$ that refers to the unaltered domain $\mathcal{M}$.

We can use a similar derivation to that of Thm. 1 to derive bounds on Δ under a shift $\sigma := do(\mathbf{z})$ in terms of $P_{d,\mathbf{r}}(\mathbf{V}), \mathbf{R} \in \mathbf{V}$ and obtain,

$$\Delta_{d_1 > d_0} \geq A(\mathbf{r}) \quad (104)$$

where,

$$A(\mathbf{r}) := \frac{\mathbb{E}_{P_{d_1,\mathbf{r}}}[Y | \mathbf{c}, \mathbf{z} \setminus \mathbf{r}] P_{d_1,\mathbf{r}}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r})}{P_{d_1,\mathbf{r}}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}) + 1 - P_{d_1,\mathbf{r}}(\mathbf{z} \setminus \mathbf{r})} - \frac{\mathbb{E}_{P_{d_0,\mathbf{r}}}[Y | \mathbf{c}, \mathbf{z} \setminus \mathbf{r}] P_{d_0,\mathbf{r}}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}) + 1 - P_{d_0,\mathbf{r}}(\mathbf{z} \setminus \mathbf{r})}{P_{d_0,\mathbf{r}}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}) + 1 - P_{d_0,\mathbf{r}}(\mathbf{z} \setminus \mathbf{r})}. \quad (105)$$

These bounds can be shown to be tight by constructing similar SCMs. For example, for the analytical lower bound consider,

$$\mathcal{M}_{d,\mathbf{r}}^1 =: \begin{cases} \mathbf{S} \leftarrow f_{\mathbf{S}}(\mathbf{u}) \\ \mathbf{R} \leftarrow \mathbf{r} \\ \mathbf{C} \leftarrow \begin{cases} f_{\mathbf{C}}(\mathbf{u}, \mathbf{s}, \mathbf{r}) & \text{if } f_{\mathbf{S}}(\mathbf{u}) = \mathbf{s} \\ 1 & \text{otherwise.} \end{cases} \\ \mathbf{D} \leftarrow d \\ \mathbf{Y} \leftarrow \begin{cases} f_{\mathbf{Y}}(d, \mathbf{c}, \mathbf{s}, \mathbf{r}, \mathbf{u}) & \text{if } f_{\mathbf{S}}(\mathbf{u}) = \mathbf{s} \\ 1 & \text{if } f_{\mathbf{S}}(\mathbf{u}) \neq \mathbf{s}, d = d_0 \\ 0 & \text{if } f_{\mathbf{S}}(\mathbf{u}) \neq \mathbf{s}, d = d_1 \end{cases} \\ P(\mathbf{U}) \end{cases} \quad (106)$$

where $\mathbf{S} = \mathbf{Z} \setminus \mathbf{R}$. Here $\{f_Z, f_C, f_Y, \mathcal{U}, P(\mathbf{U})\}$ are chosen to match the observed trajectory of agent interactions, i.e., such that $P^{\mathcal{M}_{d,r}^1}(\mathbf{v}) = P^{\widehat{\mathcal{M}}_{d,r}}(\mathbf{v})$ for all $\mathbf{v} \in \text{supp}_{\mathbf{V}}$. We could verify that this SCM evaluates to the lower bound above.

If we have multiple domains with different set of intervened variables $\{\mathbf{R}_i : i = 1, \dots, k\}$ we could use this construction to find a lower bound using samples from $\{P_{d,r_i}(\mathbf{V}) : i = 1, \dots, k\}$. A lower bound that can be constructed for an AI system grounded in $\{\mathcal{M}_{r_i} : i = 1, \dots, k\}$ is,

$$\Delta_{d_1 > d_0} \geq \max_{i,j=1,\dots,k} A(\mathbf{r}_i, \mathbf{r}_j) \quad (107)$$

where

$$A(\mathbf{r}_i, \mathbf{r}_j) := \frac{\mathbb{E}_{P_{d_1, \mathbf{r}_i}}[Y \mid \mathbf{c}, \mathbf{z} \setminus \mathbf{r}_i] P_{d_1, \mathbf{r}_i}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}_i)}{P_{d_1, \mathbf{r}_i}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}_i) + 1 - P_{d_1, \mathbf{r}_i}(\mathbf{z} \setminus \mathbf{r}_i)} - \frac{\mathbb{E}_{P_{d_0, \mathbf{r}_j}}[Y \mid \mathbf{c}, \mathbf{z} \setminus \mathbf{r}_j] P_{d_0, \mathbf{r}_j}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}_j) + 1 - P_{d_0, \mathbf{r}_j}(\mathbf{z} \setminus \mathbf{r}_j)}{P_{d_0, \mathbf{r}_j}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}_j) + 1 - P_{d_0, \mathbf{r}_j}(\mathbf{z} \setminus \mathbf{r}_j)}. \quad (108)$$

The intuition here is that we have multiple lower bounds for the preference gap, then the best lower bound can be taken to be the largest of the multiple lower bounds available.

We can show that this bound is tight in the case where the AI is grounded in two environments $\{\mathcal{M}_{r_1}, \mathcal{M}_{r_2}\}$ under a shift $\sigma := do(\mathbf{z}), \mathbf{Z} = \mathbf{R}_1 \cup \mathbf{R}_2$. According to the inequality above, we have simultaneously,

$$\Delta_{d_1 > d_0} \geq A(\mathbf{r}_1, \mathbf{r}_1), A(\mathbf{r}_1, \mathbf{r}_2), A(\mathbf{r}_2, \mathbf{r}_1), A(\mathbf{r}_2, \mathbf{r}_2). \quad (109)$$

Each of these terms can be evaluated from the available data sampled from $\{P_{d,r_1}, P_{d,r_2}\}$. Note that both $A(\mathbf{r}_1, \mathbf{r}_1)$ and $A(\mathbf{r}_2, \mathbf{r}_2)$ can be obtained with the SCM above. Without loss of generality, assume that $A(\mathbf{r}_1, \mathbf{r}_2) \geq A(\mathbf{r}_2, \mathbf{r}_1), A(\mathbf{r}_1, \mathbf{r}_1), A(\mathbf{r}_2, \mathbf{r}_2)$. We will show that we can construct an SCM compatible with $\{P_{d,r_1}, P_{d,r_2}\}$ that evaluates to $A(\mathbf{r}_1, \mathbf{r}_2)$ demonstrating that the bound is tight.

Consider the following SCM:

$$\mathcal{M}_d =: \begin{cases} \mathbf{R}_1 \leftarrow f_{\mathbf{R}_1}(\mathbf{u}_1) \\ \mathbf{R}_2 \leftarrow f_{\mathbf{R}_2}(\mathbf{u}_2) \\ \mathbf{C} \leftarrow \begin{cases} f_C(\mathbf{r}_1, \mathbf{r}_2, \mathbf{u}_1, \mathbf{u}_2) & \text{if } f_{\mathbf{R}_1}(\mathbf{u}_1) = \mathbf{r}_1, f_{\mathbf{R}_2}(\mathbf{u}_2) = \mathbf{r}_2 \\ 1 & \text{otherwise.} \end{cases} \\ \mathbf{D} \leftarrow d \\ \mathbf{Y} \leftarrow \begin{cases} f_Y(d, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2, \mathbf{u}_1, \mathbf{u}_2) & \text{if } f_{\mathbf{R}_1}(\mathbf{u}_1) = \mathbf{r}_1, f_{\mathbf{R}_2}(\mathbf{u}_2) = \mathbf{r}_2 \\ f_Y(d, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2, \mathbf{u}_1, \mathbf{u}_2) & \text{if } d = d_1, f_{\mathbf{R}_1}(\mathbf{u}_1) \neq \mathbf{r}_1, f_{\mathbf{R}_2}(\mathbf{u}_2) = \mathbf{r}_2 \\ f_Y(d, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2, \mathbf{u}_1, \mathbf{u}_2) & \text{if } d = d_0, f_{\mathbf{R}_1}(\mathbf{u}_1) = \mathbf{r}_1, f_{\mathbf{R}_2}(\mathbf{u}_2) \neq \mathbf{r}_2 \\ 0 & \text{if } d = d_1, f_{\mathbf{R}_1}(\mathbf{u}_1) = \mathbf{r}_1, f_{\mathbf{R}_2}(\mathbf{u}_2) \neq \mathbf{r}_2 \\ 0 & \text{if } d = d_1, f_{\mathbf{R}_1}(\mathbf{u}_1) \neq \mathbf{r}_1, f_{\mathbf{R}_2}(\mathbf{u}_2) \neq \mathbf{r}_2 \\ 1 & \text{if } d = d_0, f_{\mathbf{R}_1}(\mathbf{u}_1) \neq \mathbf{r}_1, f_{\mathbf{R}_2}(\mathbf{u}_2) = \mathbf{r}_2 \\ 1 & \text{if } d = d_0, f_{\mathbf{R}_1}(\mathbf{u}_1) \neq \mathbf{r}_1, f_{\mathbf{R}_2}(\mathbf{u}_2) \neq \mathbf{r}_2 \end{cases} \\ P(\mathbf{U}) \end{cases} \quad (110)$$

Notice that in $\mathcal{M}_d$ different choices of functional assignments “ f ” and $P(\mathbf{u})$ can generate any distribution $\{P_{d_1, \mathbf{r}_1}, P_{d_0, \mathbf{r}_2}\}$. That is this SCM (or a member of this family of SCMs) is compatible with the observed data.

Consider evaluating $A(\mathbf{r}_1, \mathbf{r}_2)$ under this SCM. Note that the derivations for the denominators are equivalent to those shown

in the proof of Thm. 1 so we will omit them here. The first term in the numerator,

$$\mathbb{E}_{P^{\mathcal{M}_{d_1}, r_1, r_2}} [Y \mathbb{1}_c(C)] \quad (111)$$

$$= \sum_{\mathbf{u}_2} \mathbb{E}_{P^{\mathcal{M}_{d_1}, r_1, r_2}} [Y \mathbb{1}_c(C) \mid \mathbf{u}_2] P^{\mathcal{M}_{d_1}, r_1, r_2}(\mathbf{u}_2) \quad (112)$$

$$= \sum_{\mathbf{u}_2} \mathbb{E}_{P^{\mathcal{M}_{d_1}, r_1}} [Y \mathbb{1}_c(C) \mid \mathbf{r}_2, \mathbf{u}_2] P^{\mathcal{M}_{d_1}, r_1}(\mathbf{u}_2) \quad (113)$$

$$= \mathbb{E}_{P^{\mathcal{M}_{d_1}, r_1}} [Y \mathbb{1}_c(C) \mid \mathbf{r}_2, \{ \mathbf{u} : f_{R_2}(\mathbf{u}_2) = \mathbf{r}_2 \}] P^{\mathcal{M}_{d_1}, r_1}(\{ \mathbf{u}_2 : f_{R_2}(\mathbf{u}_2) = \mathbf{r}_2 \}) \quad (114)$$

$$+ \mathbb{E}_{P^{\mathcal{M}_{d_1}, r_1}} [Y \mathbb{1}_c(C) \mid \mathbf{r}_2, \{ \mathbf{u} : f_{R_2}(\mathbf{u}_2) \neq \mathbf{r}_2 \}] P^{\mathcal{M}_{d_1}, r_1}(\{ \mathbf{u}_2 : f_{R_2}(\mathbf{u}_2) \neq \mathbf{r}_2 \}) \quad (115)$$

$$= \mathbb{E}_{P^{\mathcal{M}_{d_1}, r_1}} [Y \mathbb{1}_c(C) \mid \mathbf{r}_2] P^{\mathcal{M}_{d_1}, r_1}(\mathbf{r}_2) \quad (116)$$

$$= \mathbb{E}_{P^{\mathcal{M}_{d_1}, r_1}} [Y \mid \mathbf{c}, \mathbf{r}_2] P^{\mathcal{M}_{d_1}, r_1}(\mathbf{c}, \mathbf{r}_2) \quad (117)$$

The second term in the numerator is,

$$\mathbb{E}_{P^{\mathcal{M}_{d_0}, r_1, r_2}} [Y \mathbb{1}_c(C)] \quad (118)$$

$$= \sum_{\mathbf{u}_1} \mathbb{E}_{P^{\mathcal{M}_{d_0}, r_1, r_2}} [Y \mathbb{1}_c(C) \mid \mathbf{u}_1] P^{\mathcal{M}_{d_0}, r_1, r_2}(\mathbf{u}_1) \quad (119)$$

$$= \sum_{\mathbf{u}_1} \mathbb{E}_{P^{\mathcal{M}_{d_0}, r_2}} [Y \mathbb{1}_c(C) \mid \mathbf{r}_1, \mathbf{u}_1] P^{\mathcal{M}_{d_0}, r_2}(\mathbf{u}_1) \quad (120)$$

$$= \mathbb{E}_{P^{\mathcal{M}_{d_0}, r_2}} [Y \mathbb{1}_c(C) \mid \mathbf{r}_1, \{ \mathbf{u} : f_{R_1}(\mathbf{u}_1) = \mathbf{r}_1 \}] P^{\mathcal{M}_{d_0}, r_2}(\{ \mathbf{u}_1 : f_{R_1}(\mathbf{u}_1) = \mathbf{r}_1 \}) \quad (121)$$

$$+ \mathbb{E}_{P^{\mathcal{M}_{d_0}, r_2}} [Y \mathbb{1}_c(C) \mid \mathbf{r}_1, \{ \mathbf{u} : f_{R_1}(\mathbf{u}_1) \neq \mathbf{r}_1 \}] P^{\mathcal{M}_{d_0}, r_2}(\{ \mathbf{u}_1 : f_{R_1}(\mathbf{u}_1) \neq \mathbf{r}_1 \}) \quad (122)$$

$$= \mathbb{E}_{P^{\mathcal{M}_{d_0}, r_2}} [Y \mathbb{1}_c(C) \mid \mathbf{r}_1] P^{\mathcal{M}_{d_0}, r_2}(\mathbf{r}_1) + 1 - P^{\mathcal{M}_{d_0}, r_2}(\mathbf{r}_1) \quad (123)$$

$$= \mathbb{E}_{P^{\mathcal{M}_{d_0}, r_2}} [Y \mid \mathbf{c}, \mathbf{r}_1] P^{\mathcal{M}_{d_0}, r_2}(\mathbf{c}, \mathbf{r}_1) + 1 - P^{\mathcal{M}_{d_0}, r_2}(\mathbf{r}_1) \quad (124)$$

Combining these results we get that under $\mathcal{M}$,

$$\Delta_{d_1 > d_0} = A(\mathbf{r}_1, \mathbf{r}_2). \quad (125)$$

□

Corollary 3. *The bound from multiple domains in Thm. 2 will be at least as informative as the bound from a single domain in Thm. 1.*

Proof. We claim here that for any $\mathbf{R} \subset \mathbf{Z}$,

$$A(\emptyset) \leq A(\mathbf{r}) \quad (126)$$

This means that the bounds on Δ that we can obtain from an AI system grounded in $\mathcal{M}_{\mathbf{r}}$ are more informative than the bounds obtained from an AI system grounded in $\mathcal{M}$. A is a difference of two terms written $A(\mathbf{r}) = A_1(\mathbf{r}) - A_2(\mathbf{r})$.

$$A_1(\mathbf{r}) := \frac{\mathbb{E}_{P_{d_1, \mathbf{r}}} [Y \mid \mathbf{c}, \mathbf{z} \setminus \mathbf{r}] P_{d_1, \mathbf{r}}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r})}{P_{d_1, \mathbf{r}}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}) + 1 - P_{d_1, \mathbf{r}}(\mathbf{z} \setminus \mathbf{r})} \quad (127)$$

$$A_2(\mathbf{r}) := \frac{\mathbb{E}_{P_{d_0, \mathbf{r}}} [Y \mid \mathbf{c}, \mathbf{z} \setminus \mathbf{r}] P_{d_0, \mathbf{r}}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}) + 1 - P_{d_0, \mathbf{r}}(\mathbf{z} \setminus \mathbf{r})}{P_{d_0, \mathbf{r}}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}) + 1 - P_{d_0, \mathbf{r}}(\mathbf{z} \setminus \mathbf{r})}. \quad (128)$$

It holds that $A_1(\mathbf{r}) \geq A_1(\emptyset)$, $A_2(\mathbf{r}) \leq A_2(\emptyset)$ which then implies $A(\mathbf{r}) \geq A(\emptyset)$. To see this notice that,

$$A_1(\mathbf{r}) := \frac{\mathbb{E}_{P_{d_1, \mathbf{r}}} [Y \mid \mathbf{c}, \mathbf{z} \setminus \mathbf{r}] P_{d_1, \mathbf{r}}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r})}{P_{d_1, \mathbf{r}}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}) + 1 - P_{d_1, \mathbf{r}}(\mathbf{z} \setminus \mathbf{r})} \quad (129)$$

$$\geq \frac{\mathbb{E}_{P_{d_1}} [Y \mid \mathbf{c}, \mathbf{z}] P_{d_1}(\mathbf{c}, \mathbf{z})}{P_{d_1, \mathbf{r}}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}) + 1 - P_{d_1, \mathbf{r}}(\mathbf{z} \setminus \mathbf{r})} \quad (130)$$

$$= \frac{\mathbb{E}_{P_{d_1}} [Y \mid \mathbf{c}, \mathbf{z}] P_{d_1}(\mathbf{c}, \mathbf{z})}{1 - P_{d_1, \mathbf{r}}(\tilde{\mathbf{c}}, \mathbf{z} \setminus \mathbf{r})} \quad (131)$$

$$\geq \frac{\mathbb{E}_{P_{d_1}} [Y \mid \mathbf{c}, \mathbf{z}] P_{d_1}(\mathbf{c}, \mathbf{z})}{1 - P_{d_1}(\tilde{\mathbf{c}}, \mathbf{z})} \quad (132)$$

$$= \frac{\mathbb{E}_{P_{d_1}} [Y \mid \mathbf{c}, \mathbf{z}] P_{d_1}(\mathbf{c}, \mathbf{z})}{P_{d_1}(\mathbf{c}, \mathbf{z}) + 1 - P_{d_1}(\mathbf{z})} \quad (133)$$

$$= A_1(\emptyset), \quad (134)$$

where $\tilde{\mathbf{c}}$ stands for the combination of values of $\mathbf{C}$ that are not $\mathbf{c}$. Further,

$$A_2(\mathbf{r}) := \frac{\mathbb{E}_{P_{d_0, \mathbf{r}}} [Y \mid \mathbf{c}, \mathbf{z} \setminus \mathbf{r}] P_{d_0, \mathbf{r}}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}) + 1 - P_{d_0, \mathbf{r}}(\mathbf{z} \setminus \mathbf{r})}{P_{d_0, \mathbf{r}}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}) + 1 - P_{d_0, \mathbf{r}}(\mathbf{z} \setminus \mathbf{r})} \quad (135)$$

$$= 1 - \frac{\mathbb{E}_{P_{d_0, \mathbf{r}}} [1 - Y \mid \mathbf{c}, \mathbf{z} \setminus \mathbf{r}] P_{d_0, \mathbf{r}}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r})}{P_{d_0, \mathbf{r}}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}) + 1 - P_{d_0, \mathbf{r}}(\mathbf{z} \setminus \mathbf{r})} \quad (136)$$

$$\leq 1 - \frac{\mathbb{E}_{P_{d_0}} [1 - Y \mid \mathbf{c}, \mathbf{z}] P_{d_0}(\mathbf{c}, \mathbf{z})}{P_{d_0, \mathbf{r}}(\mathbf{c}, \mathbf{z} \setminus \mathbf{r}) + 1 - P_{d_0, \mathbf{r}}(\mathbf{z} \setminus \mathbf{r})} \quad (137)$$

$$\leq 1 - \frac{\mathbb{E}_{P_{d_0}} [1 - Y \mid \mathbf{c}, \mathbf{z}] P_{d_0}(\mathbf{c}, \mathbf{z})}{P_{d_0}(\mathbf{c}, \mathbf{z}) + 1 - P_{d_0}(\mathbf{z})} \quad (138)$$

$$= \frac{\mathbb{E}_{P_{d_0}} [Y \mid \mathbf{c}, \mathbf{z}] P_{d_0}(\mathbf{c}, \mathbf{z}) + 1 - P_{d_0}(\mathbf{z})}{P_{d_0}(\mathbf{c}, \mathbf{z}) + 1 - P_{d_0}(\mathbf{z})} \quad (139)$$

$$= A_2(\emptyset). \quad (140)$$

□

Thm. 3 restated. Consider an AI grounded in a domain $\mathcal{M}$ made aware of an (under-specified) shift on non-empty $\mathbf{Z} \subset \mathbf{V}$. Then the AI is provably not weakly (or strongly) predictable in any context $\mathbf{C} = \mathbf{c}$.

Proof. Recall that the preference gap is defined as:

$$\Delta_{d_1 > d_0} := \mathbb{E}_{\hat{P}_{\sigma, d_1}} [Y \mid \mathbf{C} = \mathbf{c}] - \mathbb{E}_{\hat{P}_{\sigma, d_0}} [Y \mid \mathbf{C} = \mathbf{c}] \quad (141)$$

Here we know that σ potentially modifies the mechanisms of the set of variables $\mathbf{Z}$ though the nature of the modification is unknown. In the worst-case, the AI's interpretation of the possible new assignment of $\mathbf{Z}$ could be arbitrary.

We will prove this theorem for the case of binary variables $Y, Z \in \mathbf{V}$. In the following, we construct two (canonical) models that entail any chosen distribution for the observed data $P_d(y, z \mid \mathbf{c})$ but evaluate to the a priori minimum and maximum value of the preference gap Δ , i.e. -1 and 1 respectively. We make use of the canonical model construction from (Jalaldoust et al., 2024) to define the following general SCM,

$$Z \leftarrow \begin{cases} 0 & \text{if } r_z = 0 \\ 1 & \text{if } r_z = 1 \end{cases}, \quad Y \leftarrow \begin{cases} 0 & \text{if } r_y = 0 \\ 0 & \text{if } r_y = 1, z = 0 \\ 1 & \text{if } r_y = 1, z = 1 \\ 1 & \text{if } r_y = 2, z = 0 \\ 0 & \text{if } r_y = 2, z = 1 \\ 1 & \text{if } r_y = 3 \end{cases} \quad (142)$$

$U = \{R_z, R_y\}$ where R_z and R_y might be correlated and with a probability $\hat{P}(U) = \hat{P}_d(U \mid \mathbf{c})$ such that $\hat{P}_d(z, y \mid \mathbf{c}) = \hat{P}(z, y \mid \mathbf{c})$. By (Jalaldoust et al., 2024, Thm. 1) this is always possible since this class of canonical models is sufficiently expressive to model any observational or interventional distribution. We can visualise the joint probability of exogenous variables using the following table:

Probabilities $\widehat{\mathcal{M}}$	$r_z = 0$	$r_z = 1$
$r_y = 0$	p_{00}	p_{10}
$r_y = 1$	p_{01}	p_{11}
$r_y = 2$	p_{02}	p_{12}
$r_y = 3$	p_{03}	p_{13}

where we have written $P_d(r_z = a, r_y = b \mid \mathbf{c}) = p_{ab}$. From these we could compute joint probabilities

$$P_d(z = 0, y = 0 \mid \mathbf{c}) = p_{00} + p_{01}, \quad (143)$$

$$P_d(z = 0, y = 1 \mid \mathbf{c}) = p_{02} + p_{03}, \quad (144)$$

$$P_d(z = 1, y = 0 \mid \mathbf{c}) = p_{12} + p_{11}, \quad (145)$$

$$P_d(z = 1, y = 1 \mid \mathbf{c}) = p_{11} + p_{13} \quad (146)$$

Here we can see that the parameter space $P_d(r_z, r_y \mid \mathbf{c})$ is very expressive. For example, without loss of generality we could set $p_{03} = p_{13} = 0$ or $p_{00} = p_{10} = 0$ and still be able to generate any observed distribution $P_d(z, y \mid \mathbf{c})$.

The given shift in the environment σ can be entirely modelled as a shift in $P_{\sigma,d}(r_z \mid \mathbf{c})$ while keeping the probability of r_y invariant, i.e., $P_{\sigma,d}(r_y \mid \mathbf{c}) = P_d(r_y \mid \mathbf{c})$. In other words, given the table above, we can change each of the cells while maintaining the row sums equal. Recall that we are interested in evaluating bounds on a probability of the form $P_{\sigma,d}(y = 1 \mid \mathbf{c})$ and $P_{\sigma,d}(y = 1 \mid z = 1, \mathbf{c})$ depending on whether Z is given as an input to the AI or not. Both these quantities can be written in terms of the probabilities of exogenous variables as follows,

$$P_{\sigma,d}(y = 1 \mid \mathbf{c}) = p_{02} + p_{03} + p_{11} + p_{13} \quad (147)$$

$$P_{\sigma,d}(y = 1 \mid z = 1, \mathbf{c}) = \frac{p_{11} + p_{13}}{p_{11} + p_{13} + p_{12} + p_{11}}. \quad (148)$$

For the lower bound on these quantities, without loss of generality assume that $p_{03} = p_{13} = 0$. Then the following table:

Probabilities $\widehat{\mathcal{M}}_\sigma$	$r_z = 0$	$r_z = 1$
$r_y = 0$	p_{00}	p_{10}
$r_y = 1$	$p_{01} + p_{11}$	0
$r_y = 2$	0	$p_{12} + p_{02}$
$r_y = 3$	0	0

is a perfectly valid model under a shift σ that respects the constraint on $P_{\sigma,d}(r_y \mid \mathbf{c}) = P_d(r_y \mid \mathbf{c})$ but for which $P_{\sigma,d}(y = 1 \mid \mathbf{c}) = 0$ as it is the sum of the 4 zero entries and $P_{\sigma,d}(y = 1 \mid z = 1, \mathbf{c}) = 0$ as it is the sum of the two 0 entries in the second column divided by the sum of entries in the second column.

If we are interested in getting an upper bound then without loss of generality assume that $p_{00} = p_{10} = 0$. Then the following

Probabilities $\widehat{\mathcal{M}}_\sigma$	$r_z = 0$	$r_z = 1$
$r_y = 0$	0	0
$r_y = 1$	0	$p_{01} + p_{11}$
$r_y = 2$	$p_{12} + p_{02}$	0
$r_y = 3$	p_{03}	p_{13}

is a perfectly valid model under a shift σ that respects the constraint on $P_{\sigma,d}(r_y \mid \mathbf{c}) = P_d(r_y \mid \mathbf{c})$ but for which $P_{\sigma,d}(y = 1 \mid \mathbf{c}) = 1$ as it is the sum of the 4 non-zero entries and $P_{\sigma,d}(y = 1 \mid z = 1, \mathbf{c}) = 1$ as it is the sum of the two non-zero entries in the second column divided by the sum of entries in the second column.

By using this construction to define lower and upper bounds for $P_{\sigma,d}(y = 1 \mid \mathbf{c})$ or $P_{\sigma,d}(y = 1 \mid \mathbf{z}, \mathbf{c})$ for $d = d_0, d_1$ we obtain a possible internal model for the AI that entails the observed external behaviour but for which the preference gap evaluates to -1 and 1 . This means that the a priori bound,

$$-1 \leq \Delta_{d>d^*} \leq 1, \quad (149)$$

is tight whenever the shift is undefined (whether we know the variables it applies to or not). Since the preference gap is unconstrained for any $\mathbf{C} = \mathbf{c}$ and any pair of decisions (d, d^*) , the AI is not predictable. $\square$

Thm. 4 restated. Consider an AI grounded in a domain $\mathcal{M}$ and $P_{\sigma,d}(\mathbf{C})$ made aware of a shift σ on $\mathbf{Z} \subset \mathbf{C}$. The AI is weakly predictable under this shift in a context $\mathbf{C} = \mathbf{c}$ if there exists a decision d^* such that,

$$1 - \frac{2 + \mathbb{E}_{P_{d^*}}[Y \mid \mathbf{c}] P_{d^*}(\mathbf{c}) - \mathbb{E}_{P_d}[Y \mid \mathbf{c}] P_d(\mathbf{c}) - 2P_d(\mathbf{z}) + P_d(\mathbf{c})}{P_{\sigma,d^*}(\mathbf{c})} > 0, \quad \text{for some } d \neq d^*. \quad (150)$$

Proof. Recall that the preference gap under a shift σ between decisions (d_1, d_0) in a situation $\mathbf{C} = \mathbf{c}$ is defined as:

$$\Delta_{d_1>d_0} := \mathbb{E}_{\hat{P}_{\sigma,d_1}}[Y \mid \mathbf{C} = \mathbf{c}] - \mathbb{E}_{\hat{P}_{\sigma,d_0}}[Y \mid \mathbf{C} = \mathbf{c}] \quad (151)$$

Here we know that σ potentially modifies the mechanisms of the set of variables $\mathbf{Z}$. The nature of the modification is unknown but we are told that after modification, the expected probability of $\mathbf{C}$ is given by $P_{\sigma,d}(\mathbf{C})$, assumed to be known and internalised by the A. This means that its internal model, whatever interpretation for the shift it chooses, generates the assumed probabilities, i.e. $\hat{P}_{\sigma,d}(\mathbf{C}) = P_{\sigma,d}(\mathbf{C})$.

We will consider the derivation of bounds on each term of this difference separately. Firstly, note that,

$$\mathbb{E}_{\hat{P}_{\sigma,d}}[Y \mid \mathbf{C} = \mathbf{c}] = \mathbb{E}_{\hat{P}_{\sigma,d}}[Y \mathbb{1}_{\mathbf{c}}(\mathbf{C})] / \hat{P}_{\sigma,d}(\mathbf{c}) \quad (152)$$

For ease of notation let us write $\mathbf{R} := \mathbf{C} \setminus \mathbf{Z}$. We could then show that,

$$\mathbb{E}_{\hat{P}_{\sigma,d}}[Y \mathbb{1}_{\mathbf{z},\mathbf{r}}(\mathbf{Z}, \mathbf{R})] = \mathbb{E}_{\hat{P}_{\sigma,d}}[Y_{\mathbf{z}} \mathbb{1}_{\mathbf{z},\mathbf{r}}(\mathbf{Z}, \mathbf{R}_{\mathbf{z}})] \quad \text{by consistency} \quad (153)$$

$$\leq \sum_{\mathbf{z}'} \mathbb{E}_{\hat{P}_{\sigma,d}}[Y_{\mathbf{z}} \mathbb{1}_{\mathbf{z}',\mathbf{r}}(\mathbf{Z}, \mathbf{R}_{\mathbf{z}})] \quad (154)$$

$$= \mathbb{E}_{\hat{P}_{\sigma,d}}[Y_{\mathbf{z}} \mathbb{1}_{\mathbf{r}}(\mathbf{R}_{\mathbf{z}})] \quad \text{marginalizing over the values } \mathbf{z}' \text{ of } \mathbf{Z} \quad (155)$$

Now once we intervene on $\mathbf{z}$ the mechanism that generate its value before hand, whether it was the shift σ or something else is irrelevant. In essence, we get an equivalence between shifted and un-shifted distributions under intervention:

$$\mathbb{E}_{\hat{P}_{\sigma,d}}[Y_{\mathbf{z}} \mathbb{1}_{\mathbf{r}}(\mathbf{R}_{\mathbf{z}})] = \mathbb{E}_{\hat{P}_d}[Y_{\mathbf{z}} \mathbb{1}_{\mathbf{r}}(\mathbf{R}_{\mathbf{z}})] \quad (156)$$

We could now take this quantity to show the following,

$$\mathbb{E}_{\hat{P}_d}[Y_{\mathbf{z}} \mathbb{1}_{\mathbf{r}}(\mathbf{R}_{\mathbf{z}})] = \sum_{\mathbf{z}'} \mathbb{E}_{\hat{P}_d}[Y_{\mathbf{z}} \mathbb{1}_{\mathbf{z}',\mathbf{r}}(\mathbf{Z}, \mathbf{R}_{\mathbf{z}})] \quad (157)$$

$$= \mathbb{E}_{\hat{P}_d}[Y_{\mathbf{z}} \mathbb{1}_{\mathbf{z},\mathbf{r}}(\mathbf{Z}, \mathbf{R}_{\mathbf{z}})] + \sum_{\mathbf{z}' \neq \mathbf{z}} \mathbb{E}_{\hat{P}_d}[Y_{\mathbf{z}} \mathbb{1}_{\mathbf{z}',\mathbf{r}}(\mathbf{Z}, \mathbf{R}_{\mathbf{z}})] \quad (158)$$

$$= \mathbb{E}_{\hat{P}_d}[Y \mathbb{1}_{\mathbf{z},\mathbf{r}}(\mathbf{Z}, \mathbf{R})] + \sum_{\mathbf{z}' \neq \mathbf{z}} \mathbb{E}_{\hat{P}_d}[Y_{\mathbf{z}} \mathbb{1}_{\mathbf{z}',\mathbf{r}}(\mathbf{Z}, \mathbf{R}_{\mathbf{z}})] \quad \text{by consistency} \quad (159)$$

$$\leq \mathbb{E}_{\hat{P}_d}[Y \mathbb{1}_{\mathbf{z},\mathbf{r}}(\mathbf{Z}, \mathbf{R})] + \sum_{\mathbf{z}' \neq \mathbf{z}} \mathbb{E}_{\hat{P}_d}[\mathbb{1}_{\mathbf{z}'}(\mathbf{Z})] \quad \text{since } Y_{\mathbf{z}} \text{ and } \mathbb{1}_{\mathbf{r}}(\mathbf{R}_{\mathbf{z}}) \text{ are } \leq 1 \quad (160)$$

$$= \mathbb{E}_{\hat{P}_d}[Y \mathbb{1}_{\mathbf{z},\mathbf{r}}(\mathbf{Z}, \mathbf{R})] + 1 - \hat{P}_d(\mathbf{z}) \quad (161)$$

$$= \mathbb{E}_{\hat{P}_d}[Y \mid \mathbf{c}] \hat{P}_d(\mathbf{c}) + 1 - \hat{P}_d(\mathbf{z}) \quad (162)$$

For the lower bound we could consider the following derivation,

$$\mathbb{E}_{\hat{P}_{\sigma,d}}[Y \mathbb{1}_{\mathbf{z},\mathbf{r}}(\mathbf{Z}, \mathbf{R})] = \mathbb{E}_{\hat{P}_{\sigma,d}}[\mathbb{1}_{\mathbf{z},\mathbf{r}}(\mathbf{Z}, \mathbf{R})] - \mathbb{E}_{\hat{P}_{\sigma,d}}[(1 - Y) \mathbb{1}_{\mathbf{z},\mathbf{r}}(\mathbf{Z}, \mathbf{R})]. \quad (163)$$

For ease of notation let us define,

$$\mathbb{E}_{\hat{P}_{\sigma,d}}[\tilde{Y} \mathbb{1}_{z,r}(\mathbf{Z}, \mathbf{R})] := \mathbb{E}_{\hat{P}_{\sigma,d}}[(1 - Y) \mathbb{1}_{z,r}(\mathbf{Z}, \mathbf{R})]. \quad (164)$$

Similar bounds apply on $\mathbb{E}_{\hat{P}_{\sigma,d}}[\tilde{Y} \mathbb{1}_{z,r}(\mathbf{Z}, \mathbf{R})]$ to get,

$$\mathbb{E}_{\hat{P}_{\sigma,d}}[Y \mathbb{1}_{z,r}(\mathbf{Z}, \mathbf{R})] \geq \mathbb{E}_{\hat{P}_{\sigma,d}}[\mathbb{1}_{z,r}(\mathbf{Z}, \mathbf{R})] - \{\mathbb{E}_{\hat{P}_d}[\tilde{Y} \mathbb{1}_{z,r}(\mathbf{Z}, \mathbf{R})] + 1 - \hat{P}_d(z)\} \quad (165)$$

$$= \mathbb{E}_{\hat{P}_{\sigma,d}}[\mathbb{1}_{z,r}(\mathbf{Z}, \mathbf{R})] - \mathbb{E}_{\hat{P}_d}[\mathbb{1}_{z,r}(\mathbf{Z}, \mathbf{R})] + \mathbb{E}_{\hat{P}_d}[Y \mathbb{1}_{z,r}(\mathbf{Z}, \mathbf{R})] - 1 + \hat{P}_d(z) \quad (166)$$

$$= \hat{P}_{\sigma,d}(c) - \hat{P}_d(c) + \mathbb{E}_{\hat{P}_d}[Y | c] \hat{P}_d(c) - 1 + \hat{P}_d(z) \quad (167)$$

Putting the lower and upper bounds together to form bounds on $\Delta_{d_1 > d_0}$ we get,

$$\Delta_{d_1 > d_0} \geq \frac{\hat{P}_{\sigma,d_1}(c) - \hat{P}_{d_1}(c) + \mathbb{E}_{\hat{P}_{d_1}}[Y | c] \hat{P}_{d_1}(c) - 1 + \hat{P}_{d_1}(z) - \{\mathbb{E}_{\hat{P}_{d_0}}[Y | c] \hat{P}_{d_0}(c) + 1 - \hat{P}_{d_0}(z)\}}{\hat{P}_{\sigma,d_0}(c)} \quad (168)$$

$$= 1 + \frac{-\hat{P}_{d_1}(c) + \mathbb{E}_{\hat{P}_{d_1}}[Y | c] \hat{P}_{d_1}(c) - 1 + \hat{P}_{d_1}(z) - \mathbb{E}_{\hat{P}_{d_0}}[Y | c] \hat{P}_{d_0}(c) + 1 - \hat{P}_{d_0}(z)}{\hat{P}_{\sigma,d_0}(c)} \quad (169)$$

$$= 1 - \frac{2 + \mathbb{E}_{\hat{P}_{d_0}}[Y | c] \hat{P}_{d_0}(c) - \mathbb{E}_{\hat{P}_{d_1}}[Y | c] \hat{P}_{d_1}(c) - 2\hat{P}_{d_1}(z) + \hat{P}_{d_1}(c)}{\hat{P}_{\sigma,d_0}(c)} \quad (170)$$

and by grounding,

$$\Delta_{d_1 > d_0} \geq 1 - \frac{2 + \mathbb{E}_{P_{d_0}}[Y | c] P_{d_0}(c) - \mathbb{E}_{P_{d_1}}[Y | c] P_{d_1}(c) - 2P_{d_1}(z) + P_{d_1}(c)}{P_{\sigma,d_0}(c)}. \quad (171)$$

This statement holds for any SCM compatible with the grounded AI's external behaviour and therefore,

$$\min_{\mathcal{M} \in \mathbb{M}} (\Delta_{d > d^*}) \geq 1 - \frac{2 + \mathbb{E}_{P_{d^*}}[Y | c] P_{d^*}(c) - \mathbb{E}_{P_d}[Y | c] P_d(c) - 2P_d(z) + P_d(c)}{P_{\sigma,d^*}(c)}. \quad (172)$$

We can establish that the AI is weakly predictable in a context $C = c$ if there exists a decision d^* such that,

$$1 - \frac{2 + \mathbb{E}_{P_{d^*}}[Y | c] P_{d^*}(c) - \mathbb{E}_{P_d}[Y | c] P_d(c) - 2P_d(z) + P_d(c)}{P_{\sigma,d^*}(c)} > 0, \quad (173)$$

for some $d \neq d^*$. □

We now continue with our inference of the AI's perceived fairness and harm of decisions in Sec. 4.3.

Thm. 5 restated. Consider an agent with utility function Y grounded in a domain $\mathcal{M}$. Then,

$$-\mathbb{E}_{P_d}[Y | z, c] \leq \Upsilon(d, c) \leq 1 - \mathbb{E}_{P_d}[Y | z, c]. \quad (174)$$

This bound is tight.

Proof. Recall that for a given utility Y , the AI's counterfactual fairness gap relative to a decision d , in a given context c , is

$$\Upsilon(d, c) := \mathbb{E}_{\hat{P}}[Y_{d,z_1} | z_0, c] - \mathbb{E}_{\hat{P}}[Y_d | z_0, c]. \quad (175)$$

And remember that $Z \in C$.

For ease of notation, write $z_1 = z, z_0 = z'$ such that,

$$\Upsilon(d, c) := \mathbb{E}_{\hat{P}}[Y_{d,z} | z', c] - \mathbb{E}_{\hat{P}}[Y_d | z', c]. \quad (176)$$

We start by considering the following derivation:

$$\hat{P}(y_{d,z} | \mathbf{c}) = \hat{P}(y_{d,z}, z_d | \mathbf{c}) + \hat{P}(y_{d,z}, z'_d | \mathbf{c}) \quad \text{by marginalization} \quad (177)$$

$$= \hat{P}(y_d, z_d | \mathbf{c}) + \hat{P}(y_{d,z}, z'_d | \mathbf{c}) \quad \text{by consistency} \quad (178)$$

and since d does not affect Z or $\mathbf{C}$, i.e. $Z_d = Z, \mathbf{C}_d = \mathbf{C}$,

$$\hat{P}(y_{d,z} | \mathbf{c}) = \hat{P}(y_d, z_d | \mathbf{c}) + \hat{P}(y_{d,z}, z' | \mathbf{c}) \quad (179)$$

which implies

$$\hat{P}(y_{d,z} | z', \mathbf{c}) = \frac{\hat{P}(y_{d,z} | \mathbf{c}) - \hat{P}_d(y, z | \mathbf{c})}{\hat{P}_d(z' | \mathbf{c})} \quad (180)$$

Therefore,

$$\mathbb{E}_{\hat{P}}[Y_{d,z} | z', \mathbf{c}] = \frac{\mathbb{E}_{\hat{P}}[Y_{d,z} | \mathbf{c}] - \mathbb{E}_{\hat{P}_d}[Y | z, \mathbf{c}]\hat{P}_d(z | \mathbf{c})}{\hat{P}_d(z' | \mathbf{c})}. \quad (181)$$

All quantities on the r.h.s are observable except for $\mathbb{E}_{\hat{P}}[Y_{d,z} | \mathbf{c}]$ which can be tightly bounded.

For the lower bound, consider the following derivation,

$$\mathbb{E}_{\hat{P}}[Y_{d,z} | \mathbf{c}] = \sum_{\tilde{z}} \mathbb{E}_{\hat{P}}[Y_{d,z} \mathbb{1}_{\tilde{z}_d}(Z) | \mathbf{c}] \quad \text{marginalizing over } z_d \quad (182)$$

$$\geq \mathbb{E}_{\hat{P}}[Y_{d,z} \mathbb{1}_{z_d}(Z_d) | \mathbf{c}] \quad \text{since summands } > 0 \quad (183)$$

$$= \mathbb{E}_{\hat{P}}[Y_d \mathbb{1}_{z_d}(Z_d) | \mathbf{c}] \quad \text{by consistency} \quad (184)$$

$$= \mathbb{E}_{P_d}[Y | \mathbf{c}, z] P_d(z | \mathbf{c}) \quad \text{by grounding and } \mathbf{C}_d = \mathbf{C} \quad (185)$$

Similarly, we can get an upper bound by noting

$$\mathbb{E}_{\hat{P}}[Y_{d,z} | \mathbf{c}] = 1 - \mathbb{E}_{\hat{P}}[(1 - Y_{d,z}) | \mathbf{c}] \quad (186)$$

$$\leq \mathbb{E}_{\hat{P}_d}[Y | \mathbf{c}, z] \hat{P}_d(z | \mathbf{c}) + \hat{P}_d(z' | \mathbf{c}). \quad (187)$$

Tightness Lower Bound For the lower bound we will consider the following SCM,

$$\mathcal{M}_d^1 =: \begin{cases} Z \leftarrow f_Z(\mathbf{u}) \\ \mathbf{C} \leftarrow f_{\mathbf{C}}(\mathbf{u}) \\ D \leftarrow d \\ Y \leftarrow \begin{cases} f_Y(d, \mathbf{c}, z, \mathbf{u}) & \text{if } f_Z(\mathbf{u}) = z \\ 0 & \text{otherwise} \end{cases} \\ P(\mathbf{U}) \end{cases} \quad (188)$$

Here $\{f_Z, f_{\mathbf{C}}, f_Y, \mathcal{U}, P(\mathbf{U})\}$ are chosen to match the observed trajectory of agent interactions, i.e., such that $P^{\mathcal{M}_d^1}(\mathbf{v}) = P^{\hat{\mathcal{M}}_d}(\mathbf{v})$ for all $\mathbf{v} \in \text{supp}_{\mathbf{V}}$.

Then, under $\mathcal{M}_d^1$,

$$\mathbb{E}_{P^{\mathcal{M}_d^1}}[Y_{d,z} | \mathbf{c}] \quad (189)$$

$$= \sum_{\mathbf{u}} \mathbb{E}_{P^{\mathcal{M}_d^1}}[Y_{d,z} | \mathbf{u}, \mathbf{c}] P^{\mathcal{M}_d^1}(\mathbf{u} | \mathbf{c}) \quad (190)$$

$$= \sum_{\mathbf{u}} \mathbb{E}_{P^{\mathcal{M}_d^1}}[Y_d | z, \mathbf{u}, \mathbf{c}] P^{\mathcal{M}_d^1}(\mathbf{u} | \mathbf{c}) \quad (191)$$

$$= \mathbb{E}_{P^{\mathcal{M}_d^1}}[Y | z, \mathbf{c}, \{\mathbf{u} : f_Z(\mathbf{u}) = z\}] P^{\mathcal{M}_d^1}(\{\mathbf{u} : f_Z(\mathbf{u}) = z\} | \mathbf{c}) \quad (192)$$

$$+ \mathbb{E}_{P^{\mathcal{M}_d^1}}[Y | z, \mathbf{c}, \{\mathbf{u} : f_Z(\mathbf{u}) \neq z\}] P^{\mathcal{M}_d^1}(\{\mathbf{u} : f_Z(\mathbf{u}) \neq z\} | \mathbf{c}) \quad (193)$$

$$= \mathbb{E}_{P^{\mathcal{M}_d^1}}[Y | z, \mathbf{c}] P^{\mathcal{M}_d^1}(z | \mathbf{c}). \quad (194)$$

This expression is the same one as the analytical bound showing that it is tight.

Tightness Upper Bound For the upper bound we will consider the following SCM,

$$\mathcal{M}_d^2 =: \begin{cases} Z \leftarrow f_Z(\mathbf{u}) \\ C \leftarrow f_C(\mathbf{u}) \\ D \leftarrow d \\ Y \leftarrow \begin{cases} f_Y(d, \mathbf{c}, z, \mathbf{u}) & \text{if } f_Z(\mathbf{u}) = z \\ 1 & \text{otherwise} \end{cases} \\ P(U) \end{cases} \quad (195)$$

Here $\{f_Z, f_C, f_Y, \mathcal{U}, P(U)\}$ are chosen to match the observed trajectory of agent interactions, i.e., such that $P^{\mathcal{M}_d^2}(\mathbf{v}) = P^{\widehat{\mathcal{M}}_d}(\mathbf{v})$ for all $\mathbf{v} \in \text{supp}_V$.

Then, under $\mathcal{M}_d^2$,

$$\mathbb{E}_{P^{\mathcal{M}_d^2}}[Y_{d,z} \mid \mathbf{c}] \quad (196)$$

$$= \sum_{\mathbf{u}} \mathbb{E}_{P^{\mathcal{M}_d^2}}[Y_{d,z} \mid \mathbf{u}, \mathbf{c}] P^{\mathcal{M}_d^2}(\mathbf{u} \mid \mathbf{c}) \quad (197)$$

$$= \sum_{\mathbf{u}} \mathbb{E}_{P^{\mathcal{M}_d^2}}[Y_d \mid z, \mathbf{u}, \mathbf{c}] P^{\mathcal{M}_d^2}(\mathbf{u} \mid \mathbf{c}) \quad (198)$$

$$= \mathbb{E}_{P^{\mathcal{M}_d^2}}[Y \mid z, \mathbf{c}, \{\mathbf{u} : f_Z(\mathbf{u}) = z\}] P^{\mathcal{M}_d^2}(\{\mathbf{u} : f_Z(\mathbf{u}) = z\} \mid \mathbf{c}) \quad (199)$$

$$+ \mathbb{E}_{P^{\mathcal{M}_d^2}}[Y \mid z, \mathbf{c}, \{\mathbf{u} : f_Z(\mathbf{u}) \neq z\}] P^{\mathcal{M}_d^2}(\{\mathbf{u} : f_Z(\mathbf{u}) \neq z\} \mid \mathbf{c}) \quad (200)$$

$$= \mathbb{E}_{P^{\mathcal{M}_d^2}}[Y \mid z, \mathbf{c}] P^{\mathcal{M}_d^2}(z \mid \mathbf{c}) + 1 - P^{\mathcal{M}_d^2}(z \mid \mathbf{c}). \quad (201)$$

We therefore find that,

$$0 \leq \mathbb{E}_{\widehat{P}}[Y_{d,z} \mid z', \mathbf{c}] \leq 1, \quad (202)$$

and ultimately,

$$-\mathbb{E}_{P_d}[Y \mid z, \mathbf{c}] \leq \Upsilon(d, \mathbf{c}) \leq 1 - \mathbb{E}_{P_d}[Y \mid z, \mathbf{c}], \quad (203)$$

as claimed. $\square$

Thm. 6 restated. Consider an agent with utility function Y grounded in a domain $\mathcal{M}$. Then,

$$\max\{0, \mathbb{E}_{P_d}[Y \mid \mathbf{c}] + \mathbb{E}_{P_{d_0}}[Y \mid \mathbf{c}] - 1\} \leq \Omega(d, d_0) \leq \min\{\mathbb{E}_{P_d}[Y \mid \mathbf{c}], \mathbb{E}_{P_{d_0}}[Y \mid \mathbf{c}]\} \quad (204)$$

and this bound is tight.

Proof. Consider an agent with internal model $\widehat{\mathcal{M}}$ and utility function Y . Recall that the agent's expected harm of a decision d with respect to a baseline d_0 , in context $\mathbf{c}$, is

$$\Omega(d, d_0) := \mathbb{E}_{\widehat{P}}[\max\{0, Y_{d_0} - Y_d\} \mid \mathbf{c}]. \quad (205)$$

We can re-write this quantity as follows

$$\Omega(d, d_0) = \mathbb{E}_{\widehat{P}}[\max\{0, Y_{d_0} - Y_d\} \mid \mathbf{c}] \quad (206)$$

$$= \int \max\{0, y_{d_0} - y_d\} \widehat{P}(y_d, y_{d_0} \mid \mathbf{c}) dy_d dy_{d_0} \quad (207)$$

Since Y_d is binary, the only time that the maximum evaluates to something greater than zero is when $Y_{d_0} = 1$ and $Y_d = 0$. Then,

$$\Omega(d, d_0) = \widehat{P}(Y_{d_0} = 1, Y_d = 0) \quad (208)$$

This quantity can be tightly bounded using the results of (Tian & Pearl, 2000, Sec. 4.2.2) giving

$$\max\{0, \mathbb{E}_{\hat{P}_d} [Y | \mathbf{c}] + \mathbb{E}_{\hat{P}_{d_0}} [Y | \mathbf{c}] - 1\} \leq \Omega(d, d_0) \leq \min\{\mathbb{E}_{\hat{P}_d} [Y | \mathbf{c}], \mathbb{E}_{\hat{P}_{d_0}} [Y | \mathbf{c}]\}. \quad (209)$$

And by grounding,

$$\max\{0, \mathbb{E}_{P_d} [Y | \mathbf{c}] + \mathbb{E}_{P_{d_0}} [Y | \mathbf{c}] - 1\} \leq \Omega(d, d_0) \leq \min\{\mathbb{E}_{P_d} [Y | \mathbf{c}], \mathbb{E}_{P_{d_0}} [Y | \mathbf{c}]\}. \quad (210)$$

□

D. Other accounts of fairness and harm

To ground definitions of fairness, several authors appeal to counterfactual thinking but some accounts, instead, are interventional in nature.

Within legal systems, counterfactual fairness (Def. 7) operationalizes a doctrine known as disparate impact doctrine focuses on outcome fairness, namely, the equality of outcomes among protected groups. On the other hand, the doctrine of disparate treatment seeks to enforce the equality of treatment in different groups, prohibiting the use of a protected attribute in the decision process, and has been formalized using interventional accounts (Barocas & Selbst, 2016).

A popular notion in the disparate treatment literature is known as direct discrimination (Barocas & Selbst, 2016; Zhang & Bareinboim, 2018). An agent is said to engage in direct discrimination if the causal influence of a sensitive attribute Z that is not mediated by other variables C is non-zero. This is a contrast between interventional expectations. We adapt this notion to define an AI’s perceived direct fairness gap as the difference in expected utilities obtained for different values of a protected attribute while holding all other variables fixed.

Definition 10 (Direct Discrimination Gap). *Let $Z \in \{z_0, z_1\}$ be a protected attribute. For a given utility Y , define an agent’s direct discrimination gap relative to a baseline value z_0 in a given context c as*

$$\Psi(d, c) := \mathbb{E}_{\hat{P}}[Y_{d,z_1,c}] - \mathbb{E}_{\hat{P}}[Y_{d,z_0,c}]. \quad (211)$$

We say that an AI “intends” to avoid direct discrimination if under any context $C = c$ and decision $D = d$ the direct discrimination gap Ψ evaluates to 0. Here, we consider this notion of fairness to illustrate the kind of inference that is possible to obtain from an AI’s external behaviour with one alternative account. The following theorem shows that, contrary to the counterfactual fairness gap, Ψ can be bounded given the AI’s external behaviour.

Theorem 9. *Consider an agent with utility Y grounded in a domain $\mathcal{M}$. Then,*

$$\Psi(d, c) \geq \mathbb{E}_{P_d}[Y | z_1, c] P_d(z_1, c) - \mathbb{E}_{P_d}[Y | z_0, c] P_d(z_0, c) + P_d(z_0, c) - 1, \quad (212)$$

$$\Psi(d, c) \leq \mathbb{E}_{P_d}[Y | z_1, c] P_d(z_1, c) - \mathbb{E}_{P_d}[Y | z_0, c] P_d(z_0, c) + 1 - P_d(z_1, c). \quad (213)$$

This bound is tight.

Proof. Let $Z \in \{0, 1\}$ be a protected attribute and z_0 a baseline value of Z . For a given utility variable Y , recall that the AI’s direct fairness gap relative to a baseline z_0 in a given context c is defined as

$$\Psi(d, c) := \mathbb{E}_{\hat{P}}[Y_{d,z_1,c}] - \mathbb{E}_{\hat{P}}[Y_{d,z_0,c}]. \quad (214)$$

Using a similar proof strategy to that in Thm. 1, we can derive tight bounds on Ψ .

Analytical Lower Bound A lower bound on the interventional expectation can be obtained using the following derivation:

$$\mathbb{E}_{\hat{P}}[Y_{z,c,d}] = \sum_{\tilde{c}, \tilde{z}} \mathbb{E}_{\hat{P}}[Y_{z,c,d} \mathbb{1}_{\tilde{c}, \tilde{z}}(C_d, Z_{c,d})] \quad \text{marginalizing over } c_d, z_{c,d} \quad (215)$$

$$\geq \mathbb{E}_{\hat{P}}[Y_{z,c,d} \mathbb{1}_{c,z}(C_d, Z_{c,d})] \quad \text{since summands } > 0 \quad (216)$$

$$= \mathbb{E}_{\hat{P}}[Y_{c,d} \mathbb{1}_{c,z}(C_d, Z_{c,d})] \quad \text{by consistency} \quad (217)$$

$$= \mathbb{E}_{\hat{P}}[Y_d \mathbb{1}_{c,z}(C_d, Z_d)] \quad \text{by consistency} \quad (218)$$

$$= \mathbb{E}_{P_d}[Y \mathbb{1}_{c,z}(C, Z)] \quad \text{by grounding} \quad (219)$$

$$= \mathbb{E}_{P_d}[Y | c, z] P_d(c, z). \quad (220)$$

Analytical Upper Bound For deriving an upper bound on the interventional expectation, we start by noting that,

$$\mathbb{E}_{\hat{P}}[Y_{z,c,d}] = 1 - \mathbb{E}_{\hat{P}}[1 - Y_{z,c,d}] \quad (221)$$

Leveraging the bounds derived above we obtain,

$$\mathbb{E}_{\hat{P}}[Y_{z,c,d}] \leq 1 - \mathbb{E}_{P_d}[(1 - Y) | c, z] P_d(c, z) \quad (222)$$

$$= \mathbb{E}_{P_d}[Y | c, z] P_d(c, z) + 1 - P_d(c, z). \quad (223)$$

By setting $z = z_1$ in the lower bound and $z = z_0$ in the upper bound of the expected utility, we obtain a lower bound on the difference of expected utilities:

$$\Psi(d, \mathbf{c}) \geq \mathbb{E}_{P_d} [Y \mid z_1, \mathbf{c}] P_d(z_1, \mathbf{c}) - \mathbb{E}_{P_d} [Y \mid z_0, \mathbf{c}] P_d(z_0, \mathbf{c}) + P_d(z_0, \mathbf{c}) - 1. \quad (224)$$

And similarly, by setting $z = z_1$ in the upper bound and $z = z_0$ in the lower bound of the expected utility, we obtain an upper bound on the difference of expected utilities:

$$\Psi(d, \mathbf{c}) \leq \mathbb{E}_{P_d} [Y \mid z_1, \mathbf{c}] P_d(z_1, \mathbf{c}) - \mathbb{E}_{P_d} [Y \mid z_0, \mathbf{c}] P_d(z_0, \mathbf{c}) + 1 - P_d(z_1, \mathbf{c}). \quad (225)$$

We now show that these bounds are tight by constructing SCMs (that is, possible world models of the AI system) that evaluate to the lower and upper bounds while generating the distribution of agent interactions P_d .

Tightness Lower Bound For the lower bound we will consider the following SCM,

$$\mathcal{M}_d^1 =: \begin{cases} Z \leftarrow f_Z(\mathbf{u}) \\ \mathbf{C} \leftarrow f_{\mathbf{C}}(\mathbf{u}) \\ D \leftarrow d \\ Y \leftarrow \begin{cases} f_Y(d, \mathbf{c}, z_1, \mathbf{u}) & \text{if } f_Z(\mathbf{u}) = z_1, f_{\mathbf{C}}(\mathbf{u}) = \mathbf{c} \\ 0 & \text{if } f_Z(\mathbf{u}) \neq z_1 \text{ or } f_{\mathbf{C}}(\mathbf{u}) \neq \mathbf{c}, \text{ and } Z = z_1 \\ f_Y(d, \mathbf{c}, z_0, \mathbf{u}) & \text{if } f_Z(\mathbf{u}) = z_0, f_{\mathbf{C}}(\mathbf{u}) = \mathbf{c} \\ 1 & \text{if } f_Z(\mathbf{u}) \neq z_0 \text{ or } f_{\mathbf{C}}(\mathbf{u}) \neq \mathbf{c}, \text{ and } Z = z_0 \end{cases} \\ P(\mathbf{U}) \end{cases} \quad (226)$$

Here $\{f_Z, f_{\mathbf{C}}, f_Y, \mathbf{U}, P(\mathbf{U})\}$ are chosen to match the observed trajectory of agent interactions, i.e., such that $P^{\mathcal{M}_d^1}(\mathbf{v}) = P^{\widehat{\mathcal{M}}_d}(\mathbf{v})$ for all $\mathbf{v} \in \text{supp}_{\mathbf{V}}$.

Then, under $\mathcal{M}_d^1$,

$$\Psi(d, \mathbf{c}) = \mathbb{E}_{P^{\mathcal{M}_d^1}} [Y_{d, z_1, \mathbf{c}}] - \mathbb{E}_{P^{\mathcal{M}_d^1}} [Y_{d, z_0, \mathbf{c}}] \quad (227)$$

$$= \sum_{\mathbf{u}} \mathbb{E}_{P^{\mathcal{M}_d^1}} [Y_{d, z_1, \mathbf{c}} \mid \mathbf{u}] P^{\mathcal{M}_d^1}(\mathbf{u}) \quad (228)$$

$$- \sum_{\mathbf{u}} \mathbb{E}_{P^{\mathcal{M}_d^1}} [Y_{d, z_0, \mathbf{c}} \mid \mathbf{u}] P^{\mathcal{M}_d^1}(\mathbf{u}) \quad (229)$$

$$= \sum_{\mathbf{u}} \mathbb{E}_{P^{\mathcal{M}_d^1}} [Y_d \mid z, \mathbf{u}, \mathbf{c}] P^{\mathcal{M}_d^1}(\mathbf{u}) \quad (230)$$

$$- \sum_{\mathbf{u}} \mathbb{E}_{P^{\mathcal{M}_d^1}} [Y_d \mid z, \mathbf{u}, \mathbf{c}] P^{\mathcal{M}_d^1}(\mathbf{u}) \quad (231)$$

$$= \mathbb{E}_{P^{\mathcal{M}_d^1}} [Y \mid z_1, \mathbf{c}, \{\mathbf{u} : f_Z(\mathbf{u}) = z_1, f_{\mathbf{C}}(\mathbf{u}) = \mathbf{c}\}] P^{\mathcal{M}_d^1}(\{\mathbf{u} : f_Z(\mathbf{u}) = z_1, f_{\mathbf{C}}(\mathbf{u}) = \mathbf{c}\}) \quad (232)$$

$$+ \mathbb{E}_{P^{\mathcal{M}_d^1}} [Y \mid z_1, \mathbf{c}, \{\mathbf{u} : f_Z(\mathbf{u}) \neq z_1 \text{ or } f_{\mathbf{C}}(\mathbf{u}) \neq \mathbf{c}\}] P^{\mathcal{M}_d^1}(\{\mathbf{u} : f_Z(\mathbf{u}) \neq z_1 \text{ or } f_{\mathbf{C}}(\mathbf{u}) \neq \mathbf{c}\}) \quad (233)$$

$$- \mathbb{E}_{P^{\mathcal{M}_d^1}} [Y \mid z_0, \mathbf{c}, \{\mathbf{u} : f_Z(\mathbf{u}) = z_0, f_{\mathbf{C}}(\mathbf{u}) = \mathbf{c}\}] P^{\mathcal{M}_d^1}(\{\mathbf{u} : f_Z(\mathbf{u}) = z_0, f_{\mathbf{C}}(\mathbf{u}) = \mathbf{c}\}) \quad (234)$$

$$- \mathbb{E}_{P^{\mathcal{M}_d^1}} [Y \mid z_0, \mathbf{c}, \{\mathbf{u} : f_Z(\mathbf{u}) \neq z_0 \text{ or } f_{\mathbf{C}}(\mathbf{u}) \neq \mathbf{c}\}] P^{\mathcal{M}_d^1}(\{\mathbf{u} : f_Z(\mathbf{u}) \neq z_0 \text{ or } f_{\mathbf{C}}(\mathbf{u}) \neq \mathbf{c}\}) \quad (235)$$

$$= \mathbb{E}_{P^{\mathcal{M}_d^1}} [Y \mid z_1, \mathbf{c}] P^{\mathcal{M}_d^1}(z_1, \mathbf{c}) - \mathbb{E}_{P^{\mathcal{M}_d^1}} [Y \mid z_0, \mathbf{c}] P^{\mathcal{M}_d^1}(z_0, \mathbf{c}) - 1 + P^{\mathcal{M}_d^1}(z_0, \mathbf{c}). \quad (236)$$

This expression is the same one as the analytical bound showing that it is tight.

Tightness Upper Bound For the upper bound we will consider the following SCM,

$$\mathcal{M}_d^2 =: \begin{cases} Z \leftarrow f_Z(\mathbf{u}) \\ C \leftarrow f_C(\mathbf{u}) \\ D \leftarrow d \\ Y \leftarrow \begin{cases} f_Y(d, \mathbf{c}, z_1, \mathbf{u}) & \text{if } f_Z(\mathbf{u}) = z_1, f_C(\mathbf{u}) = \mathbf{c} \\ 1 & \text{if } f_Z(\mathbf{u}) \neq z_1 \text{ or } f_C(\mathbf{u}) \neq \mathbf{c}, \text{ and } Z = z_1 \\ f_Y(d, \mathbf{c}, z_0, \mathbf{u}) & \text{if } f_Z(\mathbf{u}) = z_0, f_C(\mathbf{u}) = \mathbf{c} \\ 0 & \text{if } f_Z(\mathbf{u}) \neq z_0 \text{ or } f_C(\mathbf{u}) \neq \mathbf{c}, \text{ and } Z = z_0 \end{cases} \\ P(\mathbf{U}) \end{cases} \quad (237)$$

Here $\{f_Z, f_C, f_Y, \mathcal{U}, P(\mathbf{U})\}$ are chosen to match the observed trajectory of agent interactions, i.e., such that $P^{\mathcal{M}_d^2}(\mathbf{v}) = P^{\widehat{\mathcal{M}}_d}(\mathbf{v})$ for all $\mathbf{v} \in \text{supp}_{\mathbf{V}}$.

Then, under $\mathcal{M}_d^2$,

$$\Psi(d, \mathbf{c}) = \mathbb{E}_{P^{\mathcal{M}_d^2}}[Y_{d, z_1, \mathbf{c}}] - \mathbb{E}_{P^{\mathcal{M}_d^2}}[Y_{d, z_0, \mathbf{c}}] \quad (238)$$

$$= \sum_{\mathbf{u}} \mathbb{E}_{P^{\mathcal{M}_d^2}}[Y_{d, z_1, \mathbf{c}} | \mathbf{u}] P^{\mathcal{M}_d^2}(\mathbf{u}) \quad (239)$$

$$- \sum_{\mathbf{u}} \mathbb{E}_{P^{\mathcal{M}_d^2}}[Y_{d, z_0, \mathbf{c}} | \mathbf{u}] P^{\mathcal{M}_d^2}(\mathbf{u}) \quad (240)$$

$$= \sum_{\mathbf{u}} \mathbb{E}_{P^{\mathcal{M}_d^2}}[Y_d | z, \mathbf{u}, \mathbf{c}] P^{\mathcal{M}_d^2}(\mathbf{u}) \quad (241)$$

$$- \sum_{\mathbf{u}} \mathbb{E}_{P^{\mathcal{M}_d^2}}[Y_d | z, \mathbf{u}, \mathbf{c}] P^{\mathcal{M}_d^2}(\mathbf{u}) \quad (242)$$

$$= \mathbb{E}_{P^{\mathcal{M}_d^2}}[Y | z_1, \mathbf{c}, \{\mathbf{u} : f_Z(\mathbf{u}) = z_1, f_C(\mathbf{u}) = \mathbf{c}\}] P^{\mathcal{M}_d^2}(\{\mathbf{u} : f_Z(\mathbf{u}) = z_1, f_C(\mathbf{u}) = \mathbf{c}\}) \quad (243)$$

$$+ \mathbb{E}_{P^{\mathcal{M}_d^2}}[Y | z_1, \mathbf{c}, \{\mathbf{u} : f_Z(\mathbf{u}) \neq z_1 \text{ or } f_C(\mathbf{u}) \neq \mathbf{c}\}] P^{\mathcal{M}_d^2}(\{\mathbf{u} : f_Z(\mathbf{u}) \neq z_1 \text{ or } f_C(\mathbf{u}) \neq \mathbf{c}\}) \quad (244)$$

$$- \mathbb{E}_{P^{\mathcal{M}_d^2}}[Y | z_0, \mathbf{c}, \{\mathbf{u} : f_Z(\mathbf{u}) = z_0, f_C(\mathbf{u}) = \mathbf{c}\}] P^{\mathcal{M}_d^2}(\{\mathbf{u} : f_Z(\mathbf{u}) = z_0, f_C(\mathbf{u}) = \mathbf{c}\}) \quad (245)$$

$$- \mathbb{E}_{P^{\mathcal{M}_d^2}}[Y | z_0, \mathbf{c}, \{\mathbf{u} : f_Z(\mathbf{u}) \neq z_0 \text{ or } f_C(\mathbf{u}) \neq \mathbf{c}\}] P^{\mathcal{M}_d^2}(\{\mathbf{u} : f_Z(\mathbf{u}) \neq z_0 \text{ or } f_C(\mathbf{u}) \neq \mathbf{c}\}) \quad (246)$$

$$= \mathbb{E}_{P^{\mathcal{M}_d^2}}[Y | z_1, \mathbf{c}] P^{\mathcal{M}_d^2}(z_1, \mathbf{c}) + 1 - P^{\mathcal{M}_d^2}(z_1, \mathbf{c}) - \mathbb{E}_{P^{\mathcal{M}_d^2}}[Y | z_0, \mathbf{c}] P^{\mathcal{M}_d^2}(z_0, \mathbf{c}). \quad (247)$$

This expression is the same one as the analytical bound showing that it is tight. $\square$

Definitions of harm (defined with respect to a causal model) can also be split in two groups: causal and counterfactual accounts. (Beckers et al., 2022) exemplify the causal account as defining a *decision d to harm a person if and only if d is a cause of harm*. Recall that the counterfactual account has the same structure but differs in the second clause, instead defining a *decision d to harm a person if and only if she would have been better off if d had not been taken*. Here, we quantify how “good” or “beneficial” a particular situation $\mathbf{V} = \mathbf{v}$ is with a binary utility $Y \in \{y_0, y_1\}$ that we assume is tracked in experiments (it might capture, for example, the value of sensitive environmental variables). A formalisation of this causal account of harm, with respect to an AI’s internal model, is given in the following definition.

Definition 11 (Causal Harm Gap). *Consider an agent with internal model $\widehat{\mathcal{M}}$ and utility $Y \in \{y_0, y_1\}$. The agent’s expected causal harm of a decision d with respect to a baseline d_0 that obtained the non-harmful outcome y_0 in context $\mathbf{c}$, is*

$$\Omega(d_1, d_0, \mathbf{c}) := \mathbb{E}_{\widehat{P}}[Y_{d_1} | y_0, d_0, \mathbf{c}]. \quad (248)$$

This probability expresses the capacity of d_1 to produce a harmful event $Y = y_1$ that implies a transition from the absence to the presence of d_1 and y_1 , we condition the probability on situations where d_1 and y_1 are absent, i.e. $D = d_0, Y = y_0$.

Theorem 10. *Consider an agent with utility Y grounded in a domain $\mathcal{M}$. Then,*

$$\frac{P_{d_1}(y_1 | \mathbf{c}) - P(y_1, d_1 | \mathbf{c})}{P_{d_0}(y_0 | \mathbf{c})P(d_0 | \mathbf{c})} \leq \Omega(d_1, d_0, \mathbf{c}) \leq \frac{P_{d_1}(y_1 | \mathbf{c}) - P_{d_1}(y_1 | \mathbf{c})P(d_1 | \mathbf{c})}{P_{d_0}(y_0 | \mathbf{c})P(d_0 | \mathbf{c})}. \quad (249)$$

Proof. Note that the causal harm gap may be equivalently written,

$$\Omega(d_1, d_0, \mathbf{c}) := \hat{P}(y_{1,d_1} \mid y_0, d_0, \mathbf{c}). \quad (250)$$

The lower and upper bounds may be derived considering the following,

$$\hat{P}(y_{1,d_1} \mid \mathbf{c}) = \hat{P}(y_{1,d_1}, y_0, d_0 \mid \mathbf{c}) + \hat{P}(y_{1,d_1}, y_1, d_0 \mid \mathbf{c}) + \hat{P}(y_{1,d_1}, y_0, d_1 \mid \mathbf{c}) + \hat{P}(y_{1,d_1}, y_1, d_1 \mid \mathbf{c}) \quad (251)$$

$$= \hat{P}(y_{1,d_1}, y_0, d_0 \mid \mathbf{c}) + \hat{P}(y_{1,d_1}, y_1, d_0 \mid \mathbf{c}) + \hat{P}(y_{1,d_1}, y_1, d_1 \mid \mathbf{c}) \quad (252)$$

$$= \hat{P}(y_{1,d_1}, y_0, d_0 \mid \mathbf{c}) + \hat{P}(y_{1,d_1}, y_1 \mid \mathbf{c}) \quad (253)$$

$$\leq \hat{P}(y_{1,d_1}, y_0, d_0 \mid \mathbf{c}) + \hat{P}(y_{1,d_1} \mid \mathbf{c}) \quad (254)$$

$$\hat{P}(y_{1,d_1} \mid \mathbf{c}) = \hat{P}(y_{1,d_1}, y_0, d_0 \mid \mathbf{c}) + \hat{P}(y_{1,d_1}, y_1 \mid \mathbf{c}) \quad (255)$$

$$\geq \hat{P}(y_{1,d_1}, y_0, d_0 \mid \mathbf{c}) + \hat{P}(y_{1,d_1}, y_1, d_1 \mid \mathbf{c}) \quad (256)$$

$$= \hat{P}(y_{1,d_1}, y_0, d_0 \mid \mathbf{c}) + \hat{P}(y_{1,d_1}, d_1 \mid \mathbf{c}) \quad \text{by consistency} \quad (257)$$

$$= \hat{P}(y_{1,d_1}, y_0, d_0 \mid \mathbf{c}) + \hat{P}(y_{1,d_1}, d_1 \mid \mathbf{c}) \hat{P}(d_1 \mid \mathbf{c}). \quad (258)$$

$\hat{P}(d_1 \mid \mathbf{c})$ stands for the AI's policy in the source environment, i.e., the probability it uses for choosing decision d_1 in situation $\mathbf{c}$. Re-arranging these equations this implies,

$$\frac{\hat{P}(y_{1,d_1} \mid \mathbf{c}) - \hat{P}(y_{1,d_1} \mid \mathbf{c})}{\hat{P}(y_{0,d_0} \mid \mathbf{c}) \hat{P}(d_0 \mid \mathbf{c})} \leq \Omega(d_1, d_0, \mathbf{c}) \leq \frac{\hat{P}(y_{1,d_1} \mid \mathbf{c}) - \hat{P}(y_{1,d_1} \mid \mathbf{c}) \hat{P}(d_1 \mid \mathbf{c})}{\hat{P}(y_{0,d_0} \mid \mathbf{c}) \hat{P}(d_0 \mid \mathbf{c})}. \quad (259)$$

And by grounding,

$$\frac{P_{d_1}(y_1 \mid \mathbf{c}) - P(y_{1,d_1} \mid \mathbf{c})}{P_{d_0}(y_0 \mid \mathbf{c}) P(d_0 \mid \mathbf{c})} \leq \Omega(d_1, d_0, \mathbf{c}) \leq \frac{P_{d_1}(y_1 \mid \mathbf{c}) - P_{d_1}(y_1 \mid \mathbf{c}) P(d_1 \mid \mathbf{c})}{P_{d_0}(y_0 \mid \mathbf{c}) P(d_0 \mid \mathbf{c})}. \quad (260)$$

□